Adaptive Image Classification on Mobile Phones

von

Erich Bruns

Dissertation eingereicht an der

Fakultät Medien

zum Erreichen des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

an der

BAUHAUS-UNIVERSITÄT WEIMAR

Gutachter	Univ. Prof. Dr.-Ing. habil. Oliver Bimber Johannes Kepler Universität Linz
	Kari Pulli, Ph.D. Nokia Research Center, Palo Alto Zweitgutachter
Tag der Einreichung	19.05.2010

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Weitere Personen waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlung- bzw. Beratungsdiensten (Promotionsberater oder anderer Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Ich versichere, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.

Erich Bruns, Weimar 19.05.2010

ABSTRACT

The advent of high-performance mobile phones has opened up the opportunity to develop new context-aware applications for everyday life. In particular, applications for context-aware information retrieval in conjunction with image-based object recognition have become a focal area of recent research.

In this thesis we introduce an adaptive mobile museum guidance system that allows visitors in a museum to identify exhibits by taking a picture with their mobile phone. Besides approaches to object recognition, we present different adaptation techniques that improve classification performance.

After providing a comprehensive background of context-aware mobile information systems in general, we present an on-device object recognition algorithm and show how its classification performance can be improved by capturing multiple images of a single exhibit. To accomplish this, we combine the classification results of the individual pictures and consider the perspective relations among the retrieved database images. In order to identify multiple exhibits in pictures we present an approach that uses the spatial relationships among the objects in images. They make it possible to infer and validate the locations of undetected objects relative to the detected ones and additionally improve classification performance. To cope with environmental influences, we introduce an adaptation technique that establishes ad-hoc wireless networks among the visitors' mobile devices to exchange classification data. This ensures constant classification rates under varying illumination levels and changing object placement. Finally, in addition to localization using RF-technology, we present an adaptation technique that uses user-generated spatio-temporal pathway data for person movement prediction. Based on the history of previously visited exhibits, the algorithm determines possible future locations and incorporates these predictions into the object classification process. This increases classification performance and offers benefits comparable to traditional localization approaches but without the need for additional hardware.

Through multiple field studies and laboratory experiments we demonstrate the benefits of each approach and show how they influence the overall classification rate.

DEUTSCHE ZUSAMMENFASSUNG

Einleitung

Als Bill Schilit 1994 [130] den Begriff des ”context-aware computing” geprägt hat, ist ein neues Forschungsfeld entstanden, das sich zum Ziel gesetzt hat, neue Techniken und Algorithmen zu entwickeln, um intuitive Interaktionen zwischen Menschen, Plätzen und Objekten zu ermöglichen [51]. Dies führte im Laufe der Zeit zu einer Vielzahl von unterschiedlichen Anwendungen in der Forschung und Industrie, in denen z. B. Häuser auf ihre Einwohner eingehen [121], Mautstellen vorbeifahrende Autos registrieren [46], Touristen auf umliegende Restaurants aufmerksam gemacht werden [107] oder Tablettendosen erkennen, ob ihr Inhalt durch Patienten eingenommen wurde [14].

Die Entwicklung solcher kontextsensitiver Applikationen wurde von einigen Schlüsseltechnologien vorangetrieben, welche maßgeblich zu der Realisierung der Vision einer intelligenten Welt, in der jedes Element sich seiner Umgebung bewusst ist, beigetragen haben. Einer der wichtigen Schritte in Richtung der ”ubiquitous computing”-Vision von Mark Weiser [191] war die Entwicklung von RFID-Tags, die in den letzten Jahren klein und günstig genug wurden, um sie an beliebige Objekte anzubringen[1]. Durch die Verwendung von entsprechenden mobilen Lesegeräten können diese Tags ausgelesen und die übertragenen Informationen verwendet werden, um das Objekt zu identifizieren. In Verbindung mit integrierten Sensoren ist es den Tags zudem möglich, zusätzliche Informationen, wie Temperatur oder Beleuchtung, zu übermitteln, um Anwendungen in Echtzeit zu adaptieren.

Ebenso wichtig für die Verbreitung von ”context-aware computing” war die kontinuierliche Weiterentwicklung von Mobiltelefonen zu leistungsstarken Multimediasystemen. Heutzutage haben solche Geräte ausreichende Rechenleistung, Speicher, Kommunikationsschnittstellen (z. B. WLAN, Bluetooth), Sensoren (z. B. GPS, Beschleunigungssensor) und Visualisierungsmöglichkeiten (z. B. Display), um eine Vielzahl von unterschiedlichen kontextsensitiven Anwendungen zu ermöglichen.

Besonders die Entwicklung von Handys mit Internetverbindung und GPS-Empfänger, um die Position von Benutzern zu bestimmen, führte zum Anwendungsangebot der ortsbasierten Dienste (Location Based Services, LBS), mit denen Benutzer Zugriff auf Informationen über ihre nähere Umgebung erhalten können. Z. B. können sie Hinweise auf gastronomische oder kulturelle Ange-

[1] IDTechEx berichtete dass ca. 2.35 Mrd. RFID-Tags in 2009 verkauft wurden, gegenüber 1.97 Mrd. in 2008, 1.74 Mrd. in 2006 und 1.02 Mrd. in 2005 [91].

bote abrufen, herausfinden wo sich der nächste Bankautomat befindet oder wo der nächste Bus abfährt. Darüberhinaus ist es ihnen möglich mit ihrer Umgebung zu interagieren, um z. B. ein Busticket zu kaufen oder Sitzplätze im Theater zu buchen. Das Marktanalyseunternehmen Gartner stuft LBS als zweitwichtigste Mobilapplikation in ihrem Bericht "Dataquest Insight: The Top 10 Consumer Mobile Applications in 2012" ein. Grund hierfür ist der hohe Nutzwert von LBS und ihr Einfluss auf die Loyalität der Nutzer. Gartner prognostiziert, dass die Anwenderzahl von LBS von 96 Mio. in 2009 auf 526 Mio. in 2012 ansteigen wird [74].

Doch wie Albrecht Schmidt schon 1999 festgestellt hat, gibt es mehr Möglichkeiten um den Kontext einer Person zu bestimmen als nur über dessen Position [156]. Durch die kontinuierliche Weiterentwicklung von Mobiltelefonen in den Bereichen Rechenleistung, Speicher und Display in Kombination mit der Integration von Kameras, sind Handys nun in der Lage akzeptable Objekterkennungsalgorithmen durchzuführen. Dies ermöglicht das Identifizieren von Objekten in ihrer Umgebung durch einfaches Fotografieren. Jene, im Bereich "mobile human-computer interaction" als "Pointing" bekannte Interaktionstechnik, erscheint als eine natürliche und intuitive Möglichkeit mit seiner Umgebung zu kommunizieren [153]. Darüber hinaus ist diese Interaktion, im Vergleich zu ortsbasierten Systemen, eindeutiger, da Benutzer nur Informationen über das Objekt erhalten, was sie wirklich identifizieren möchten.

Anwendungsbereiche, in denen solch eine Interaktion vorteilhaft und vielversprechend scheint, sind digitale Museums- oder Stadtführer: Anwender fotografieren mit ihrem Handy ein Ausstellungsstück oder eine Sehenswürdigkeit und betreffende Multimedia-Informationen werden kurze Zeit später auf dem Gerät abgespielt.
Solch eine Funktionalität in einem unkontrollierten Umfeld und unter unterschiedlichem Nutzerverhalten zu gewährleisten, ist jedoch eine Herausforderung. Im Kontext eines mobilen Museumsführers präsentiert diese Arbeit Verfahren und Lösungen, um diese Herausforderung anzugehen.

Motivation

Die Einführung von Mobiltelefonen mit eingebauten Sensoren wie Kameras, GPS oder Beschleunigungssensoren, sowie Kommunikationstechniken wie Bluetooth oder WLAN ermöglicht die Entwicklung neuer kontextsensitiver Anwendungen für das tägliche Leben. Insbesondere Applikationen im Bereich kontextsensitiver Informationsbeschaffung in Verbindung mit bildbasierter Objekterkennung sind in den Fokus der aktuellen Forschung geraten.
Erste Ansätze in diesem Gebiet erlaubten einem Benutzer Gebäude [71] oder Ausstellungsstücke

[81] mit seinem Handy zu fotografieren. Das Bild wurde dann zu einem Server übertragen, um die Objekterkennung durchzuführen. Das Ergebnis wurde an das Handy zurückübermittelt und entsprechende Informationen auf dem Gerät abgespielt. Durch die kontinuierliche Leistungssteigerung aktueller Handys in letzter Zeit ist es nun auch möglich geworden, die Bilderkennung direkt auf dem mobilen Endgerät durchzuführen [169, 84, 52].

Unabhängig von der verwendeten Hardware (Handy oder PC) ist jedoch eine fehlerfreie Objekterkennung schwierig, besonders wenn tausende Objekte zuverlässig identifiziert werden müssen: Sie können ihr Erscheinungsbild in Abhängigkeit der Beleuchtung ändern und wirken anders, wenn man sie aus unterschiedlichen Perspektiven und Entfernungen betrachtet. Dies führt folglich zu einer unendlichen Zahl von verschiedenen zwei-dimensionalen Bildern [138]. Darüberhinaus können Objekte überlagert werden oder sich stark ähneln. Im Bereich der mobilen Objekterkennung sind diese Voraussetzungen umso schwerwiegender, da Mobiltelefone im Vergleich zu PCs über verminderte Systemressourcen und Rechenleistung verfügen. Dies hat zur Folge, dass keine beliebigen rechenintensiven Objekterkennungsalgorithmen oder Bildauflösungen verwendet werden können [87].

Zur Behebung der genannten Probleme, stellt der Forschungsbereich des maschinellen Sehens eine Vielzahl von unterschiedlichen Verfahren zur Objekterkennung bereit. Die Entwicklung und Auswahl von Objekterkennungsalgorithmen für mobile Geräte steht jedoch zurzeit noch am Anfang. Es ist allerdings wahrscheinlich, dass viele Algorithmen, die ursprünglich auf dem PC entwickelt worden sind, in Zukunft portiert und auf die gegebenen Systemvoraussetzungen der Handys adaptiert werden.

Die grundlegenden Herausforderungen der Objekterkennung bestehen jedoch weiterhin und sind unabhängig von der verwendeten Hardware. Aufgrund dessen betrachten wir das Problem von einer anderen Seite: Anstatt die Nachteile (z. B. Systemvoraussetzungen) von Mobiltelefonen gegenüber PCs zu analysieren und zu kompensieren, richten wir den Fokus auf deren Vorteile gegenüber stationären Computern. Basierend darauf können wir Techniken entwickeln, die beliebige Objekterkennungsalgorithmen unterstützen, um deren Erkennungsrate zu verbessern. Der Vorteil hierbei ist, dass diese ”Adaptionstechniken” zum größten Teil unabhängig von der verwendeten Objekterkennung sind. Schreitet also die Entwicklung von Handys weiter voran, können neue Objekterkennungsalgorithmen eingesetzt werden, dessen Qualität durch die Adaptionstechniken weiterhin verbessert werden kann.

Ein beliebtes Beispiel für solch eine Adaptionstechnik ist die Bestimmung der aktuellen Position von Benutzern durch GPS oder Bluetooth [71, 116, 75, 21]. Die resultierende Ortsinformation wird mit dem Objekterkennungsalgorithmus verknüpft, um die Erkennungsrate zu verbessern. Dies wird gewährleistet, indem nur die Objekte als mögliche Ergebnisse berücksichtigt werden, die im Umkreis des Benutzers liegen. Da jene nur einen kleinen Teil der Gesamtanzahl in der Objektdatenbank ausmachen, verbessert sich die Objekterkennung merklich.

Jener Sachverhalt motivierte uns, weitere Algorithmen und Techniken zu entwerfen, die über die traditionellen Verfahren hinausgehen, um die Objekterkennung auf Handys zu verbessern. Die Entwicklung und Evaluierung solcher Techniken ist der Gegenstand der vorliegenden Arbeit.

Ergebnisse der Arbeit

Der Beitrag dieser Arbeit ist die Entwicklung eines bildbasierten, mobilen Museumsführersystems, welches unterschiedliche Adaptionstechniken verwendet, um die Objekterkennung zu verbessern.

Es wird gezeigt, wie Objekterkennungsalgorithmen auf Mobiltelefonen realisiert werden können und wie die Erkennungsrate verbessert wird, indem man ad-hoc Netzwerke einsetzt oder Bewegungsvorhersagen von Personen berücksichtigt. Wir bezeichnen diese Funktionalität als ”adaptive image classification”.
Nachfolgend werden die relevantesten Beiträge dieser Arbeit aufgelistet:

- Übersicht und Klassifizierung von existierenden Techniken, die den Entwurf und die Entwicklung eines kontext-sensitiven mobilen Informationssystems ermöglichen.
- Entwicklung eines Objekterkennungsalgorithmus, der ein Ausstellungsstück identifiziert, indem es mehrere Bilder während einer kurzen Kamerabewegung aufnimmt. Neben einer kombinierten Bilderkennung wird das System darüberhinaus verbessert, indem es die Bildrelationen der gefundenen Datenbankbilder berücksichtigt.
- Entwicklung eines Objekterkennungsalgorithmus, der mehrere Ausstellungsstücke in einem Bild identifiziert. Hierfür werden die räumlichen Beziehungen der Objekte untereinander betrachtet, was zu einer Verbesserung der Erkennungsrate sowie der Erkennungsdauer führt.
- Entwicklung einer Adaptionstechnik, die es Mobiltelefonen erlaubt, untereinander automatisch Klassifikationsdaten auszutauschen. Dies ermöglicht die Unabhängigkeit der Bilderkennung von Umwelteinflüssen, wie wechselnden Beleuchtungssituationen oder Objektmanipulationen.

- Entwicklung einer Adaptionstechnik, die räumliche und zeitliche Bewegungsmuster von Besuchern im Museum verwendet, um die Objekterkennung zu verbessern.
  - Entwicklung einer Methode, um Bewegungen von Personen vorherzusagen. Hierbei wird die räumliche Lage von Ausstellungsstücken verwendet, um einen Klassifikator zu generieren. Basierend auf schon besichtigten Objekten, versucht der Algorithmus zukünftige Orte der Besucher vorherzusagen. Die Ergebnisse werden dann mit der Bildererkennung verknüpft.
  - Entwicklung einer Methode, die Bereiche im Museum identifiziert, in den Besucher nicht mehr zurückkehren, wenn sie sie einmal verlassen haben. Sobald ein Besucher sich aus solch einem Bereich entfernt hat, werden die Objekte in diesem Gebiet für zukünftige Objekterkennungen nicht mehr als mögliche Ergebnisse berücksichtigt.

Ein Grundkonzept dieser Verfahren ist der Verzicht auf jegliche zusätzliche Hardware (z. B. Sensoren), die während des Einsatzes des Systems nötig wäre. In Verbindung mit ihrer grundsätzlichen Unabhängigkeit von dem angewendeten Objekterkennungsalgorithmus weisen diese Verfahren die nötigen Voraussetzungen auf, um sie auf andere Systeme zu übertragen. Alle Verfahren wurden durch Benutzerstudien in Museen verifiziert, die die Vorteile unserer Verfahren im Vergleich zu nicht adaptierten Objekterkennungssystemen aufzeigen. Im Folgenden werden die einzelnen Beiträge dieser Arbeit näher erläutert.

Um diese Arbeit thematisch einzuordnen, haben wir zunächst den Begriff ”context-aware mobile information systems” geprägt. Er umschreibt Systeme, die den Kontext eines Benutzers erfassen und darauf basierend Informationen auf einem mobilen Gerät bereitstellen. Um den Kontext zu definieren, existieren, basierend auf der Analyse verwandter Arbeiten, zwei grundsätzliche Techniken: Auf der einen Seite gibt es Systeme, die die Position eines Benutzers ermitteln, um dessen Umgebung zu bestimmen. In Innenräumen basieren diese Systeme hauptsächlich auf unterschiedlichen Funktechniken (z. B. RFID, Bluetooth, WLAN), die zur Bestimmung der Position genutzt werden. Um solche Systeme thematisch weiter zu gliedern, haben wir aufgrund dessen die unterschiedlichen Signalreichweiten der Emitter als zusätzliches Kriterium zur Unterteilung verwendet. Auf der anderen Seite existieren mobile Anwendungen, die Bilderkennung einsetzen, um spezifische Ausstellungsstücke in der Umgebung durch Fotografieren zu identifizieren. Diese können wiederum eingeteilt werden in assistenzbasierte Systeme, die Referenzpunkte wie Marker oder Barcodes fotografieren müssen, um das dazugehörige Objekt zu klassifizieren, sowie in natürliche Bilderkennungsmethoden, die das Exponat erkennen, indem man es direkt aufnimmt.

Bei der Betrachtung der verwandten Arbeiten zeigte sich, dass Methoden, die auf mobiler Bildererkennung beruhen, nur die Position sowie die Orientierung des Benutzers in Betracht ziehen, um die Bilderkennung zu adaptieren. Da weitere Ansätze nicht berücksichtigt wurden, obwohl die vorgestellten Erkennungsraten der Verfahren Raum für Verbesserungen boten, bestätigte dies unsere Motivation, weitere Adaptionstechniken zu entwerfen.

Zunächst haben wir in dieser Arbeit unterschiedliche Verfahren entwickelt und vorgestellt, die ein oder mehrere fotografierte Objekte identifizieren können. Die Basis-Bilderkennung ist schnell, einfach sowie speichereffizient und erlaubt die Klassifizierung eines einzelnen Objekts mit Hilfe globaler Farbmerkmale und einem künstlichen neuronalen Netz. Dieses Verfahren haben wir bereits vor dieser Arbeit entwickelt.
Neben der Identifikation eines Objekts durch ein einzelnes Bild haben wir ein Verfahren implementiert, welches mehrere Bilder verwendet, um ein Ausstellungsstück zu erkennen. Diese Bilder werden durch eine kurze Bewegung des Handys von nah zu fern relativ zu dem Exponat aufgenommen. Neben einer kombinierten Klassifizierung der einzelnen Bilder durch Mehrheitsentscheidung werden die Bildrelationen der gefundenen Datenbankbilder untersucht. Nur die Bilder, die die gleichen Relationen wie die aufgenommenen Fotos aufweisen, werden für die Klassifizierung verwendet. Im Vergleich zur Objekterkennung durch ein einzelnes Bild, konnte im Zuge einer Benutzerstudie gezeigt werden, dass sich die Erkennungsrate um ca. 15% verbessert hat. Darüberhinaus führt die Berücksichtigung der Bildrelationen zu einer Beschleunigung der Klassifizierung von bis zu 35%, wobei im Durchschnitt 27,7% weniger Bilder benötigt wurden.

Um mehrere Objekte (sog. Subobjekte) in einem Bild zu erkennen, haben wir ein Verfahren entwickelt, das die räumlichen Beziehungen der aufgenommenen Subobjekte im Bild untereinander berücksichtigt. Diese räumlichen Beziehungen garantieren es, Suchregionen ausgehend von bereits gefundenen Subobjekten zu definieren. Dadurch ist es möglich, die Erkennungsgeschwindigkeit sowie die Erkennungsrate zu verbessern. Durch eine Benutzerstudie konnte ermittelt werden, dass durchschnittlich alle Subobjekte mit einer Wahrscheinlichkeit von 85,9% richtig erkannt werden. Ein erfahrener Benutzer, der die Subobjekte aus ähnlichen Perspektiven aufgenommen hat, wie sie trainierten worden sind, hat eine Erkennungsrate von 94,4% erzielt. Im Vergleich zu Ansätzen, die keine räumlichen Beziehungen verwendet haben, weist dies eine Verbesserung von ca. 10% auf. Die Klassifikationsgeschwindigkeit konnte in diesem Kontext um bis zu 68% verbessert werden. Neben der Definition von Suchregionen werden die räumlichen Beziehungen zudem verwendet, um die Position von Subobjekten zu erschließen, die vollständig durch andere Objekte verdeckt sind, was durch keine Objekterkennung zu gewährleisten ist.

Der Einsatz von Farbmerkmalen in unserem Verfahren macht die Objekterkennung anfällig gegen starke Beleuchtungsänderungen. Um dies auszugleichen, haben wir ein Adaptionsverfahren entwickelt, das Klassifikationsdaten unter den Mobiltelefonen austauscht.
Sobald ein Ausstellungsstück fotografiert wurde, liefert das neuronale Netz der Objekterkennung eine Sequenz von Klassifikationsdaten. Jedes Element der Sequenz weist einem Objekt, das trainiert worden ist, eine Wahrscheinlichkeit zu, dass es fotografiert wurde. Die gesamte Sequenz beschreibt das aktuelle Erscheinungsbild des fotografierten Objekts eindeutig und kann verwendet werden, um einen zweiten Klassifikationsschritt durchzuführen. Hierfür werden alle Sequenzen von allen Handys akkumuliert und durch ad-hoc Netzwerke über Bluetooth unter den einzelnen Geräten ausgetauscht. Die gesammelten Sequenzen dienen auf den Mobiltelefonen als Datenbank, um eine Nächste-Nachbarn-Klassifikation zwischen der aktuellen Sequenz der Bilderkennung und der Datenbank durchzuführen. Durch die kontinuierliche Synchronisation ist gewährleistet, dass jedes Mobiltelefon aktuelle Sequenzdatenbanken zur Verfügung hat. Durch das Untersuchen von unterschiedlichen Testfällen in einem Simulationsprogramm konnten wir feststellen, dass nach ca. 25 min, nachdem eine Beleuchtungsänderung stattgefunden hat, sich die Bilderkennung adaptieren konnte (abhängig von der Anzahl der Besucher). Dies bedeutet, dass die Objekterkennung die gleiche Erkennungsrate aufweisen konnte, wie vor der Änderung. Um realistische Ergebnisse in der Simulation zu erzielen, sammelten wir unterschiedliche Daten, wie Bilder und Positionen einzelner Objekte, Grundrisse oder räumliche Besonderheiten im Museum.

Eine häufig genutzte Adaptionstechnik, um die Objekterkennung zu verbessern, ist die Bestimmung der Position von Benutzern. Bei einer Objekterkennung werden dann nur die Objekte als mögliche Ergebnisse in Betracht gezogen, die sich in der näheren Umgebung des Museumsgasts befinden. Um die Bestimmung des Ortes in Innenräumen zu gewährleisten, werden häufig Funkwellen-Sender an bekannte Positionen angebracht. Erhält ein mobiler Empfänger Signale einer oder mehrerer Sender, kann er seinen aktuellen Ort bestimmen. Diese Art der Positionsbestimmung hat jedoch einige Nachteile: Zunächst müssen Sender flächendeckend in Gebäuden verteilt werden, was zum einen Kosten verursacht und zum anderen geschultes Personal zum Warten benötigt. Darüber hinaus beansprucht das kontinuierliche Suchen nach Funkwellen-Sendern die Akkulaufleistung des Handys enorm.
Aufgrund dessen haben wir ein Verfahren entwickelt, das die Vorteile einer Positionsbestimmung hat, aber keine zusätzliche Hardware einsetzt. Um dies zu gewährleisten, haben wir zunächst eine Benutzerstudie über vier Monate in einem Museum durchgeführt, um zu bestimmen, wie Besucher durch Museumsräume laufen. Basierend auf diesen Daten haben wir zwei Techniken entwick-

elt. Die erste Methode versucht Vorhersagen über Ausstellungsstücke im Museum zu treffen, die Besucher als nächstes aufnehmen werden, basierend auf vorherigen Fotografien. Hierfür wird ein neuronales Netz mit den Wegrouten unterschiedlicher Besucher trainiert. Beschrieben werden diese Wegrouten durch die fotografierten Objekte, sowie über solche, die in der näheren Umgebung liegen. Dadurch ist es möglich, richtige Vorhersagen zu treffen, auch wenn der Pfad, den der jeweilige Besucher gegangen ist, niemals zuvor beschritten wurde. Dieses Verfahren macht sich dabei zu Eigen, dass es z. B. wichtig ist, dass ein Besucher vor einer Vitrine stand, es jedoch unbedeutend ist, welches Objekt er in der Vitrine fotografiert hat. Das Ergebnis der Vorhersagen wird dann mit dem Ergebnis der Bilderkennung verknüpft, um das Gesamtresultat zu erhalten. Das zweite Verfahren basiert auf der Annahme, dass Bereiche im Museum existieren, in die Besucher nicht mehr zurückkehren, wenn sie sie einmal verlassen haben. Die Exponate in diesen Bereichen werden dann für zukünftige Objekterkennungen nicht mehr als mögliche Ergebnisse berücksichtigt. Um solche Bereiche automatisch zu ermitteln, haben wir die gesammelten Wegrouten als gewichteten, gerichteten Graphen umformuliert. Durch Tiefensuche können nun Teilgraphen gefunden werden, deren Kanten zu anderen Teilgraphen entweder alle nach innen oder nach außen zeigen. Für die entsprechenden Objekte in jeden Teilgraphen wird dann ein neuronales Netz generiert. Verlässt ein Besucher also solch einen Bereich, wird für zukünftige Objekterkennungen das neuronale Netz gewählt, das die Objekte aus diesem Bereich nicht mehr trainiert hat.
Eine Benutzerstudie hat gezeigt, dass die Erkennungsrate in Verbindung mit diesen beiden Verfahren um ca. 20% gesteigert werden konnte. Somit erzielen sie ähnliche Ergebnisse wie die gleiche Objekterkennung in Verbindung mit einer Positionsbestimmung durch Bluetooth-Sender – jedoch ohne den Einsatz zusätzlicher Hardware.

Durch Verknüpfung aller Verfahren in dieser Arbeit konnte unsere Basis-Objekterkennung um knapp 40%, von $\sim$50% auf $\sim$89%, gesteigert werden. Ein Vergleich hat gezeigt, dass dies durch den Einsatz eines komplexeren Objekterkennungsalgorithmus, der keine Adaptionstechniken verwendet, nicht zu erreichen ist.

ACKNOWLEDGMENTS

First of all I would like to thank my supervisor Prof. Oliver Bimber for his support, guidance and continuous assistance throughout the years of my studies and Ph.D. I would also like to thank Kari Pulli for being my second supervisor and for allowing my project to be funded by Nokia, which would have not been possible without his personal efforts.

I am grateful to the Stiftung für Technologie, Innovation und Forschung Thüringen for funding the PhoneGuide project. I am indebted to Alf Rößner, director of the Stadtmuseum Weimar and Matthias Hartmann, director of the Naturkundemuseum Erfurt, who both supported our project by allowing us to conduct experiments in their museums.

I would also like to thank Tobias Langlotz for discussing my thesis and for his reliable availability whenever I needed mental distraction, Benjamin Brombach for his contributions and comments, Anselm Grundhöfer for daily conversations, Chrystoph Toll for supplying cake, Stefanie Zollmann for reading my thesis, Julian Reisenberger for proof-reading and everyone else in the lab for making life pleasant these past years.

Finally, for the support, encouragement and patience over the years, I would like to thank Theres.

CONTENTS

1 Introduction 1

1.1 Motivation . . . 2

1.2 Contribution . . . 4

1.3 State of Research prior to the Dissertation . . . 8

1.4 Publications . . . 10

1.5 Outline . . . 11

2 Background and Related Work 13

2.1 Location-Based Information Systems . . . 15

2.1.1 Short-Range Localization . . . 15

2.1.2 Long-Range Localization . . . 19

2.2 Vision-Based Information Systems . . . 22

2.2.1 Assisted Vision-Based Techniques . . . 23

2.2.2 Natural Vision-Based Techniques . . . 24

2.3 Related Research Areas . . . 29

2.3.1 Object Recognition . . . 30

2.3.2 Wireless Ad-hoc Networks . . . 33

2.3.3 Person Movement Prediction . . . 35

3 Adaptive Mobile Museum Guidance System 37

3.1 Mobile Museum Guidance . . . 37

3.2 PhoneGuide . . . 38

3.2.1 Basic Object Recognition . . . 38

3.2.2 Adaptation through User Feedback . . . 40

3.2.3 Adaptation through Bluetooth Tracking . . . 42

3.3 Preliminaries . . . 44

4 Multi-Image Classification 46

4.1 Relational Multi-Image Classification . . . 47

4.2 Off-line Preprocessing . . . 49

4.2.1 Detecting Image Relations . . . 49

4.2.2 Generating Image Classifier . . . 50

4.3 On-line Relational Multi-Image Classification . . . 51

4.4 Graphical User Interface 52

4.5 Evaluation 53

4.6 Discussion 57

5 Subobject Recognition 59

5.1 Off-line Preprocessing 60

5.1.1 Registration and Tracking of Subobjects 61

5.1.2 Generation of Subobject Classifiers 62

5.1.3 Extraction of Spatial Relationships 63

5.2 On-line Subobject Recognition 64

5.2.1 Detection of Anchor Subobjects 65

5.2.2 Detection of Remaining Subobjects 68

5.3 Graphical User Interface 69

5.4 Evaluation 69

5.4.1 Performance Analysis 71

5.4.2 Field Experiment 74

5.5 Discussion 75

6 Adaptation through Phone-to-Phone Communication 77

6.1 Adaptation through Ad-hoc Networking 78

6.2 Synchronization 81

6.3 Implementation 82

6.4 Simulation 84

6.5 Evaluation 87

6.6 Discussion 90

7 Adaptation through Pathway Analysis 92

7.1 Data Acquisition 93

7.2 Generation of the Pathway Classifier 95

7.3 Ensemble Classification 98

7.4 Segmentation through Pathways 98

7.5 Evaluation 100

7.6 Discussion 104

8 Evaluation 106

8.1 System Architecture 106
8.2 User Interface 110
8.3 Classification Performance 111
8.4 Discussion 114
8.4.1 Advantages and Disadvantages 114
8.4.2 Transferability 116

9 Conclusion and Future Work 118

9.1 Conclusion 118
9.2 Future Work 121

A Appendix 124

A.1 Questionnaire for Evaluating Multi-Image Classification 124
A.2 Questionnaire for Evaluating Subobject Recognition 125
A.3 Questionnaire for Analyzing Pathways 127
A.4 Estimation of Classification Rate 129

References 130

LIST OF FIGURES

1.1 Illustration of the basic concept of our museum guidance system. Adaptation techniques such as Bluetooth tracking or phone-to-phone communication are visualized in a museum scenario. 5

2.1 Overview of the main techniques for realizing context-aware mobile information systems such as mobile museum guides. 14

2.2 Selection of different barcodes. 23

2.3 The three conceptual areas of mobile natural vision-based techniques: client-server based, on-device classification based using high-performance devices and on-device classification based using smart phones. In addition, release dates of popular mobile phones (with CPU clock rate) and the different research steps of the whole PhoneGuide project. 25

3.1 A selection of different audio guides that use either reference numbers (a), electronic lists (b) or RFID tags (c) to identify exhibits in a museum. 38

3.2 User interface representing the basic object recognition approach. A video stream allows visitors to point at an exhibit. A floor plan indicates the rough location of the visitor in the museum (a). The classification result is presented as a probability-sorted list of result candidates visualized as thumbnails. The selected object is previewed above the list. An information icon indicates what kind of multimedia content (video, image, audio) is available (b). After selecting the correct object, multimedia content is displayed (c). 41

3.3 Bluetooth emitter [175] with battery pack (scale in millimeter). The Bluetooth beacons are distributed throughout the museum to ensure cell-based localization. . 43

3.4 The phones used for evaluating the different approaches: Nokia 6670 (CPU: TI OMAP 1510 123MHz; display: 2.1”, 176x208; camera: 1.0 MPixel; Bluetooth 1.1, released: October 2004). Nokia 6630, Nokia 6680 (both, CPU: TI OMAP 1710 220MHz; display: 2.1”, 176x208; camera: 1.3 MPixel; Bluetooth 1.1, released: November 2004, May 2005). Nokia N95 (CPU: TI OMAP 2420 332 MHz; display: 2.6”, 240x320; camera: 5.0 MPixel; Bluetooth 2.0, released: March 2007) [49]. . . 44

4.1 A visitor in a museum captures a sequence of images with his mobile phone by moving the device backwards. Afterwards distinct keyframes are extracted and matched against a database. Only those database images that exhibit the same near-far relations as the captured video frames are used to derive the final classification result. 46

4.2 Preprocessing steps for retrieving image relations: After recording videos of each exhibit, keyframes are extracted. For each possible pair of keyframes, SIFT features are computed and correspondences in both images are detected. The relation between each pair of images is then derived from the motion vectors defined by the correspondences. If motion vectors, originating in at least two different quadrants, point towards or away from the center of the image, a near-far movement is detected (a). The image relations are then encoded and stored in a relation table (b). Ambiguous image relations are resolved by inferring their relations based on other known image pairs (c). One relation table is stored for each exhibit. 48

4.3 Flow chart showing the user interface for single and multi-image classification. If the user takes a single photograph of an exhibit (1a), an image classification is carried out and the result is presented as a probability-sorted list of objects (1b). The correct object can be selected with a minimum number of clicks in order to retrieve the corresponding multimedia information (1c). If a set of images was recorded (2a) the relational multi-image classification is carried out for each keyframe. Intermediate results are visualized through object thumbnails in combination with the corresponding classification rate on the screen (2b). Users are allowed to select the correct result during the classification phase or by choosing from a list of objects that is presented at the end of the classification process to retrieve the multimedia information (2c-2d). 52

4.4 Result of a multi-image classification experiment, using a near-far camera movement. The multi-image classification was applied with and without relational reasoning and its performance is shown for different numbers of keyframes. The performance of single-image classification is independent of the number of keyframes and therefore constant. 54

4.5 Comparison between near-far camera movement (a,b) and digital zoom from the far position (c). 55

4.6 Result of a multi-image classification experiment, using a digital zoom of one shot. The multi-image classification was applied with and without relational reasoning and its performance is shown for different numbers of keyframes, comparable to figure 4.4. 56

4.7 Single-image classification and the different elements of our relational multi-image classification and their impact on the classification rate depending on the number of extracted keyframes. 58

5.1 A visitor takes 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, the known spatial relationships among the subobjects still allow a correct identification. 59

5.2 Flow chart of the off-line preprocessing: After a video of a scene with subobjects is 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). 62

5.3 Predicted search area of an undetected subobject (B) relative to a detected subobject (A). The corresponding ring sector is defined by the minimum and maximum distances and angles that were extracted during subobject tracking in the off-line preprocessing. 64

5.4 After an image was captured, the subobject detection searches for the first anchor subobject by varying the size of the search mask (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 cross sections that define reliable search areas of the remaining subobjects (g). This is repeated until all subobjects are detected (h). 66

5.5 Flow chart showing 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 list of objects (1b). The correct object can be selected with a minimum number of clicks for receiving multimedia information (1c). If a group of subobjects is captured rather than a single object (2a), the user first has to acknowledge the scene classification (2b), before a consecutive subobject classification can be 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). 70

5.6 Average recognition rate and number of classifications for a brute-force search, a brute-force search 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. 72

5.7 Number of classifications required for brute-force search, brute-force search 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. . . . 73

6.1 Adaptation example (sequence numbers encircled): Object 1 (chair) is first successfully recognized on phone A (1). The voted classification sequence is time-weighted with f_t^2 and stored in A's *LODB*. Then object 2 (plate) is successfully recognized on phone A, and the *LODB* is expanded accordingly (2). Phone B fails in recognizing object 1 (3), and stores the weighted vote sequence in its local *LODB*. Then phones A and B are able to establish a connection. They both update their *GODB*s first (4+5) before synchronizing them (6). Next, a connection between phones B and C is established and *GODB*s are synchronized (7). The classification of object 1 on phone C (8) would fail when relying solely on the feedback from the neural network, but it will succeed if the nearest neighbor classification is performed together with the *GODB*. 79

6.2 Steps for synchronizing one *GODB* row: If phone A synchronizes with phone B, A's *GODB* row has to be updated before being transmitted. After receiving it at B, this row replaces the existing row in B's *GODB*. Then, B has to perform an update and sends the result back to A, where it replaces the corresponding *GODB* row. After doing this for each row, the *GODBs* of A and B are synchronized and contain the most up-to-date information. 82

6.3 Propagation example of classification and feedback data in a mobile ad-hoc network (sequence numbers encircled): Initially, all phones store their local *LODB*s only. First, phone C synchronizes with phone D (1), then C with B (2), then A with B (3), and finally A with C (4). After these steps, A, B, C share the same and most up-to-date information that have been collected by all phones. Phone D would also be up-to-date if it would now synchronize with either A,B, or C. 83

6.4 Simulation screenshot: The floor plan of the City Museum of Weimar with the locations of the exhibited objects (blue boxes). For each successful ad-hoc synchronization during simulation, the locations of both phones have been tagged (red dots). The majority of successful data exchanges are in the proximities of closely located objects, since visitors gather together in these areas for a longer time. The solid green lines define barriers that do not transmit the Bluetooth radio signal. The dotted yellow frame outlines the area in which we tested our approach with real image data and under realistic illumination conditions. 85

6.5 User interface of our simulation application. Left: the necessary parameters to configure the simulation. Right: On-going simulation. Bottom: real-time recording of the classification rate. 86

6.6 Seven hemispherically aligned light sensors with internal memory to record the illumination state over multiple days. 87

6.7 Average recognition rates for an unadapted (blue) image classification, and an image classification that is adapted through ad-hoc synchronization (dark red: 12 visitors/h and light red: 36 visitors/h), simulated over the opening hours of the museum under measured lighting changes (green). The adapted classification always performs better than the non-adapted classification, is less fragile to dynamic environmental changes, and recovers quickly to sudden changes. 88

6.8 Same as in figure 6.7, but with constantly decreasing illumination. While the recognition rate of the non-adapted classification decreases, the recognition rate of the adapted classification drops too – but is always better. If the performance of the non-adapted classification stagnates, the recognition rate of the adapted classification improves through ad-hoc synchronization. 89

7.1 Color-coded pathway visualization of 132 museum visitors in the Natural History Museum Erfurt, Germany. Each transition between two objects is visualized for the first floor (a). The main routes for all floors are visualized in b-d. MVT denotes the minimum visitor transitions. Edges are visualized if the number of visitor transitions between the corresponding objects is larger or equal than MVT. Several sample exhibits are illustrated on the right. 94

7.2 Algorithm for translating pathways (a) into a 2D map representation (b). Each entry of the pathway map represents one exhibit. Their weights reflect the current segment of a visitor's path. The highest weight is assigned to the current (captured) object, and gradually decreased values indicate neighbor objects within range (c). To decode the order of visited exhibits, the pathway maps are normalized in such a way that the weights of previously captured objects are attenuated over time (d-e). . 96

7.3 Illustration of pathway clustering by applying a depth-first search for two different starting nodes (a, b). 99

7.4 Classification rates for different approaches: image classification only, image classification and Bluetooth (BT) localization, image classification and localization through pathway clusters, image classification and pathway classification, image and pathway classification with localization via pathway clusters, image and pathway classification with Bluetooth localization. 101

7.5 Illustration of the classification rate of the image classifier with and without the pathway classifier and pathway clustering for each visitor. It proves the pathway classifier's independence of the path length (number of successively photographed exhibits) as well as its consistent benefit for image classification. 102

7.6 The probability of selecting the correct next object is significantly higher using prediction (red line) as opposed to a random selection (blue line). 104

8.1 Overview of our adaptive mobile museum guidance system: Classification data (image and pathway classifiers, spatial relationships, relation tables, rules) and multimedia in combination with the front-end application are initially retrieved by the phones from a server when entering the museum. During runtime, no server connection is necessary. The visitors interact with the classification system through a user interface by taking photographs and providing interaction feedback. Through the sensor interface, phones can search for external sensors to determine their approximate location while the phone-to-phone interface makes it possible to exchange classification data in order to cope with sudden environmental changes. Instead of sensors, the pathway classifier can be applied to predict the location of users. When leaving the museum, adaptation parameters are transmitted to the server to help continuously improve the system over time. 108

8.2 Flow diagram showing the different elements of our user interface if either a single exhibit is photographed or if multiple objects in a single image are captured. . . . 110

8.3 The benefits of our adaptation approaches for the classification rate of the object recognition approach (red), of the object recognition approach when the top 3 results are considered (purple) and of the subobject recognition technique (green). 112

8.4 Overview of the advantages and disadvantages of the different approaches presented in this thesis. 115

1 INTRODUCTION

Since Bill Schilit first coined the term ”context-aware computing” in 1994 [130], a new field of research has emerged that tries to establish technologies and algorithms to facilitate seamless and intuitive interaction between humans, places and objects [51]. This has led to a variety of different applications in research and industry where, for example, houses have become aware of their inhabitants [121], toll gates aware of passing cars [46], tourists made aware of nearby restaurants [107] and pill dispensers in a hospital aware of the respective patients [14].

The development and evolution of such context-aware applications has been driven by several key technologies and are significant for their contribution to the realization of the vision of a smart environment where every entity becomes intelligent. One such key step towards the vision of ”ubiquitous computing” [191] is the continuous miniaturization and improved performance of microelectronics. One popular example is the introduction of radio frequency identification (RFID) tags which have become cheap and small enough to attach them at arbitrary objects[2]. By utilizing mobile readers these RFID emitters are read out and the transmitted information is used to identify the object. In combination with accompanying sensors the emitter can also provide further information about the surroundings, such as the current temperature or illumination state which allows applications to adapt in real-time.

Another key factor for the emergence of context-aware computing was the continuous evolution of mobile phones from a minor-functional low-performance device into a high-performance multimedia system. These devices now have sufficient computing power, storage, communication interfaces (e.g., WLAN, Bluetooth, etc.), sensors (e.g., GPS, accelerometer) and visualization capabilities (display) to serve a variety of applications. In addition, mobile phones have become a part of daily life and are always with us. People now have easy access to a device that allows them to interact with the environment and which they can use ”as a window onto the information space” [65]. This opens up new ways for context-aware applications.

In particular, the advent of mobile phones with fast and wireless Internet connections in combination with GPS receivers to determine the location of users has become a fundamental concept for context-aware applications and has led to the field of Location Based Services (LBS) where users get access to a variety of information related to their environment. For example, users can retrieve information about places to visit in their vicinity, where the next ATM is located or the next bus

[2] IDTechEx reports that approximately 2.35 billion tags were sold in 2009 versus 1.97 billion in 2008; 1.74 billion in 2006 and 1.02 billion in 2005 [91].

departs. In addition, they can interact with these co-located objects via their mobile phone, for example to buy a ticket for the bus or book seats for the theater. Gartner ranks LBS as the second most important mobile application in their report ”Dataquest Insight: The Top 10 Consumer Mobile Applications in 2012” because of ”its perceived high user value and its influence on user loyalty”. They predict that the ”LBS user base will grow globally from 96 million in 2009 to more than 526 million in 2012” [74].

But ”there is more to context than location” as Albrecht Schmidt pointed out as far back as 1999 [156]. As a result of the significant performance improvements in recent years in combination with integrated cameras[3], mobile phones are now able to carry out acceptable object recognition algorithms. This has opened up the opportunity to identify objects in the environment simply by photographing them. This emerging and promising interaction technique also known as ”pointing” in the field of mobile human-computer interaction seems to be an intuitive behavior for communicating with one's context [153]. In addition, the interaction is more distinct compared to location-based applications since users only retrieve information about what they really want to identify.

Areas where this kind of interaction is both beneficial and popular are digital museum or city guidance applications: Users just point their mobile phone at an object and corresponding multimedia information is presented within seconds.
However, to ensure such functionality in uncontrolled and public environments with varying user behavior is challenging and an emerging research field. This thesis presents approaches to accomplishing these challenges and demonstrates solutions for designing a mobile museum guidance system.

1.1 Motivation

The advent of high-performance mobile phones in combination with built-in sensors such as cameras, GPS or accelerometers, as well as wireless communication technology such as Bluetooth or WLAN, has opened up the opportunity to develop new context-aware applications for everyday life. Especially applications for context-aware information retrieval in conjunction with image-based object recognition have become a focal area of recent research.
Initial approaches in this field allowed users to photograph buildings [71] or exhibits in museums [81] with their mobile hand-held device. The corresponding images were transferred to remote

[3] InfoTrends reports that the worldwide shipments of camera phones will jump from over 700 million units in 2007 to surpass 1.3 billion in 2012 [92].

servers to carry out the object classification. The results were then relayed back to the device and related multimedia information was presented. Recently, the performance of mobile phones has improved to such a degree that it is now possible to develop methods for performing identification directly on the device by applying object recognition algorithms [169, 84, 52].

Independent of the hardware used (mobile phone or PC), visual object recognition is a very challenging task, especially if thousands of objects need to be distinguished and identified reliably: objects change their appearance as the illumination varies and look different from varying perspectives and at arbitrary scales which result in ”an infinite number of different 2D images” [138]. Additionally, they can be occluded or they share a similar appearance by nature. In the context of mobile object recognition techniques, this task is even more demanding as mobile devices have comparatively limited system resources such as processing power, memory and power supply compared with desktop-PCs [87]. This means that it is not possible to use arbitrary high-performance object recognition algorithms or image resolutions efficiently.

In the field of computer vision the amount of different approaches that address the challenges mentioned above is tremendous [174]. At present, the development of object recognition methods for mobile phones is in its infancy and we believe that in future many different object recognition algorithms from desktop platforms will be ported to mobile phones and adjusted with respect to their current hardware limitations.

The fundamental challenges of object recognition, however, still remain regardless of the hardware used. Thus, we will take another viewpoint: rather than considering the drawbacks (system resources) of mobile phones compared with desktop-PCs we will focus on their advantages and differences to traditional PCs. From these, we can then design techniques which support and improve arbitrary object classification methods on mobile phones. The resulting advantage is that these ”adaptation techniques” are mainly independent of the applied object recognition algorithm: as the performance of mobile phones increases further, new object recognition algorithms can be used without renouncing the benefits of these methods.

One popular example of such an adaptation technique is the ability to determine the current location of users via, e.g., GPS or Bluetooth tracking. The resulting location information is then combined with object recognition algorithms in order to improve classification performance [21, 71, 75, 116]: Location awareness makes it possible to scale down the number of potential result candidates in the object database to the exhibits that are in the vicinity of the user. Since the remaining objects are only a fraction of the entire database, this method increases the classification rate significantly. This gave us the motivation to develop object recognition algorithms as well as further techniques that go beyond traditional approaches to improve the object recognition performance on mobile devices. In the context of a mobile museum guidance system the design and evaluation of such algorithms and adaptation techniques is the subject of this thesis.

1.2 Contribution

The contribution of this thesis is the design and development of a vision-based mobile museum guidance system that utilizes different adaptation techniques to improve the overall classification rate.

This work shows how object recognition[4] can be realized on mobile phones and how classification performance can be improved by employing techniques such as ad-hoc networking or person movement prediction. To comprise this functionality we coin the term ”adaptive image classification”. Figure 1.1 gives an impression of the basic concept of our system.

The following list provides a brief overview of the most relevant contributions of this thesis:

- Overview and Classification of existing techniques for designing and developing context-aware mobile information systems.
- Development of an object recognition algorithm that classifies a single exhibit by using multiple photos captured during a short camera movement. Alongside combined image classification, classification performance is further increased by using image relations between the retrieved database images as an additional classification constraint.
- Development of an object recognition algorithm that detects multiple objects in a single image by identifying spatial relationships among the exhibits in the captured photo. This improves the classification rate as well as the classification duration.
- Development of an adaptation technique that allows mobile phones to seamlessly exchange classification data among each other to ensure robustness against environmental changes such as variations in illumination levels or object manipulation.
- Development of an adaptation technique that utilizes user-generated spatio-temporal pathway data to unobtrusively improve classification performance.

[4] The terms object classification, image classification or image recognition are used synonymously throughout this thesis.

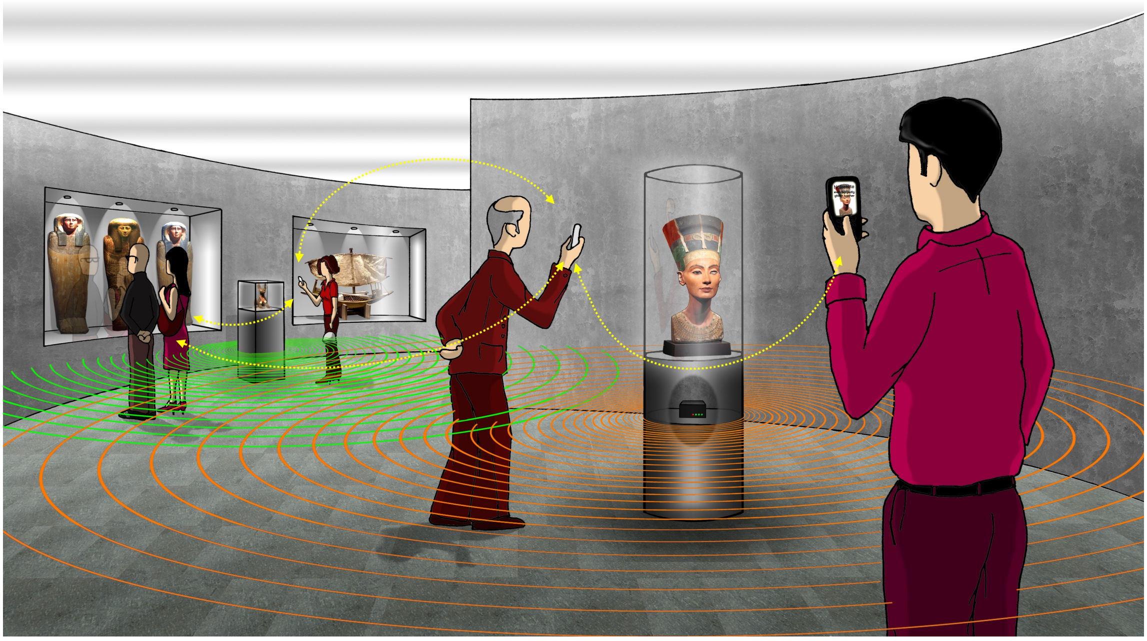


Figure 1.1: Illustration of the basic concept of our museum guidance system. Adaptation techniques such as Bluetooth tracking or phone-to-phone communication are visualized in a museum scenario.

- Development of a person movement prediction method that takes into account the spatial co-locations of visited objects for learning and inference. Based on the history of previously visited exhibits, the algorithm determines future locations and the predictions are integrated into the object classification process.
- Development of a technique to find physical areas in the museum that visitors do not return to after leaving. The corresponding exhibits in these areas are excluded as possible result candidates during the classification process after visitors have left them.

A core concept of these approaches is that they do not require any additional hardware such as sensors for adaptation during the application of the system. In combination with the basic independence of the image features or object recognition algorithm used, they ensure that the system can in future fulfill the basic requirement of transferability. All of these approaches have been verified through multiple field studies in realistic museum environments. The experiments revealed the benefits of our methods compared to basic object recognition algorithms.

The remainder of this section provides details of these contributions and the results of the techniques developed.

To provide a comprehensive background of the underlying topic of this thesis, we define the term ”context-aware mobile information systems” and give an overview of the techniques used to develop such systems.

We divide them into location- as well as vision-based applications. Both fields are characterized and are further subdivided into short- and wide-range localization as well as into assisted and natural vision-based techniques. For each area we provide an extensive selection of related work and discuss further research areas that we cover with our approaches. The investigation of related mobile vision-based techniques revealed that to improve classification performance only the location and the orientation of the corresponding mobile phone are considered as additional adaptation parameters. However, the classification rates of many approaches showed room for improvement and motivated the development of further adaption techniques as presented in this thesis.

Besides the identification of an object through a single picture, we developed an object recognition algorithm that classifies a single exhibit by utilizing multiple photos.

These photos are captured during a short near-far camera movement. Through combined image classification, a user study in a museum revealed that we enhance the recognition rate in comparison to single-image classification by up to $\sim$15%. In addition we extended this technique by considering the relations among the retrieved database images. If they do not follow the near-far relationship, they are not used during the classification phase. This increases the classification speed by up to 35% and saves on average 27.7% keyframes if the number of available keyframes is low. We show that a digital zoom technique is not an adequate alternative to near-far camera movement.

As a consecutive classification step based on the basic object recognition task, we developed an algorithm that detects multiple objects in a single image by identifying the spatial relationships among the exhibits in the captured picture. To accomplish this, a search mask spirally searches in the center of the image for known objects. If an exhibit is found, the spatial relationships span search areas relative to the detected object so that the search for remaining objects is carried out only in the predefined image areas. This reduces the number of classification steps and increases the classification rate since it narrows down the number of possible locations of the exhibit.

In laboratory experiments we evaluated this approach in comparison to two brute-force methods that use the same object recognition algorithm but without the application of spatial relationships. We achieved an average classification rate of 94.4% which outperforms both brute-force methods by $\sim$10%. Furthermore, our approach outperforms the related brute-force methods by up 68% in terms of classification time.

In the course of a field study our technique achieved an average classification rate of 85.9% under realistic conditions in a museum.

In order to cope with environmental changes such as variations in illumination levels or revised object placement, we developed an adaptation technique that allows mobile phones to seamlessly exchange classification data between each other.
After an object was photographed, the image classifier outputs a classification sequence that assigns each object a probability that it was photographed. This classification sequence is unique for each object and serves as a feature vector that describes the current appearance of the object. By distributing these sequences, they serve as an additional database for carrying out a second classification step on phones that subsequently approach this object. Synchronization is ensured by establishing ad-hoc networks among co-located mobile phones via Bluetooth. Through simulated test cases we show that this approach makes it possible to adapt to environmental changes after a short period of time depending on the number of visitors.

As an alternative to more conventional indoor localization systems such as Bluetooth tracking, which we have implemented in the past, we have developed an adaptation technique that makes use of the pathways of museum visitors.
To accomplish this, we carried out a user study over four months to determine how visitors move through a museum. Waypoints of visitors were defined by exhibits they find interesting and which they should photograph to simulate the use of our museum guidance system. Based on this data we designed two different adaptation methods: first, we developed a person movement prediction technique that takes into account the spatial co-locations of photographed objects for generating a pathway classifier. Based on the history of already visited exhibits, the classifier predicts future locations. The results are combined with the image classifier through average voting to receive a final outcome. For the second technique, we determined physical areas in the museum that visitors do not return to after visiting them by translating the pathway data into a weighted directed graph. This makes it possible to find solitary areas in the museum. After visitors have left these areas, the corresponding exhibits are excluded as possible result candidates during classification.
For improving the quality and realism of the evaluation, we have also natively identified the exhibits that are of interest to the visitors by examining the pathway data of the user study. These exhibits were used for the evaluation, and the corresponding user study showed that the recognition rate of image-based classification can be improved by up to 20% when combined with our pathway prediction methods. In addition we demonstrate that these techniques are competitive compared to image-based classification using Bluetooth localization –but without requiring auxiliary hardware.

By applying all the techniques we developed in conjunction, we demonstrate that the recognition rate of our basic object recognition system can be improved by up to ∼40% to ∼89% in a real-world scenario. We show that this significantly outperforms results of a high-performance object classification system that do not apply additional adaptation techniques.

An overview of the different adaptation techniques reveals their pros and cons in terms of benefit and practicability for future applications. Furthermore, we offer an outlook of additional techniques that were not realized in this thesis. For example, an ad-hoc phone-to-phone localization technique that determines the position of visitors through the relative distances to other users. Finally, we outline what kind of improvements could be beneficial in future.

1.3 State of Research prior to the Dissertation

The approaches of this thesis are based on the results of a research project called PhoneGuide. Following this relationship we use PhoneGuide as a synonym for the results of this work throughout this thesis. The state of the PhoneGuide research project prior to this dissertation is outlined in this section. Further details on these approaches are given in section 3.2.

The PhoneGuide project has started in 2004 with the goal of designing an application that applies object recognition algorithms on mobile phones for museum guidance purposes: after installing the PhoneGuide software on their mobile devices, visitors in a museum are able to use their mobile phone as an electronic information retrieval system. Each time they want to know more about an exhibit, they take a photo of it. Moments later, the software has classified the captured image and related multimedia content, e.g., audio, video, text or images, is presented.

To accomplish this application, the first approach of PhoneGuide applied 14 global color features (mean, variance in RGB channels, 4-bin histogram) for image description and a 1-layer artificial neural network (Perceptron) for classification. The classifier was trained directly on the mobile device [67]. One drawback of this approach was the lack of scalability: If the number of different exhibits was too large the classification rate dropped. To overcome this, PhoneGuide was extended to locate itself indoors [36]. Therefore, small battery-equipped Bluetooth emitters were distributed in each room in the museum. They transmitted a unique ID with a maximum range of 10 m. The mobiles phones scanned continuously for these emitters and if one or more were detected, only the exhibits within signal range were trained.

In 2007, the first step towards an adaptive mobile museum guidance system was mastered by developing a user interface that allows for continuously collecting image data through user feedback [30]. After each classification, a probability-sorted list of result candidates (encoded as image thumbnails) was presented to the user. If the classification failed, the visitor could select the correct image by browsing through this list. This final means of selection provides a way of validly mapping the captured picture to the actual object. Consequently, these images were stored and utilized to train the image classifiers and improve them over time.

In addition, the captured photos were separated into predefined patches and for each patch, a 3-layer neural network was trained on a server in advance to master the growing amount of data. The classification, however, was still performed on the mobile device by transferring the neural networks to the phones at the beginning of the museum visit.

This state of development provides the basis for this dissertation.

1.4 Publications

Over the last 6 years of research, different parts of our PhoneGuide project have been presented at various conferences and published in international journals. In the following list, the highlighted entries were published during the course of the Ph.D.:

Journal Papers

E. Bruns and O. Bimber. Localization and Classification through Adaptive Pathway Analysis. IEEE Pervasive Computing, minor revision, April 2010

E. Bruns and O. Bimber. Adaptive Training of Video Sets for Image Recognition on Mobile Phones. Journal of Personal and Ubiquitous Computing, vol. 13, no. 2, pp. 165-178, February 2009

E. Bruns, B. Brombach and O. Bimber. Mobile Phone Enabled Museum Guidance with Adaptive Classification. IEEE Computer Graphics and Applications, vol. 28, no. 4, pp. 98-102, July-August 2008

E. Bruns, B. Brombach, T. Zeidler and O. Bimber. Enabling Mobile Phones To Support Large-Scale Museum Guidance. IEEE MultiMedia, vol. 14, no. 2, pp. 16-25, April 2007

Conference Papers

E. Bruns and O. Bimber. Mobile Museum Guidance through Relational Multi-Image Classification. Accepted at International Conference on Multimedia and Ubiquitous Engineering (MUE'10), 2010

E. Bruns and O. Bimber. Adaptation Techniques for Mobile Image Classification. Proceedings of Wireless Communication and Information (WCI'09), pp. 235-247, 2009

B. Brombach, E. Bruns and O. Bimber. Subobject Detection through Spatial Relationships on Mobile Phones. Proceedings of International Conference of Intelligent User Interfaces (IUI'09), pp. 267-276, 2009

E. Bruns and O. Bimber. Phone-to-Phone Communication for Adaptive Image Classification. Proceedings of International Conference on Advances in Mobile Computing & Multimedia (MoMM'08), pp. 276-281, 2008

P. Föckler, T. Zeidler, B. Brombach, E. Bruns and O. Bimber. PhoneGuide: Museum Guidance Supported by On-Device Object Recognition on Mobile Phones. Proceedings of International Conference on Mobile and Ubiquitous Computing (MUM'05), pp. 3-10, 2005

The *PhoneGuide* project was supported by the Stiftung für Technologie, Innovation und Forschung Thüringen (STIFT) and Nokia.

1.5 Outline

In chapter 2 we discuss existing related work. First, we describe approaches which are related to the entire Phoneguide system. These can be separated into location- and vision-based mobile information systems and several related work approaches are presented. Second, we provide an overview of approaches that relate to specific research elements of this thesis such as wireless ad-hoc networking or person movement prediction.

In chapter 3 we provide further background information on our mobile museum guidance system and outline the approaches that were developed prior to this thesis in the context of the PhoneGuide project. In addition, preliminaries are given on this thesis.

In chapter 4 we demonstrate how capturing multiple images of the same object leads to higher object recognition rates by employing combined image classification. A user study reveals the benefits of this approach.

In chapter 5 we explain how the spatial relationships of multiple objects can be utilized to improve and accelerate classification performance. This technique is compared with related brute-force methods and the results of a user study are presented.

In chapter 6 we show how the seamless exchange of classification data among mobile phones via Bluetooth can improve classification performance. A simulation with real world data reveals the advantages of this approach.

In chapter 7 we present a person movement prediction method that takes into account the spatial co-locations of visited objects for learning and prediction. Based on the history of already visited exhibits, the algorithm determines future locations and the predictions are integrated into the object classification process. Furthermore we explain how physical locations are detected that visitors do not return to after they have left them in order to reduce the number of possible result candidates.

Finally, the results of a user study are presented.

In chapter 8 we provide an overview of our system architecture as well as of our entire user interface. We then show how the different adaptation techniques presented in this thesis influence the classification rate. Finally, we present a summary of all approaches and outline their pros and cons as well as their potential for transferability.

In chapter 9 we present our conclusion and offer an insight into potential future work areas.

2 Background and Related Work

This work is composed of several elements that originate from various different research fields. For example, mobile image classification has its roots in object recognition, respectively computer vision. On the other hand, the overall paradigm of user-driven interaction with the real world by means of mobile devices originates from context-aware mobile computing. In the following, we will try to give an overview of the different research fields and their taxonomy.

In general, ”context-aware computing is a mobile computing paradigm in which applications can discover and take advantage of contextual information” such as user location, time of day, nearby people and devices and user activity [40]. The term ”context-awareness” is broadly used in many different fields of research. It originates from the area of ubiquitous computing established by Mark Weiser [191] and was first introduced by Schilit et al. [130]. It describes software that ”[...] adapts according to the location of use, the collection of nearby people, hosts, and accessible devices, as well as to changes to such things over time. A system with these capabilities can examine the computing environment and react to changes to the environment.” A general definition of context was provided by Dey et al. [51] who defined context as ”any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.”

In this work we define ”context-aware mobile information systems”[5] as all kinds of applications that utilize techniques for identifying (parts of) the user’s environment with a mobile device. Based on this knowledge, information and/or services are provided which relate to the user’s context.

The approaches that we present in this thesis were designed for realizing a mobile museum guidance application that allows visitors to identify exhibits by taking a picture of them. Besides object recognition algorithms, we present different adaptation techniques which support and improve classification performance. This kind of application can be categorized as context-aware mobile information systems since it characterizes a system which provide information about objects in museums that are co-located to the user.

The simplest approach to retrieving information about things in the environment in general, is to identify them manually. In such cases, for example, users type reference numbers or URLs that are attached to an object into a mobile device (e.g., audio guides, multimedia guides, mobile

[5] We have used the abbreviation ”information systems” throughout this thesis.

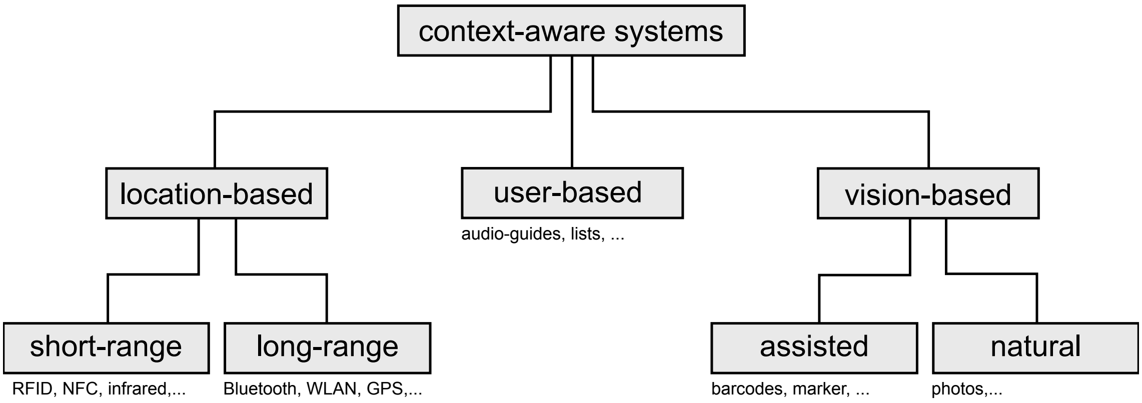


Figure 2.1: Overview of the main techniques for realizing context-aware mobile information systems such as mobile museum guides.

phones) to obtain further information. In a museum environment, users can also browse through a(n) (electronic) list of labeled exhibits [25, 79] to find information corresponding to an object. Such user-mediated approaches have been commercially available for a long time.

More sophisticated information systems that facilitate a more user-friendly and smarter interaction with the environment can be separated into two main research areas: We call the first area ”location-based information systems” that apply localization (and orientation) techniques to determine the location and thus the environment of users. Based on this data, information corresponding to the surrounding objects is presented. Significant related work in this field is described in section 2.1.
The second area comprises ”vision-based information systems” which identify the context using image processing techniques. This area is the most related field to this work and will be investigated further in section 2.2. An overview of the main techniques and their classification is illustrated in figure 2.1.
A separation of context-aware computing from an interactive point of view is proposed by Rukzio et al. [152].

In addition to these main areas, parts of this thesis originated from various independent research disciplines including object recognition, wireless ad-hoc networks and person movement prediction which are described and explained in section 2.3.

2.1 Location-Based Information Systems

According to our definition, location-based information systems allow people to retrieve information or services concerning nearby objects by using their mobile devices. To provide such information the locations of the users is determined by using localization or wireless communication technologies.

There are two different ways of describing the location of users: On the one hand, a location can be absolute. This means that the position of the user is defined by global geographical coordinates determined using positioning systems such as GPS. On the other hand locations can be relative, which means that the position is measured based on distributed sensors. These sensors are typically wireless communication transmitters or receivers such as RFID, Bluetooth or WLAN technology (an overview of the various hardware techniques can be found in [163, 127]) that are distributed at known locations. These sensors then continuously broadcast identification data that is received by the users' mobile devices. The devices process this data and determine their relative location to the emitter. It is worth noting that the communication technologies were not intentionally designed for this task. They are misused for localization which can result in practical challenges as we will discuss later. These local positioning systems [101], also known as indoor positioning systems, are used especially inside buildings since global positioning techniques perform less well indoors.

To provide a better overview, we have divided the local positioning systems into two categories based on the hardware used and the corresponding complexity of the localization algorithms: Short-range localization techniques use sensors with a signal range of up to few meters so the location of a user is determined by the emitter they receive a signal from. Long-range localization methods have a signal range or resolution of up to several hundred meters. Therefore, additional algorithmic refinements are necessary to retrieve the precise location information.

The amount of related work in both categories is tremendous and several surveys on location-based information systems (independent of the localization technique used) have been published in the past [40, 85, 20, 157, 122]. In the following we will characterize and describe the differences of both categories.

2.1.1 Short-Range Localization

Short-range emitters like Radio Frequency Identification (RFID) [64], Near Field Communication (NFC) [69] or infrared [10] have a signal range of only a few meters. Of course, advanced RFIDs or infrared technologies exist that achieve much higher signal ranges but in this section we shall concentrate on approaches that apply this sort of technology to support short-range communication only.

To determine the location of a user through the application of short-range emitters, the sensors are attached to or close to (e.g., on the ceiling) an object of interest. When a receiver device recognizes the signal emitted by a sensor, the corresponding user can be located as within signal range and thus close to the object. This sort of location technique is straightforward and is known as ”cell-based localization” [45] because no further algorithmic refinement or processing of the signal is carried out: the location is determined solely as a binary indication that the signal of an emitter was received or not. The area that is covered by the corresponding signal range is called ”cell”. In the case of short-range emitters this technique is reasonable since the cells are small enough to precisely locate the visitor. Where long-range emitters are used, this method becomes fairly imprecise as we will see later. Since the location of the user is only related to the specific object one can alternatively argue that the device just identifies the corresponding object instead of determining the device’s position. However, for consistency we also call this localization.

In the following we will give a short description of each hardware technology and afterwards describe significant research work that uses these techniques for context-aware applications.

Radio Frequency Identification (RFID) systems are made up of a transponder (also known as a tag) and a reader that communicates with one or multiple transponders. The tags can be either passive or active. Passive transponders have no battery and retrieve their energy from the electromagnetic waves of a reader and consequently have a lower signal range than active tags. Active transponders have their own energy source which increases their size and cost and reduces their lifetime [187]. RFID systems can operate on various different frequencies ranging from 135 kHz to 5.8 GHz and achieve signal ranges of between a few millimeters and up to 15 meters [64].

Huang et al. [89], for instance, introduced a PDA- and RFID-based museum guide in combination with a recommendation system. With this system, visitors identify exhibits by scanning for RFID tags attached to or near objects to call up corresponding multimedia information from a remote server via Bluetooth or WLAN. In addition the paths of visitors are tracked and matched against a database using collaborative filtering to recommend related exhibits that may be of interest to the current visitor. Rashid et al. [144, 143] designed a real world Pacman game called Paclan. RFID equipped plastic discs are distributed around a large outdoor area. If a user activates such a disc with a RFID reader attached to his mobile phone the current location is identified and sent to a

remote server via GPRS in combination with Pacman related game information such as a kill timer, etc. The authors chose to use RFIDs because GPS was too inaccurate and slow if the signal was lost as a result of interferences caused by buildings.
Chon et al. [43] installed RFID tags along roads to improve the accuracy of automotive navigation systems. The precisely known location of the RFID tag is combined with GPS information to refine the location data, especially in challenging areas such as tunnels or downtown areas. In addition to providing location information, these tags can provide further location-based services such as the presence of nearby museums or restaurants which are stored in a permanently updated transponder database. Chang et al. [38] designed a RFID-based indoor guidance system. With this system, the location and navigation details are displayed on a PDA that determines its position based on distributed RFID tags. Välkkynen et al. [182] used sensor devices called SoapBoxes [173] which host several integrated sensors. In addition to scanning for RFID chips in the vicinity (which are embedded in each SoapBox), a SoapBox can be triggered by a mobile device through infrared or laser light (via an integrated light sensor). A wireless connection between the corresponding SoapBox and the mobile device is established and context data, such as URLs are transmitted.

Infrared-based communication makes it possible to transmit electronic data via infrared light. Since it operates on a similar wavelength it shares many of the properties of visible light. For example, it cannot pass through walls or other obstacles. In contrast to radio-frequency techniques, infrared techniques require a direct line of sight between emitter and receiver [16].

One of the first localization systems to be implemented was the active badge system [188]. People in office buildings wear infrared beacons which are recognized by infrared readers mounted on the wall or the ceiling. The location data for all the users is collected by a centralized server which provides an interface for retrieving the location of specific persons. Comparable approaches (infrared beacons on the ceiling) were introduced by Ward et al. [189] to accurately track the position and orientation of people in an Augmented Reality environment or by Bederson et al. [23] to establish an Augmented Reality museum guide. Furthermore, Hiyama et al. [86] introduced a localization approach in a museum environment by attaching infrared transmitters to the ceiling of an exhibition hall. The transmitter broadcasted a unique ID to encode the current location of visitors and present related information about the exhibits.
Another well-known indoor location-based guidance system based on infrared technology is called Cyberguide [3]. Simple maps with outlines of buildings and context information are displayed on a hand-held device. Beside this outdoor application, they applied infrared beacons to estimate the

device's position within buildings. Oppermann et al. attached infrared beacons directly to exhibits in a museum to identify them [132].

Near field communication (NFC) is an extension of RFID that operates at 13.56 MHz and transfers data within several centimeters at up to 424 Kbits/second [136]. In contrast to RFID, NFC allows for bi-directional device to device transfer [8].

Rudametkin et al. [150], for instance, used NFC technology to design an open-source middleware for identifying museum exhibits. The tags are attached to paintings which transmit a unique ID to an NFC equipped mobile phone if it is located close to the exhibit. The mobile phone then sends the identification data to a remote server via Bluetooth which in turn plays object related multimedia information on a nearby display. Rukzio et al. [59] introduced a framework that utilizes different interaction techniques such as visual markers, Bluetooth, RFID or NFC to interact with the physical environment and discusses the requirements for implementing such a system.

All of these systems discussed above determine the location of users relative to one object. This is either accomplished by attaching a short-range emitter directly to an object, i.e., the object is selected directly by moving a mobile reader close to it, or beacons are distributed in the vicinity of an exhibit (e.g., mounted at the ceiling). If such a beacon is detected by a device of an approaching visitor, the location is defined by the closest entity in the environment (e.g., an exhibit in a museum). Consequently, appropriate information can be presented about it.

From our point of view, such techniques produce several practical challenges and drawbacks that arise when they are used to identify a single object, e.g., as part of a mobile guidance system in museums. For example, museums have to bear the hardware costs to equip their building area-wide with RF-technology which requires additional maintenance by appropriately trained staff. Furthermore, if many densely arranged objects are located in a small area in the museum (e.g., in showcases), it is not possible to reliably select distinct objects because of the coarse accuracy of the localization techniques. This means that RF-signals from several emitters attached to different objects are simultaneously detected by one reader and these signals cannot be separated. As a result, such approaches have to display a list of possible exhibits in the vicinity of the user who then has to additionally manually select the final object.

In contrast to this, our approach makes it possible to directly select a physical object by taking a picture of it. We believe that this is an intuitive, economic and effective technique for selecting and interacting with physical objects such as exhibits in a museum in order to retrieve related information.

2.1.2 Long-Range Localization

Long-range localization systems for indoor environments apply RF-hardware such as Bluetooth or WLAN with signal ranges of up to 100 meters. In comparison to short-range emitters, the positioning is therefore much more imprecise if simple cell-based approaches are used, because the signal area covered by one emitter is much larger. This has led to the development of various algorithms in the past designed to increase the localization accuracy.

As with the previous section we will provide a short description of each hardware technology first, followed by an account of significant research work that applies these techniques for context-aware applications.

WLAN comprises a family of radio protocols (e.g., 802.11 protocols also known as Wi-Fi) for building low-cost wireless networks [66]. The majority of protocols operate in a frequency band of 2,400 to 2,485 GHz and achieve transfer rates of between 11 (802.11b) and 600 MBit/s (802.11n). The signal range heavily depends on the power of the emitter, the receiver, the application area (obstacles), etc. On average, the signal range of standard WLAN devices varies between 30m to 90m indoors [172].

Comparable to the related work that uses short-range emitters for determining the position of a user on a signal cell level, several approaches also use this technique but based on WLAN technology. For instance, Cheverst et al. [42] introduced a city guidance system on a tablet PC that combines user feedback with location information. For recognizing objects, users have to manually provide a rough indication of how far they are away. The location information is then estimated by detecting nearby WLAN hotspots (beacons). Kim et al. [98] proposed a method for discovering semantically meaningful places (”colloquially labeled representations of locations such as ’Home’, ’My Office’,...” [98]) by scanning for WLAN beacons to determine the rough location of people indoors.

However, due to the high coverage of a single WLAN beacon, the cell-based method is fairly imprecise especially where several objects in the immediate vicinity of people need to be identified. Accordingly, many research projects utilize refinement techniques to enhance the localization methods. Mainly, they utilize the signal strength of one or multiple RF-emitters to deduce the current

position of users more accurately. An overview of different signal-strength-based techniques can be found in [110].
One popular refinement approach for indoor localization based on signal strength (independent of the hardware used) is called ”fingerprinting” [45]. In these cases, the signal strength of one or multiple base stations is determined for each possible location in an indoor environment and used as a fingerprint to derive the current location of users. For instance, Bahl et al. [12] took measurements of signal strength from 3 different WLAN stations for 70 distinct physical locations and 4 directions (north, south, east, west) in an office area of 980 m^2. The current location is then retrieved by determining the signal strength of the available WLAN beacons and by finding the k-nearest neighbors in the fingerprinting database stored during the calibration phase. With this technique they reported an accuracy of 2-3 m. Comparable approaches that use fingerprinting techniques based on WLAN beacons can be found in [18, 110, 97, 27].

Bluetooth (BT) is a wireless RF-technology that operates on the same frequency as WLAN (2.4 GHz) [88]. The data transfer rate varies between 2.1 Mbit/s (Bluetooth 2.0) and 24 Mbit/s (Bluetooth 3.0). The signal range depends on the power class and range between 1 m (class 1) and 100 m (class 3). However, almost all mobile devices such as mobile phones, laptops and headsets that utilize Bluetooth are class 2 devices which achieve a maximal signal range of 10 m. Due to the high market coverage of Bluetooth-enabled devices, Bluetooth has become popular not only as a ubiquitous data transfer method but also as a localization technology.

Bruno et al. [29] proposed a cell-based localization approach by distributing BT-emitters inside a building where each room represents one cell. People in the building carry BT-equipped mobile devices. The BT-beacons either scan periodically for these devices or establish a connection to them until the connection is aborted (because the user moves out of the signal range) to determine the current cell of the user. Another cell-based approach was introduced by Bay et al. [21] where BT-emitters are distributed in a museum environment. In contrast to Bruno et al. [29], the mobile device scans for BT-emitters in the vicinity to determine its approximate location. To increase the accuracy of cell-based methods, many approaches distribute multiple BT-emitters close to each other to create several overlapping RF-regions to refine the localization.
Filho et al. [63], for instance, distributed multiple BT-beacons in a building and constructed a probability map that indicates how often each BT-node is detected at certain locations. The position during the localization process is then derived by a decision rule based on the probabilities of the observed BT-nodes. A strategy to efficiently place such BT-beacons in indoor environments for

optimal coverage and localization is proposed in [39].

Wendtland et al. [192] introduced a fingerprinting technique based on signal strength of Bluetooth. In an office building they distribute several BT-beacons in a room to determine the current location of users carrying a mobile phone. The position is derived by combining the probability density functions of each received BT-beacon. The probability density function represents the probability that a BT-emitter is detected at a certain location. The accuracy depends on the number of BT-emitters used and was on average 2-3 m. Bargh et al. [15] also applied a fingerprinting technique. However, instead of using signal strength as the measure, they utilize the inquiry response rate. It is defined as the percentage of inquiry responses to total inquiries (in the context of Bluetooth, an inquiry describes the act of scanning for BT devices). The localization accuracy of this system was at room level. Kotanen et al. [103] utilized the signal strength to measure the distance between a mobile device and distributed BT-beacons in a building. With this system, they convert the signal strength values into power levels and use a radio wave propagation model to derive the distance. They achieved an accuracy of $\sim$4 m.

Besides WLAN and Bluetooth, which are the most common techniques for estimating user locations indoors, several other approaches exist that utilize additional hardware for localization. The well-known Cricket project [141], for instance, uses RF-technology in combination with ultrasound for localization: Beacons in the environment continuously transmit a RF as well as an ultrasound signal. Since the RF signal travels much fast than the ultrasound signal, a user's mobile device can derive its distance to the beacon by determining the time difference between the receipt of both signals (also known as ”time of arrival” technique). The location granularity of this system was 4x4 feet and several enhancements to this system were published. For example, Priyantha et al. [142] extended the system to determine the orientation of the mobile device. Piontek et al. [139] improved the localization accuracy. A comparable system was introduced by Harter et al. [82] which could determine the position of a mobile device within 9 cm by distributing 100 beacons to cover approximately 280 m^3.

Ravi et al. [145] proposed a localization technique based on fingerprinting that utilizes the intensity of light. With this system, users wear light sensors that measure the amount of incident light. Since each location in a room shows a slightly different illumination pattern, the location can be estimated. Prior to that, the same authors introduced a method based on image recognition to derive the location of users [146]. For this, people have to wear a smart phone as a pendant which continuously captures photos of the environment. These images are then sent to a remote server where they are compared with a database of images captured from known locations to compute the current position. Arth et al. [9] propose a system for estimating the full 6 degrees of freedom from wide-area 3D reconstructions based on natural image features. Here, an initial estimate of the position of the user has to be provided using an arbitrary localization technique.

GSM (Global System for Mobile Communications) can also be utilized for localization. The signal strength of multiple receiving cell-towers are recorded to compute the position based on trilateration [45]. Of course, the fingerprinting technique based on GSM signals can also be applied for localization in outdoor environments [106] as well as indoors [176].
Furthermore, many approaches exist that utilize GPS (global positioning system) to provide context-aware information [149, 104]. This is straightforward outdoors, but cannot be applied indoors because the signal strength is too low to penetrate a building [40].

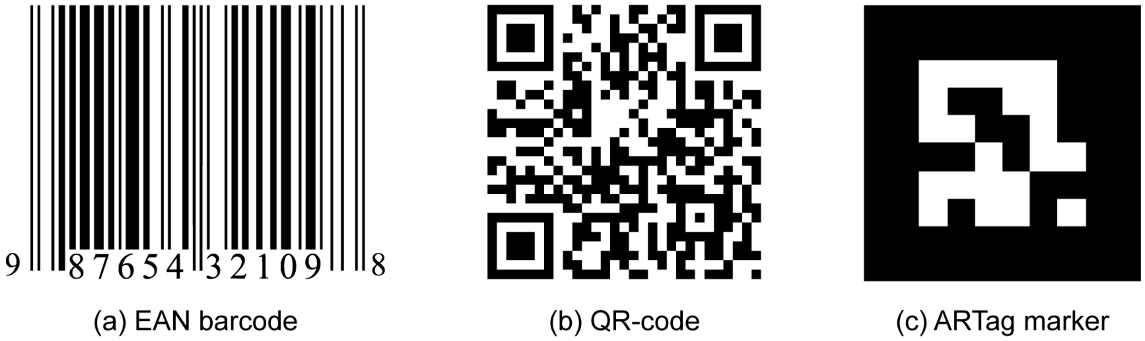
Finally, many researchers combine different localization techniques. For instance, they improve the accuracy by using a combination of RFID, Bluetooth, and WLAN [183], facilitating the seamless handover between indoor to outdoor environments for continuous localization [19, 80], or they combine localization with orientation information [135].

As already mentioned in the previous section, localization techniques based on RF technology have several drawbacks such as hardware and maintenance costs. Furthermore, the localization accuracy of long-range localization techniques can be much more inaccurate (unless a significant amount of specialized hardware is used) and error-prone (e.g., emitters can drop out) which leads to a less precise definition of the surrounding environment. Consequently, such techniques cannot provide unambiguous results of the object of interest and often return a selection of possible result candidates.

Again, we believe that taking pictures of objects is a more intuitive and effective technique for interacting with physical objects that overcomes several drawbacks of location-based information systems.

2.2 Vision-Based Information Systems

We define vision-based information systems as applications that allow users to retrieve information about objects. These objects are either identified by taking pictures directly of them or by capturing an image of a representative entity (e.g., a reference tag) with a mobile device. This kind of interaction technique is also known as ”pointing” [153, 152] or ”point and shoot” [13] as proposed



(a) EAN barcode

(b) QR-code

(c) ARTag marker

Figure 2.2: Selection of different barcodes.

by related work.

The variety of applications of such systems ranges from simple mobile applications that decode barcodes on goods to find the lowest price on the internet [125], to museum or city guidance systems that classify exhibits or buildings by photographing them directly. We have separated this research field into two main areas: The first area comprises approaches that identify objects by recognizing reference tags such as barcodes or fiduciary markers (fiducials). These tags are normally attached directly to the object or in its vicinity. We call these methods ”assisted vision-based techniques”. The second area comprises ”natural vision-based techniques” that identify the object directly by photographing it.

2.2.1 Assisted Vision-Based Techniques

Assisted vision-based techniques comprise all methods that use additional visual references to identify the object. Instead of the object, the reference tag that is attached (close) to the object, is photographed for classification.

The application of barcodes for classifying objects is straightforward. In general, barcodes are small labels which consist of black and white or colored patterns that encode identification information. The high contrast images allow for a unique and reliable identification through computer vision techniques. One dimensional markers such as EAN barcodes [76] or 2-dimensional markers such as QR-codes [190] are attached to objects (cf. figure 2.2). After taking a picture of the barcode the mobile application decodes it and retrieves either a unique ID (1D barcode) which identifies the object, or an URL (2D barcode) that links to further information.

Two dimensional barcodes are also used in the research field of Augmented Reality (AR) [11]. In

this field, they are called fiducials and serve as identifiers as well as a point of reference to augment the real world with virtual content. This technique has been applied in several museum guidance systems:
One of the first AR systems for museum guidance was introduced by Rekimoto et al. [148]. Color-coded markers are attached to paintings to identify them and to augment them with related information. The overlaid information is either presented by a head mounted display (HMD) or by a PDA. Schmalstieg [155] proposed an AR museum guide that uses markers not only for identifying objects and presenting context-sensitive information but also for storytelling and location-based games where the virtual content of physical objects has to be manipulated. Choudary et al. [44] attached colored markers to prehistoric cave engravings. By pointing a camera-equipped mobile phone towards a wall, the barely visible engravings are augmented by professional drawings in real-time. Comparable to this, Izkara et al. [94] designed a historical guidance system where AR-markers are attached to buildings for identification. A tool for creating AR applications in a museum context was introduced by Koleva et al. [100]. By utilizing AR-markers, museum experts could easily merge both physical and digital elements in a museum environment.

The drawback of these techniques is that they require reference markers which have to be located close to the objects. In our opinion, this distracts from the overall appearance of well-arranged exhibits in museums. Especially where many small and densely located objects exist, this technique is not optimal and visually clutters the exhibits.

2.2.2 Natural Vision-Based Techniques

In contrast to assisted vision-based approaches, mobile natural object recognition methods identify an object directly by photographing it. Research developments into these techniques can be divided into three conceptual areas as illustrated in figure 2.3. In the following, we will describe these three areas and present corresponding related work approaches.

The first context-aware systems to apply natural vision-based classification on mobile devices were based on a client-server architecture. The mobile device only serves as a front-end application to capture images of objects such as exhibits in museums or buildings. The photo is then transferred via wireless technology to a remote server where the image classification is carried out. Finally, the corresponding result is relayed back to the mobile device which presents the related multimedia content.

For instance, Fritz et al. [71, 116] introduced a city guide for mobile phones: datasets includ-

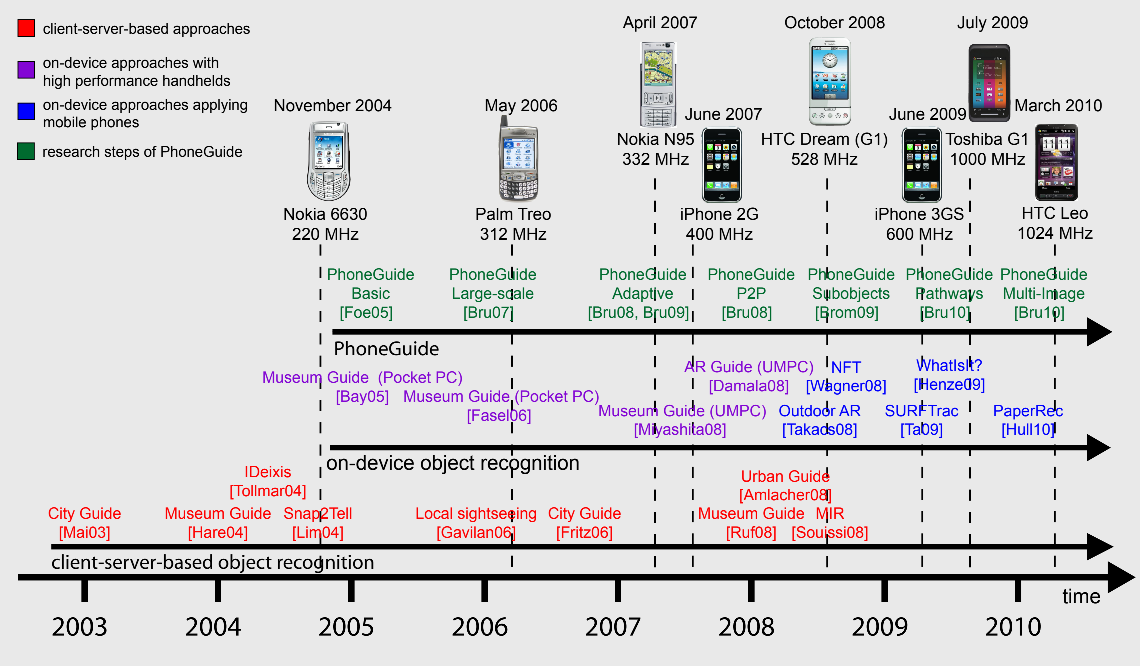


Figure 2.3: The three conceptual areas of mobile natural vision-based techniques: client-server based, on-device classification based using high-performance devices and on-device classification based using smart phones. In addition, release dates of popular mobile phones (with CPU clock rate) and the different research steps of the whole PhoneGuide project.

ing 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, SIFT features [115] (further information on the SIFT algorithm can be found in section 2.3.1) are extracted from the images and are compared with a database of image features of known sights. Finally, the corresponding multimedia data is sent back to the user's phone after the objects have been classified. Hare et al. [81] developed a museum guide for pocket PCs. Photographed images of paintings are transferred to a remote server to compute SIFT features. For classification, however, an adapted text retrieval technique is applied. Nonetheless, the recognition is comparable to that of Fritz et al. [71]. Albertini et al. [6] implemented a museum guide to recognize parts of a fresco. While visitors move a camera-equipped PDA over the fresco, images are captured and continuously transferred to a remote server. Based on color histograms, these images are classified and corresponding information is sent back to the mobile device. This approach later became part of the long-term PEACH project [165] for developing a novel integrated framework for museum visits.

Gavilan et al. [75] introduced an image retrieval method for mobile phones. A captured image

in combination with GPS coordinates is sent to a remote server to retrieve similar images. Here, the image is separated into segments via blob detection and features are extracted which describe these blobs such as the center of gravity, volume, color, etc. Through a nearest neighbor approach, related images from a database, which were separated into four categories, were retrieved. The GPS location of where each image was captured is stored in the database. During retrieval, all result candidates are neglected which are outside of a predefined radius of the GPS coordinates of the query image. A comparable approach based on image segmentation is proposed by Souissi et al. [164].
Lee et al. [108] presented an application to recognize landmarks with mobile phones. The captured image is classified on a remote server by extracting corner features [83] in combination with a distance voting scheme. Besides the image, the location as well as the orientation of the mobile device are transferred to the server. Consequently, only database images are considered during classification which hold similar location and orientation information. This significantly reduces the number of possible result candidates. A survey on mobile landmark recognition for information retrieval can be found in [41].

Ruf et al. [151] designed a museum guidance system that recognizes paintings on a server. For image description, they evaluated SIFT [115] and Speeded Up Robust Features (SURF) [22] (further information on SURF can be found in section 2.3.1) with different image resolutions. Besides a nearest neighbor approach based on Euclidean distance they implemented an approximated nearest neighbor method based on k-means clustering to improve the retrieval time. They revealed that SIFT outperforms the SURF algorithm for any resolution considered.
Lim et al. [112] introduce a client-server architecture where captured images in combination with the current location of the user are sent as MMS to a remote server. The location was either retrieved via triangulation based on a group of base stations or GPS. The information retrieval is carried out by computing local or global color histograms in HSV color space and applying histogram intersection [166] for classification. The appropriate multimedia content is then send back to the phone via MMS.
Mai et al. [118] introduce a system that makes it possible to identify buildings with a PDA. Images are captured using a mobile device and sent to a remote server in combination with the location (via GPS) and orientation (orientation sensor) of the user. By means of a predefined 3D city model the virtual viewport of the user is rendered based on the location and orientation. The captured image is then matched against the rendered image to unambiguously recognize the photographed building via Hough transform [53] and image segmentation. Further improvements to this approach can be

found in [119].
Doeller et al. [52] introduce a location-based tourism system for mobile phones. After an image was captured they extract different image features and store them in an MPEG-7 document. In order to save bandwidth this document, rather than the image itself, is then sent to a server for matching purposes. Tollmar [171] developed two different information retrieval techniques based on captured images of outdoor scenes. In the first approach, they display relevant webpages containing the matching images that were found using content-based image retrieval. In the second approach they carry out a Google search on the mobile device based on the keywords they found on the webpages of the first technique to retrieve further information. Both methods were compared to a traditional location-based system that utilizes GPS. A user study revealed that the second method performed best in terms of quickly finding appropriate information.

Beside these related approaches in research, several commercial products exist such as ”Nokia Point and Find” [140], ”Kooaba” [102], ”Snaptell” [162], LookTel [114] or ”Google Goggles” [77] that try to recognize selections of CD/DVD covers, movie posters, book covers, exhibits or sights by taking a picture of them.

All of these approaches apply a client-server architecture which has, from our point of view, several drawbacks in comparison to on-device solutions: First, they are centralized which means that the response time for classification requests depends on the hardware used and available bandwidth. Low-cost solutions for museums including appropriate server with consistent network coverage (e.g., through WLAN) are especially challenging. Second, in outdoor environments, users have to bear the connections costs if images have to be transferred via MMS or through a web interface which we wanted to avoid. Finally, the wireless transfer of image or feature data is energy consuming which leads to shorter usage times compared with on-device approaches.

The second area of vision-based information systems comprises approaches that apply high performance hardware to carry out the image classification directly on the mobile device. Consequently, no connection to a server is necessary. This area can be seen as an intermediate step between client-server based approaches and object recognition carried out on mobile phones.

Bay et al. [22], for example, introduced a museum guide based on a tablet PC. The identification is performed directly on the device so no server communication is established. For image classification they apply SURF [22]. This approach was further improved by Basel et al. [61] by applying Gaussian image intensity attenuation and a foveation-based preprocessing approach in order to make it possible to focus the interest point extraction towards the center of an image. In earlier work, Bay et al. [21] distributed Bluetooth emitters to determine the users' locations and consequently narrow down the set of possible result objects. Amlacher et al. [7] utilizes a Ultra-Mobile PC (UMPC) to classify images via i-SIFT [71] in a large urban area. To reduce the visual search space, they apply GPS to determine the rough location of the user. The distances within a predefined local spatial neighborhood around the user were then weighted using an exponential function. The resulting geographical distributions were combined with the distributions of the image classification to weight objects that are located closer to the user more strongly. Miyashita et al. [124] introduce an Augmented Reality museum guide that applies a UMPC in combination with a hybrid markerless tracking technique. Context-related information is presented using a magic lens setup where the visitors use the mobile PC as a digital window to see augmentations superimposed over real world exhibits.

Finally, the third area of vision-based techniques applies mobile phones for image classification. The main difference compared with the devices used in the previous area is that mobile phones unify all necessary hardware components such as a camera, GPS and RF-communication into one handy device for context-aware systems. In addition, mobile phones are much more widespread than, for example, UMPCs. Consequently, from an economic point of view, the step towards the third development stage was important because they made context-aware applications publicly available. From an algorithmic point of view, however, the differences between those two areas are diminishing as the performance of mobile phones continues to improve (cf. figure 2.3).

Takacs et al. [169], for example, implemented a performance-improved version of SURF [22] on a mobile phone for outdoor Augmented Reality applications. To increase the classification rate they consider only database images that were captured in the vicinity of the user. Henze et al. [84] presented a technique for recognizing poster segments with a mobile phone. They applied an adapted SIFT [115] algorithm for feature extraction and a vocabulary tree for classification. A comparable approach is presented by Hull et al. [90] that recognizes pages in newspapers. Ta et al. [167] introduced a tracking and continuous object recognition system for mobile phones. They perform object recognition in a similar way to Takacs et al. [169] to classify the first frame of a video stream. For consecutive frames they only compute the interest points and estimate the corresponding camera pose based on previous images. A comparable approach is introduced by Wagner et al. [184] that allows natural feature tracking on mobile phones based on SIFT and Ferns [198]. Although the emphasis of the last two methods lies on real-time tracking, such techniques could also be utilized for image classification.

All of these approaches, independent of the architecture used (client-server based or on-device) make only little use of additional adaptation techniques to support and improve image classification, although the classification results of the presented methods show room for improvement [81, 151, 52, 169]. Many approaches use only location information to reduce the number of possible result candidates by considering exclusively the objects that are in the vicinity of the user. This is either carried out in a weighted manner where the distance between user and potential real-world result candidates is analyzed [7], or unweighted where each object in the vicinity is considered equally [71, 116, 75, 164, 112, 21, 169, 108, 118]. In addition, some techniques add orientation information from the mobile device to further decrease the image search space [108, 118].

However, to the best of our knowledge, no related work exists that applies additional constraints which take advantage of the unique characteristics of mobile phones as proposed in this work. For example, no approach makes use of the ability to collect the captured images in combination with user feedback in order to continuously retrain and adapt the image classifiers over time. Furthermore, no related work carries out this adaptation on-the-fly by automatically exchanging classification data among the mobile phones. In addition, they do not support the ability to capture multiple images with mobile devices to automatically conduct combined image classification. Finally, none of the related work approaches determine the individual movement patterns of users to predict objects that will be photographed next. This enhances image classification and compensates for the drawbacks of traditional localization techniques such as poor GPS readings, RF-signal noise or high energy consumption.

2.3 Related Research Areas

Besides the main related research area of context-aware systems, parts of this work rely on several different research fields. In the following subsections we will detail significant research results from the following research areas: In section 2.3.1 we provide a basic overview of the field of object recognition and outline approaches which relate to our object recognition (chapter 3), subobject recognition (chapter 5) as well as multi-image classification (chapter 4) methods. In section 2.3.2 we discuss different approaches towards working with wireless ad-hoc networks and illustrate how this relates to our phone-to-phone communication approach (chapter 6). Finally, in section 2.3.3 we present several techniques that analyze person movement patterns in order to predict the future locations of users which relates to our pathway awareness method described in chapter 7.

2.3.1 Object Recognition

Object recognition is part of the research area of computer vision that aims to create machines that see and understand the perceived structure [28]. A full discussion of all the different areas of computer vision is beyond the scope of this work. Instead we will focus on providing a basic overview of how photographed objects can be recognized in an image and outline the techniques which were applied by related work approaches mentioned in section 2.2.

The most important part of an object recognition algorithm is a precise and unique description of digitally captured images of objects. Therefore, numerical representations of the image, called features, are computed which ensure a distinct, comparable and compact representation of the photographed object or the entire image. In general, the variety of different features can be separated into global and local features as explained in [113]. Global image features which encode color distribution, shape or texture describe the whole picture which means that the image description can be formulated as one feature vector. The advantages of such global features are the compact representation of images as well as their fast computation. A drawback is their sensitivity to occlusion and clutter.

Local features by contrast do not describe the entire image but only image patches or distinct image points. They are more robust with regard to occlusion or clutter but require specific classifiers since the number of feature vectors for each image varies. In addition they are generally computationally more expensive and need more memory and storage which is particularly challenging for mobile applications.

A popular local feature extraction technique was introduced by Lowe et al. [115] in 1999 called scale-invariant feature transform (SIFT). Distinct keypoints in the image are detected by convolving the image with a Gaussian filter at different scales and computing the difference of successive Gaussian-blurred images. Minima or maxima of the difference of Gaussians at all scales finally define the location of the keypoints. Each keypoint which does not lie on an edge is then described by a histogram (descriptor). This descriptor is computed from the magnitude and the orientation of a predefined region (16x16 pixels) around the keypoint. A straightforward classification is carried out by applying a k-nearest neighbor search based on Euclidean distance to find the best match in a pre-collected database. Since this is computationally expensive, it can be substituted by an approximated nearest neighbor search, such as the Best-Bin-First algorithm [24] proposed by the same authors, or enhanced, e.g., by a vocabulary tree [129].

Due to its invariance to image scale and rotation and its robustness against occlusion, noise and illumination changes, the application of the algorithm became popular. This algorithm was applied particularly frequently in the field of client-server based object recognition approaches as we have shown in section 2.2. However, since the feature computation is expensive, several performance-improved versions of SIFT have been developed. One of them is called SURF (Speeded Up Robust Features) [22] that applies box filters in combination with integral images instead of Difference of Gaussians for image convolution. Since this enhances the feature computation, initial research approaches have emerged that implement adapted versions of this algorithm on mobile devices, e.g., [169].

As we will explain in chapter 3 we apply global color features such as histograms as well as mean and variance in the color channels [30] to describe the captured photo. Although color histograms are fairly simple image features, they compete well in many areas against more sophisticated algorithms [50] and they are especially useful if they are applied on low-performance mobile phones. However, since the subject of this work relies mainly on the adaptation techniques to improve the object recognition rather than on the applied image features, we believe that we can exchange the basic object recognition algorithm, if necessary, without losing the benefit of our adaptation techniques as further discussed in section 8.4.2.

In chapter 4 we propose a method of capturing multiple pictures of a single object for combined image classification. In content-based image retrieval (CBIR) the application of multiple query images has been examined by different researchers in the past. Tahaghogi et al. [168] tested three different color spaces (LAB, LUV, YUV) for feature extraction and two different distance measures (Manhattan and Euclidean) for their multi-image classification technique. For each picture, one database query is performed and the results were combined by either a sum, minimum or maximum function. Their experiments revealed that adding a second query image increases the retrieval performance by 9-20% depending on the selected approach. Applying more than three images did not improve the retrieval quality any further. The latter conclusion was picked up by Jin et al. [95] but could not be validated. They describe a CBIR system where visually different but semantically similar images are clustered by providing multiple pictures using relevance feedback (the method of rating results of a query to perform a new query in order to continuously refine the outcome; an overview of relevance feedback techniques can be found in [197]). Related work that applies relevance feedback can also be viewed as multi-image classification technique. However, in these systems users have to manually preselect an intermediate result to refine the final outcome. This involves greater user effort which we would like to avoid. Basak et al. [17] designed a system for facial image retrieval where features of multiple query instances are combined to find related database entries. Iskandar et al. [93] developed a CBIR system that applies an image, an image region or multiple image regions to carry out a query. They have shown that utilizing multiple regions outperforms the application of a single region or the whole picture in terms of retrieval performance.

Although all of these approaches use multiple images for classification, none of these techniques consider relations among the images as we did, to improve the classification performance and speed. In addition, the area of mobile object recognition is much more challenging since the computational performance of the devices is limited and the images are photographed on-the-fly by inexperienced users. This frequently results in images of low recording quality as well as redundant data.

In chapter 5 we describe a technique to detect multiple objects (called subobjects) in a single image based on their spatial relationships. In general, 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 depending on their application. For instance, topological relations [55] distinguish the relationship between two objects by analyzing the intersections of their boundaries and interiors (e.g., occluded, partly occluded, or disjunct). A further example is illustrated by directional relations [134] which are described by directional attributes such as north, west, south-east, etc.

Spatial relationships, however, are not only used for separating individual objects but also to describe different parts within a single object. Pham et al. [137] 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 the 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. Due to higher flexibility with respect to object distortion, such a detector configuration is able to achieve a higher recognition rate than a single fixed template-based detector. Agarwal et al. [4] presented an approach for detecting object classes in images using a sparse, part-based representation. The parts are extracted automatically from gray-scaled sample images by applying the Förstner interest operator [68] and by separating small image patches around the interest points which serve as ”vocabulary” for the object class. A fixed number of possible spatial relationships (consisting of a combination of five discretized distance bins and four direction ranges) are defined between two different parts.

Each training image is represented as a binary feature vector containing one value for each single part occurrence and each relationship occurrence between any two parts. Therefore, the potential number of features is high but only a limited number of features will be non-zero. To detect an instance from the object class, a fixed-size window is moved over the image (single-scale) or image scale pyramid (multi-scale). For each window, a feature vector is computed by testing the vocabulary (using computationally intensive normalized correlation) and by extracting the spatial relationships between all detected patches. Although the results are promising, this approach is too time-consuming to be carried out on current mobile phones effectively.

Spatial relationships are also utilized to generate a spatial orientation graph [58]. 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 [193], 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 of specific distances and angles. The initial location of the faces must be known. In [123], spatial relationships verify the classification of regions (e.g., sky, tree, street) after an image segmentation. In a post-processing step, all classified regions are 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. Accordingly, 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 faster subobject detection and reduces misclassifications from the beginning. The more spatial relationships have been found, the more robust the classification becomes.

2.3.2 Wireless Ad-hoc Networks

In chapter 6 we present an approach for exchanging classification data among the mobile phones of museum visitors to continuously adapt the image classifier over time. To ensure this, a wireless ad-hoc network has to be established between the mobile devices. In the following we describe related approaches in the area of wireless ad-hoc networks as well as distributed classification.

A popular example for wireless ad-hoc networks is car-to-car (inter-vehicle) communication. This research field investigates the exchange of dynamic traffic and other information between moving vehicles using WLAN signals and GPS tracking [128, 57, 60]. In addition, several applications for mobile devices exist that utilizes wireless technologies, such as Bluetooth, for ad-hoc communication.

Aalto et al. [2], for instance, propose a push-based advertising system that detects other Bluetooth devices in the proximity of a stationary Bluetooth sensor. The sensor retrieves the address of the mobile devices and forwards it to an advertisement server. This server then sends location-aware advertising directly to the mobile device via WAP.

A similar system called BlueTorrent, has been developed by Jung et al. [96]. It supports peer-to-peer (P2P) file sharing based on Bluetooth-enabled devices. Instead of downloading a whole file from a Bluetooth access point, the data is separated into blocks that are broadcasted to nearby pedestrians. Each block is then transferred further from user to user while moving to different locations. This enables cooperative file downloads (e.g., movie trailers) of up to 10 Mbytes.

Murakami et al. [126] implemented a simplified version of the hierarchical ad-hoc on-demand distance vector method for efficient data routing in large mobile phone networks. Based on Bluetooth, their system addresses applications such as mobile games, tracking applications and mobile emergency systems. Wang et al. [186] categorize games with respect to the way game engines are updated in mobile ad-hoc networks (e.g., asynchron, synchron, real-time, etc.). They also discuss challenges that arise when developing mobile P2P games in Java ME using the available Bluetooth API. Related issues are described in [147].

In general, our approach is similar to BlueTorrent [96] except that the ad-hoc communication we propose is combined with adaptive image classification. Instead of supporting cooperative file downloads, we focus on increasing the recognition rate and on reacting to dynamic changes quickly in a cooperative manner. Similar strategies can be found in the area of distributed classification techniques.

The goal of distributed classification approaches is to outsource the complex classification task to multiple networked nodes. Each node performs a sub-classification that is merged into the final result.

Besides various approaches that are based on static networks, Luo et al. [117] introduce a distributed classification method for P2P networks. In contrast to client-server systems, their method is scalable by adding new nodes. Each node builds its local classifiers based on a modified version of the pasting bites method, while the results of all classifiers are combined by plurality voting.

Wolff et al. [195] executes a sequential association rule mining (ARM) algorithm on local databases of each node in a P2P network. Each node then participates in distributed majority voting to retrieve the combined result of all connected peers. Siersdorfer et al. [160], as another example, describes a method for distributed document classification in P2P networks. Each node generates classifiers for locally stored documents, which are propagated through the network to help other nodes in improving their own classifiers.

This last example is similar to our approach. Each node adapts and improves its local classifier through information collected by and retrieved from other nodes. The main difference to conventional distributed classification systems, however, is that in our case the network connections are highly dynamic and synchronizations are based on ad-hoc communication among multiple phones that are located within signal range.

2.3.3 Person Movement Prediction

In chapter 7 we propose a method to predict the future locations of museum visitors based on the history of already visited exhibits in order to improve the overall classification performance. This procedure originates from the research area of person movement prediction.

A well-known work on person movement prediction is the MavHome project [48] in which a house is designed as an acting rational agent. The agent tries to predict the mobility patterns and device usages of its inhabitants in order to maximize their comfort and minimize operation costs. Movements are encoded by a string of symbols that represent zones in the house. A predictor is modeled by creating a dictionary of corresponding zone IDs. The location prediction of the inhabitants is carried out by traversing a trie that contains the dictionary's phrases and its probabilities.
Kubach et al. [105] presented a technique to predict the information that a user will need in future while using a location information system before it is accessed (also known as ”hoarding”). It is based on the assumptions that mobile devices can store only a certain amount of data and that this data can only be updated at certain locations (WLAN infostations) in a city. If a user enters the signal range area of such an infostation, he uploads the history (table) of the locations he visited and where he accessed information. Tables of multiple visitors as well as external knowledge (e.g., some locations are tagged as not accessible) are combined to derive visiting probabilities for each location. These probabilities are used to predict for each location the probability that the user will visit it in future. Consequently, only information for the top ranked locations are transferred. In contrast to our work, this approach evaluated its hoarding mechanism only in a simulation. In addition its performance relies mainly on predefined external knowledge to predict the future locations

of visitors. In our approach, everything relies on real pathway data that is used for training a classifier and we show that this can be employed to enhance mobile image classification.
Fetter et al. [62] propose an approach that tries to predict the location of a user at a given time. For this they continuously collect the GPS coordinates and the duration of stay for users in outdoor environments. Through clustering, they determine places where visitors stay for a longer period of time and predict future places based on a Hidden Markov Model. A comparable technique based on Bluetooth tracking is presented by Eagle et al. [54]. They mainly concentrate on finding relationships among users, social patterns or socially significant places based on location analysis.
Vintan et al. [180] introduced an approach for predicting person movement by applying neural networks. The locations of people are encoded as binary codes and serve as training and classification samples for the neural network. With this technique they could achieve a prediction rate of $\sim$87% but took only a small number of possible locations (< 16) into account. Sas et al. [154] also used a neural network for person movement prediction, but in a virtual environment. The users' trajectories are described by their positions, orientations and distances to the next landmark in virtual space. Based on these features, the network was trained to predict the next location with an accuracy of $\sim$68% within a predefined tolerance. In order to save bandwidth and resources when accessing wireless services, Akoush et al. [5] introduced an approach to predict the future locations of mobile users in wireless cellular networks. They apply a hybrid technique of Bayesian inference and artificial neural networks. As training data, they use network cell IDs, the cell history, as well as date and time. An overview of different approaches to user modeling can be found in [70].

In contrast to these approaches, we propose a method that describes pathways by the physical location of photographed object as well as by the objects that are located in the vicinity. The consideration of co-located objects allows us to train a 3-layer neural network that is able to predict future locations of visitors even if their past pathway has never been traced before. This is especially useful in large-scale public environments such as museums where only a small amount of pathway data is available in comparison to the amount of different locations. In these cases, related approaches fail as we will explain in detail in section 7.2.
Furthermore, to the best of our knowledge, we are the first to combine user movement prediction with mobile image classification to improve the overall classification performance of object recognition.

3 ADAPTIVE MOBILE MUSEUM GUIDANCE SYSTEM

As mentioned in the introduction, this thesis presents approaches for realizing adaptive image classification on mobile phones: besides object recognition algorithms, we present different adaptation techniques which support and improve classification performance. The chosen practical application for this object recognition system is a mobile museum guide. In the following we will explain our motivation for this selection and give further information on the background of this thesis.

3.1 Mobile Museum Guidance

Today, audio guides are widely established in many museums. They are used to provide audible information to museum visitors as an alternative or in addition to readable text labels that are mounted close to exhibits. To retrieve the corresponding information visitors either browse through electronic lists or type reference numbers that are located close to the object into a mobile device. In recent solutions, the devices are equipped with RF-readers. They play automatically related information about exhibits as visitors approach the signal range of RFID tags attached to or close to the object in question (cf. figure 3.1).

From our point of view, such audio guides are suboptimal and can be improved in terms of information presentation as well as usability. In comparison to reference numbers, the selection of objects would be more intuitive if visitors were able to just point at an object with their mobile phone's camera to retrieve further information. This is a more unobtrusive technique and particularly beneficial where information about many densely located exhibits are available. When using numbers, lots of reference tags would have to be attached to the objects, cluttering the overall appearance of the exhibits and making it difficult for users to assign the numbers to the correct objects.

As already discussed in section 2, the application of RFID tags would fail too when many objects are located close to each other. A specified reader would receive multiple signals simultaneously and would not be able to unambiguously determine which object is meant. In addition, RFID equipment have to be maintained by appropriately trained staff and have to be acquired, both of which incurs costs.

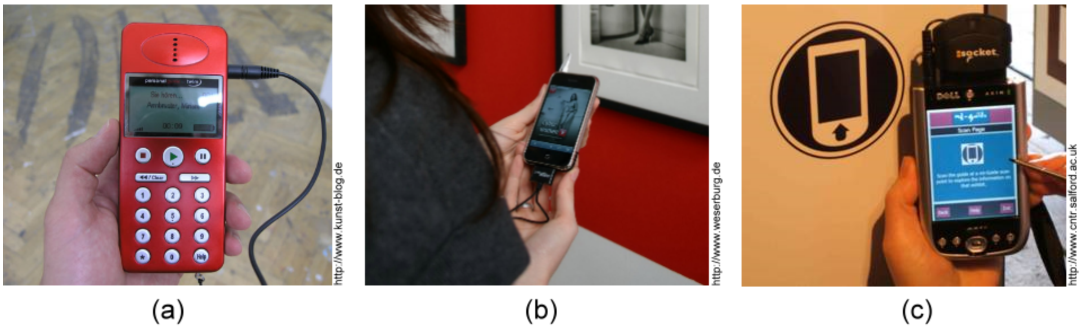


Figure 3.1: A selection of different audio guides that use either reference numbers (a), electronic lists (b) or RFID tags (c) to identify exhibits in a museum.

3.2 PhoneGuide

The approaches that we present in this thesis are based on the results of the PhoneGuide project that we outlined in section 1.3: PhoneGuide was aimed at developing an adaptive mobile museum guidance system that allows visitors in a museum to identify exhibits with their mobile phone. Through on-device object recognition enhanced with additional adaptation techniques this image is identified and appropriate multimedia content is presented to the user.
To provide a comprehensive foundation for understanding the approaches developed in this thesis, we will describe the most important elements and algorithms of our system prior to this work and show how we adjusted them throughout the years.

3.2.1 Basic Object Recognition

The basic object recognition system is based on 100 global color features and a 3-layer neural network for classifying a single object that was photographed with a mobile phone. The color features consist of a 30-bin histogram of each color channel (3x30 = 90) where each normalized bin represents one feature. Furthermore, 5 features f_{0-4} are computed based on the mean of each color channel:

$$f_{0,1,2} = \frac{\sum_{i=0}^{N-1} I_{[0,1,2],i}}{N}$$

where N is the number of pixels in the captured image and $I_{[0,1,2]}$ denotes the intensity in the red, green and blue color channel of each pixel. 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. The color ratio of the remaining channel is given explicitly and is redundant.

Another 5 features f_{5-9} are computed based on the variance:

$$f_{5,6,7} = \frac{\sum_{i=0}^{N-1} \left(I_{[0,1,2],i} - f_{0,1,2}\right)^2}{N}$$

The variances in the color channels can be set in relation in a similar way to 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 ratio of the remaining channel is redundant. At the end, the features are normalized. The entire set of features is then composed to a single feature vector.
These features differ slightly in each approach presented in this thesis as explained in section 3.3. They vary mainly in terms of the number of histogram bins and image resolution as a result of the hardware used. In chapter 4 we omitted the variance features to decrease the overall computation time.

For classification we apply 3-layer artificial neural networks[6] with a hyperbolic tangent sigmoid activation function at the input layer and linear activation functions at the hidden and output layer. In the first layer we apply 14 neurons and at the hidden layer 12 neurons. These values were selected empirically and represent a compromise between generalization and specialization. We decided to use neural networks for recognition because they are memory efficient and the classification process is fast. They increase in size only slightly with the amount of objects and, unlike instance-based classifiers such as nearest neighbor search, do not increase with the number of images per object, which is especially beneficial for our continuous learning approach as explained in the next section.
For our multi-image classification approach, we apply a k-nearest neighbor search for classification. Further information about the reasons are explained in chapter 4.

[6] For simplification we call them ”neural networks” throughout this thesis.

To ensure a basis for the object recognition algorithm we first record videos of each exhibit in the museum that we want to recognize. These videos consist of an arbitrary number of images that show the object from different perspectives and scales. In our evaluations this number varied between 80 to 160 frames. The image resolution depends on the hardware used and was set to 160x120 on low-level mobile phones (e.g., Nokia 6630) and 240x180 on newer phones (Nokia N95, cf. figure 3.4).
After recording, we extracted keyframes from the set of images. This keyframe extraction is necessary to eliminate equal frames and to prevent uncontrolled specialization of the neural networks. Each keyframe of the videos is then separated into $M = 12$ equally-sized patches and for each patch we compute the color features mentioned above. The resulting feature vectors serve as training data to generate M neural networks. This entire pre-processing is carried out once off-line on a PC that we have termed the ”server” throughout this thesis. The resulting neural networks in combination with the front-end application are then transferred to the mobile phones. During the recognition process, no connection to the server is necessary.

To classify an exhibit, a user has to take a picture of it. The corresponding image is separated into M patches and M feature vectors are computed. They serve as input for the neural networks. Using average voting [185] the outputs of the neural networks are combined. In order to interpret the output values as probabilities, a softmax activation function is applied which converts the values so that they lie between 0 and 1. The final outcome is a probability-sorted list of result candidates. The elements of this list are visualized as thumbnails on the mobile phone's display as illustrated in figure 3.2. The object with the highest output is preselected and proposed as the final result. If the classification is correct, the user acknowledges the selection. If the classification has failed, the visitor can browse through the list and select the correct object to retrieve the corresponding multimedia content. The classification performance of the basic image classification is discussed in section 8.

Further information about the basic object recognition algorithm can be found in our publications ”PhoneGuide: Museum Guidance Supported by on-device Object Recognition on Mobile Phones” [67] and ”Adaptive Training of Video Sets for Image Recognition on Mobile Phones” [30].

3.2.2 Adaptation through User Feedback

To ensure reliable object recognition, image data of the different objects has to be collected as mentioned in the previous section. In our approaches, we record a video of each exhibit from different

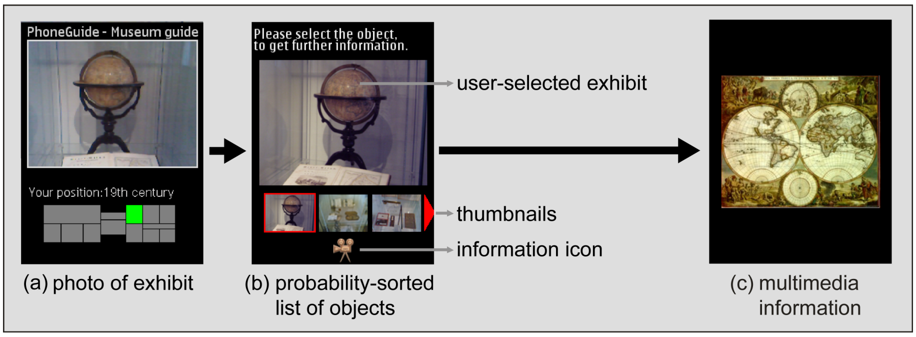

PhoneGuide - Museum guide
Your position:19th century
Please select the object,
to get further information.
user-selected exhibit
thumbnails
information icon
(a) photo of exhibit
(b) probability-sorted list of objects
(c) multimedia information

Figure 3.2: User interface representing the basic object recognition approach. A video stream allows visitors to point at an exhibit. A floor plan indicates the rough location of the visitor in the museum (a). The classification result is presented as a probability-sorted list of result candidates visualized as thumbnails. The selected object is previewed above the list. An information icon indicates what kind of multimedia content (video, image, audio) is available (b). After selecting the correct object, multimedia content is displayed (c).

perspectives and at different scales in order to cover a wide range of possible user locations. These images reflect the subjective behavior of the corresponding expert. However, visitors may act differently and capture the object from perspectives that were not initially recorded. Consequently, this can result in misclassifications.

To overcome this we have proposed a system where the pre-recorded images only serve as an initial guess. New image data is then continuously collected during the usage of our PhoneGuide system. As mentioned before, the visitor selects the correct exhibit from the user interface after image classification has been carried out. For this he browses through the list of objects and compares the pictures displayed with the real exhibit. This selection provides us with a valid mapping between the captured image and the actual object that we store on the visitors' mobile phones. At the end of the museum visit, these images are transferred to our server (e.g., wirelessly or via a memory card) where they are added to the existing image data. This entire data is then used to retrain our classifiers. This technique has two advantages: First, the image classifiers become more robust over time because more data is available for training. Second, the classifiers are being continuously specialized. This means that not every perspective is weighted equally as is the case with the initial training data. Through the continuous collection of new pictures, images of some perspectives will be produced more frequently than others. Consequently, the neural networks become specialized, adapting to the users' behavior, i.e., how visitors approach particular objects. To facilitate this behavior, the keyframe extraction is only applied to the initial dataset.

Since retraining relies on user feedback, there is a possibility that correlations between captured images and the real objects are incorrectly assigned. For example, a user might take photos of something completely different (e.g., the floor or a wall) and associate this with an exhibit. These images would have a negative effect on the training and recognition performance. To overcome this, we eliminate unlikely pairings by clustering the image set of an object and deleting suspicious elements.

In addition to collecting image data during the usage of our PhoneGuide system, we can also collect additional data to continuously improve the overall system. For example in section 7 we track the movements of visitors in a museum. This data is also stored on the mobile phone and transferred to the server to continuously improve the specific classifier over time. We call this entire set of data that we collect during the use of PhoneGuide ”adaptation parameters”. In section 8.1 we will give an overview of these parameters and explain them in more detail.

Further information about the adaptation through user feedback can be found in our publication ”Adaptive Training of Video Sets for Image Recognition on Mobile Phones” [30].

3.2.3 Adaptation through Bluetooth Tracking

The more objects a mobile object recognition system has to recognize, the higher the probability of classification errors. As we have seen in the related work section, many approaches overcome this problem by determining the location of users. Consequently, only the objects that are in the vicinity are considered as potential result candidates.

In our adaptation technique we distribute Bluetooth beacons (cf. figure 3.3) throughout the museum to support a cell-based localization: Depending on its signal strength, every emitter can cover a limited area. Since we apply class 2 Bluetooth emitters the signal range is a maximum of 10 m and can be affected by reflections and absorptions resulting from artifacts such as walls or people. Depending on the granularity of the distribution, signals from multiple emitters can overlap and are consequently detected simultaneously. As a result, an unstructured grid of emitters partitions the environment into different spatial cells of superimposed and single signals. By identifying the physical area of each cell in advance, carried out by an expert, the individual exhibits can be assigned to appropriate signal cells. The number of objects in these cells is significantly lower compared to the entire data set which reduces the number of possible result candidates. For each possible location cell, one neural network is trained by only considering the image data of the objects in this area. If the location of the visitor is determined, the appropriate neural network can

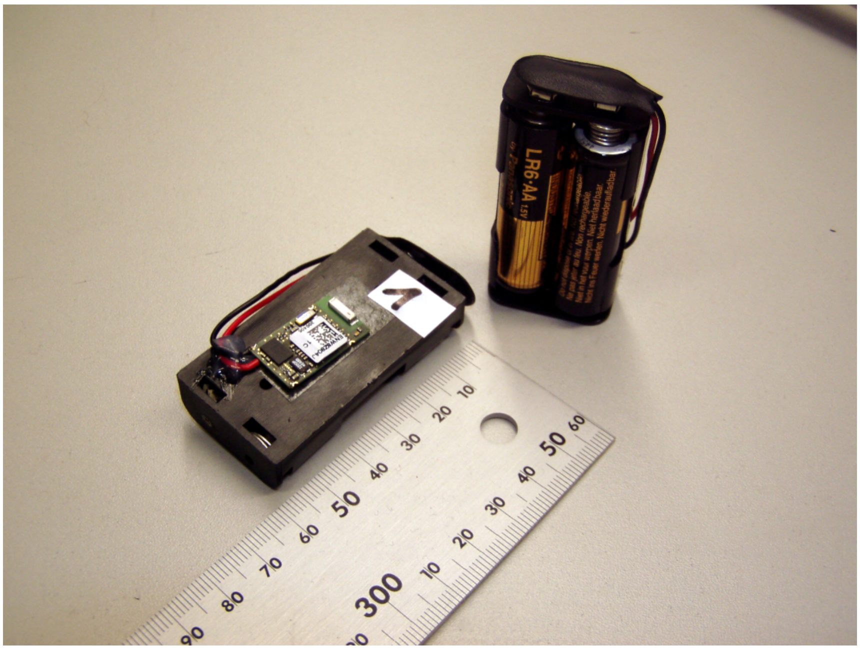


Figure 3.3: Bluetooth emitter [175] with battery pack (scale in millimeter). The Bluetooth beacons are distributed throughout the museum to ensure cell-based localization.

be selected and is used for image classification.

To determine the location of a visitor, his mobile phone continuously scans for nearby Bluetooth emitters. If one or multiple Bluetooth emitters are detected the location is derived through a fault-tolerant implication (= OR relation). This means that if multiple signals are received, we do not consider the intersection of both signals but the entire area covered by both signal ranges. A conjunction (= AND relation) would provide a more precise location estimation but dynamic absorption and reflection effects caused, for example, by people, may mean that it would not always be possible to detect every theoretically visible Bluetooth emitter. Consequently, the wrong neural network classifier may be selected and the classification would fail.

With the proposed configuration we achieved approximate accuracy at room level. The location of the visitor is visualized on the mobile phone by highlighting the current room on a map as illustrated in figure 3.2a.

Further information about the localization approach can be found in our publication ”Enabling Mobile Phones To Support Large-Scale Museum Guidance” [36].

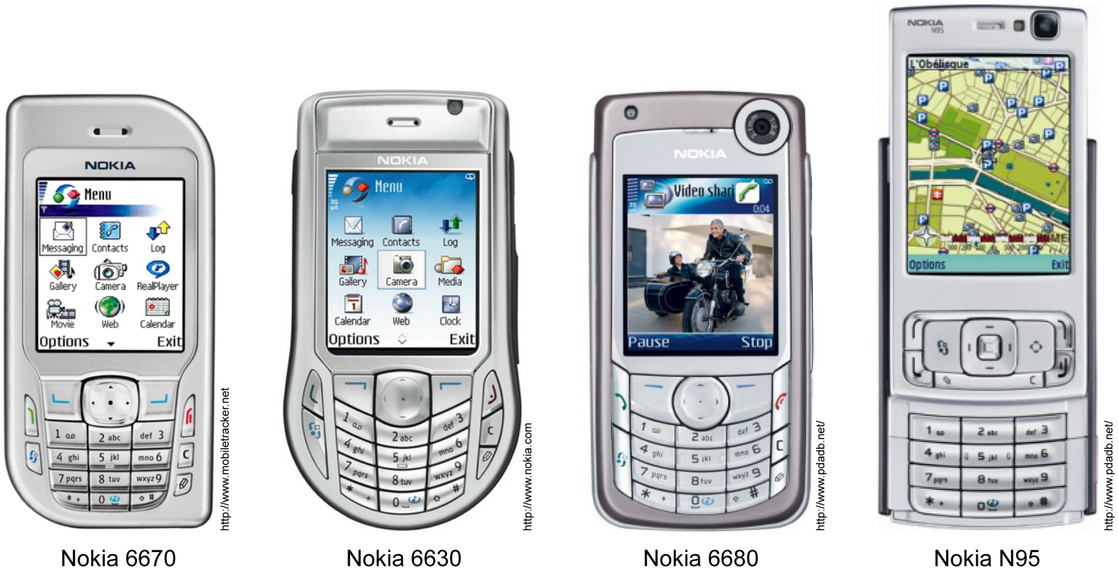


Figure 3.4: The phones used for evaluating the different approaches: Nokia 6670 (CPU: TI OMAP 1510 123MHz; display: 2.1", 176x208; camera: 1.0 MPixel; Bluetooth 1.1, released: October 2004). Nokia 6630, Nokia 6680 (both, CPU: TI OMAP 1710 220MHz; display: 2.1", 176x208; camera: 1.3 MPixel; Bluetooth 1.1, released: November 2004, May 2005). Nokia N95 (CPU: TI OMAP 2420 332 MHz; display: 2.6", 240x320; camera: 5.0 MPixel; Bluetooth 2.0, released: March 2007) [49].

3.3 Preliminaries

Based on the PhoneGuide system, this thesis presents four different approaches that facilitate adaptive image classification using mobile phones. Although these methods are part of the entire system, they are thematically unrelated to each other. Consequently, we will explain them in separated sections including their corresponding evaluations. At the end of this dissertation, we provide an overview of all the methods and show how they influence the classification rate in conjunction.

Each corresponding section of these approaches starts with a motivation (e.g., a problem that has to be solved or an enhancement) which explains the reason for developing this method. Afterwards, preliminary information is given where necessary and the algorithms are explained in detail. By presenting the results of one or multiple user studies, we discuss the benefits of the different techniques.

Due to the rapid evolution of mobile phones in terms of CPU power, memory or display size, and the increased availability of newer phones, we used different mobile phones (cf. figure 3.4) for our field studies that we present in this work. Consequently, we continuously adjusted the image description and image resolution slightly to achieve better classification performance. In each evaluation subsection, we briefly describe what kind of device and image description we have used for this approach. The image features outlined in the section 3.2.1 denote the current state of feature extraction.

In addition, in each relevant section we show how the user interface was enhanced compared with the previous techniques. At the end of each section we discuss the method we developed and propose improvements.

4 Multi-Image Classification

In the previous chapter we have explained how to identify an exhibit through object recognition by capturing a single photo of it. Since this task is challenging in general, the image classification can fail if many objects have to be recognized from various perspectives and scales. To improve classification performance, an obvious and straightforward solution would be to consider multiple pictures from different perspectives rather than just one photo for recognition. In contrast to single-image methods, where one photo is matched against a database to identify the corresponding object, multi-image classification approaches combine the classification results of multiple image instances to help identify the object. In the following we will show how this solution can be enhanced further.

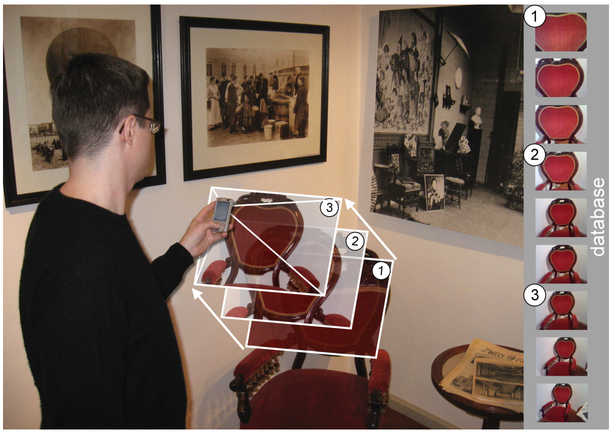


Figure 4.1: A visitor in a museum captures a sequence of images with his mobile phone by moving the device backwards. Afterwards distinct keyframes are extracted and matched against a database. Only those database images that exhibit the same near-far relations as the captured video frames are used to derive the final classification result.

The concept of combined multi-image classification has already been the subject of research as presented in section 2.3.1 [158, 168, 95]. But in contrast to related approaches, we developed a solution that is especially designed for mobile phone applications and that is enhanced by a method that we call ”relational reasoning”. For this technique visitors have to move their phones physically from near to far as illustrated in figure 4.1.

The series of images that are captured during this camera movement are then individually classified. The near-far relations (e.g., the best-match database image of the first captured image is closer to the object than the best-match database image of the second captured image and so on) of the images stored in the database are then examined. These have to be consistent with the near-far relations of the images captured by the user. Classification results of individual pictures that fail this near-far relation condition are not considered for the combined image classification.

Through this relational reasoning, we can achieve high classification rates with a minimum number of frames since several potential result candidates with wrong near-far relations can be discarded.

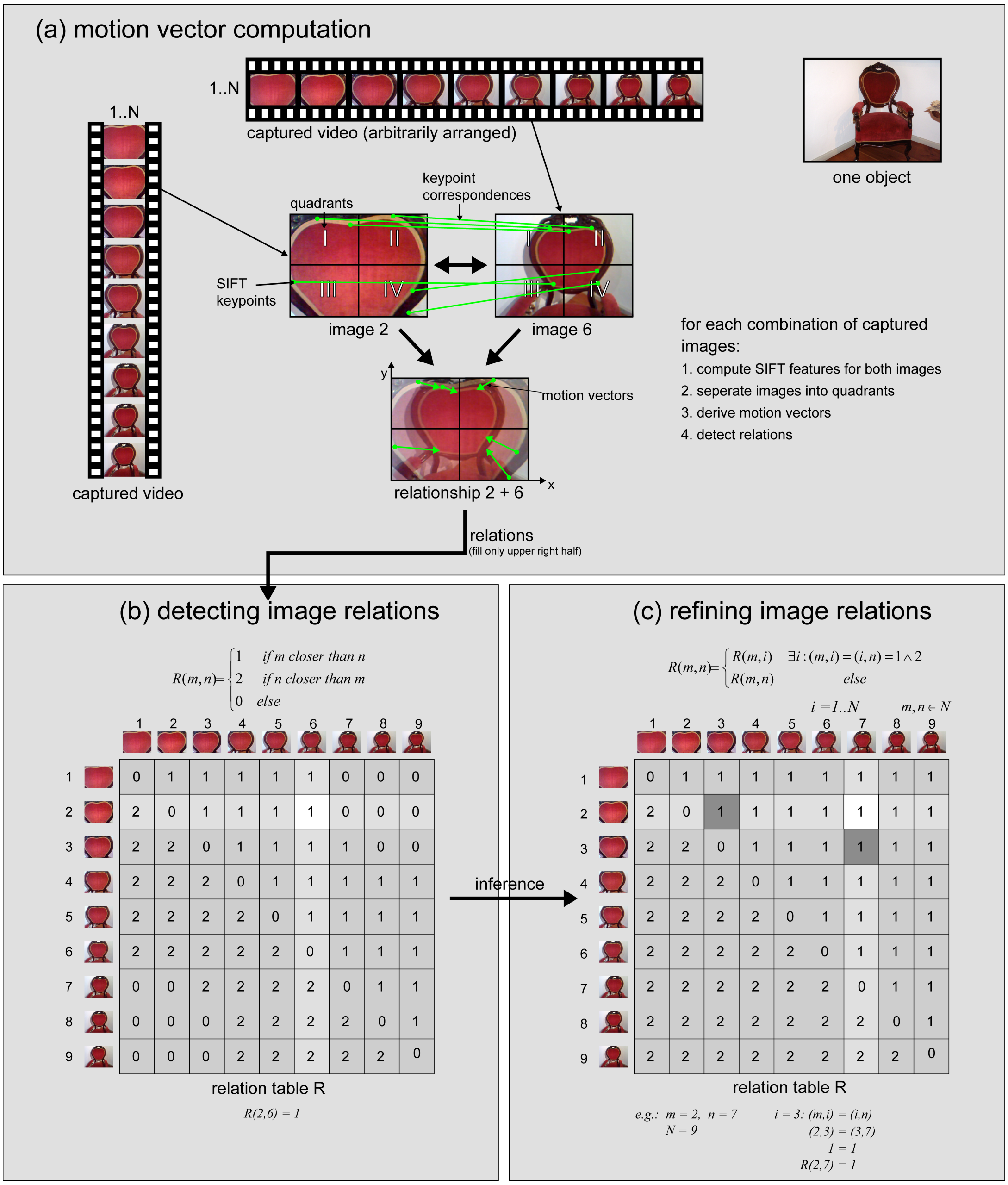
The research presented in this chapter is based on our publication ”Mobile Museum Guidance through Relational Multi-Image Classification” [34].

4.1 Relational Multi-Image Classification

In contrast to single-image classification where only one image is considered for identifying an object, we combine the results of multiple images in order to enhance the classification performance. In our approach, users are requested to record a short video sequence of an object by moving the camera phone from near to far while centering on the exhibit during recording. We decided to apply a near-far movement because of two reasons: First, it seems to be more convenient for the visitors than moving the phone from left to right or up to down under the restriction to keep the exhibit centered. Second, the near-far camera movement describes the object more properly. Therefore, in initial tests, we recorded videos of 30 different objects from left to right and determined the variance as well as the entropy of computed color features over the entire set of all images. In this experiment, both values were outperformed by the corresponding counterparts of the near-far movement so we concluded that near-far camera movement is more effective.

In general, our method works as follows: After classifying a selected keyframe of the recorded video sequence through a nearest neighbor search, each keyframe has one corresponding nearest neighbor in the image database that belongs to one particular object. How to determine the keyframes is described in section 4.2.1. The results are combined by assigning a vote (an incrementing number) to each object whose image was detected. Following a majority voting scheme [185], the object with the highest vote count is classified as the final result.

In order to further improve the classification rate and to reduce the number of necessary keyframes, we developed a technique that takes the near-far relations of the retrieved database images into account. For each database image identified during classification, its near-far relation to previously retrieved database images of the same object is examined. If it does not have a near-far relation to previously recognized database images, the corresponding classification result of this image is not considered for majority voting. In the following we will describe the off-line preprocessing steps that are necessary for our approach. These are carried out only once on our server and the results are used on the mobile device for classification as explained in section 4.3.



for each combination of captured images:

1. compute SIFT features for both images
2. seperate images into quadrants
3. derive motion vectors
4. detect relations

(b) detecting image relations

$$R(m,n)=\begin{cases}1 & \text{if } m \text{ closer than } n\\ 2 & \text{if } n \text{ closer than } m\\ 0 & \text{else}\end{cases}$$

	1	2	3	4	5	6	7	8	9
1	0	1	1	1	1	1	0	0	0
2	2	0	1	1	1	1	0	0	0
3	2	2	0	1	1	1	1	0	0
4	2	2	2	0	1	1	1	1	1
5	2	2	2	2	0	1	1	1	1
6	2	2	2	2	2	0	1	1	1
7	0	0	2	2	2	2	0	1	1
8	0	0	0	2	2	2	2	0	1
9	0	0	0	2	2	2	2	2	0

relation table R

$R(2,6) = 1$

(c) refining image relations

$$R(m,n)=\begin{cases}R(m,i) & \exists i:(m,i)=(i,n)=1\wedge 2\\ R(m,n) & \text{else}\end{cases}$$

$i=1..N \qquad m,n \in N$

	1	2	3	4	5	6	7	8	9
1	0	1	1	1	1	1	1	1	1
2	2	0	1	1	1	1	1	1	1
3	2	2	0	1	1	1	1	1	1
4	2	2	2	0	1	1	1	1	1
5	2	2	2	2	0	1	1	1	1
6	2	2	2	2	2	0	1	1	1
7	2	2	2	2	2	2	0	1	1
8	2	2	2	2	2	2	2	0	1
9	2	2	2	2	2	2	2	2	0

relation table R

e.g.: $m = 2,\ n = 7$ $\quad i = 3: (m,i) = (i,n)$
$N = 9$ $\quad (2,3) = (3,7)$
$1 = 1$
$R(2,7) = 1$

Figure 4.2: Preprocessing steps for retrieving image relations: After recording videos of each exhibit, keyframes are extracted. For each possible pair of keyframes, SIFT features are computed and correspondences in both images are detected. The relation between each pair of images is then derived from the motion vectors defined by the correspondences. If motion vectors, originating in at least two different quadrants, point towards or away from the center of the image, a near-far movement is detected (a). The image relations are then encoded and stored in a relation table (b). Ambiguous image relations are resolved by inferring their relations based on other known image pairs (c). One relation table is stored for each exhibit.

4.2 Off-line Preprocessing

In order to ensure the relational reasoning approach we need to determine the near-far relations among the database images for each object that has to be classified. For building the database, we record videos that show each exhibit from arbitrary perspectives and scales. For each video, we determine the near-far relations among the unordered video frames once on a server as explained in the following section. In section 4.2.2 we describe the necessary steps for creating the image classifier that are used on the mobile device.

4.2.1 Detecting Image Relations

To discover the near-far-relations among the unsorted set of database images that we collected in advance, we extract keyframes in an off-line preprocess on our server first: Therefore color features are computed for each frame. The first frame of each video sequence is selected as an initial keyframe. Each subsequent frame is then compared to all keyframes that have been identified previously by computing the Manhattan distance between the corresponding feature vectors. If the distance to all existing keyframes is larger than a predefined threshold, the current frame is defined as additional keyframe. This threshold is chosen to be very small to consider as many images as possible and to prevent occurrences of equal keyframes since these would only increase the overall amount of data without providing additional information for classification. Equal frames occur if similar perspectives are recorded (e.g., if the phone is not moved). For each keyframe, we retrieve distinct image keypoints by computing SIFT features [115]. Based on these keypoints we determine the motion vectors between any two keyframes by finding keypoint correspondences (cf. figure 4.2a). We assume that the motion vectors for backward motions mainly point towards the image center. We define that a near-far relation between two keyframes is detected if motion vectors, originated in at least 2 different quadrants of one keyframe, are pointing to or away from the center of the other keyframe. In addition, motion vectors that point in different directions are allowed in only one quadrant at the most. We determined these constraints empirically to handle outlier motion vectors. The determined near-far relations for each image combination are stored in a relation table R (cf. figure 4.2b) that is defined by equation 4.1:

$$R(m,n) = \begin{cases} 1 & \text{if } m \text{ closer than } n \\ 2 & \text{if } n \text{ closer than } m \\ 0 & \quad\quad \textit{else} \end{cases} \tag{4.1}$$

$$m,n \in N$$

where N is the number of keyframes. For example, $R(2,6) = 1$ denotes that image 6 was captured further away from the exhibit than image 2 (thus, $R(6,2) = 2$). If no distinct near-far relation could be detected for two keyframes the corresponding entry equals 0. This occurs, if for instance, our motion vector conditions are not fulfilled or not enough keypoint correspondences could be detected. The latter can happen because the image perspectives are either too different or the images are motion blurred. Since this occurs frequently, we developed a technique that refines R by inference. For each table entry that equals 0 we try to infer its near-far relation based on the known relations of other image pairs using equation 4.2:

$$R(m,n) = \begin{cases} R(m,i) & \exists i : (m,i) = (i,n) = 1 \wedge 2 \\ R(m,n) & else \end{cases} \tag{4.2}$$

$$i = 1..N$$

An unknown relation $R(m,n)$ can be resolved if two pairs of images (m,i) and (i,n) exist that have equal relations. This relation is then used instead of the unknown relation. An example is illustrated in figure 4.2c: since image 2 is closer than image 3 and image 3 is closer than image 7, image 2 has to be closer than image 7. Consequently, $R(2,7) = 1$. This procedure is carried out for each undefined image pair. For the inference process we try to ensure that the determined near-far relations in the first step are highly reliable in order to prevent the propagation of wrong relations. Therefore, keypoint correspondences are only taken into account where their matching probability is high. The entries for table cells which could still not be resolved remain set to 0. The detection of the image relations of the first step only has to be processed for the upper right part of the matrix since it is symmetrical. For each exhibit one NxN relation table is stored. These tables are initially uploaded to each phone together with the image classifiers. How to generate them is discussed next.

4.2.2 Generating Image Classifier

As outlined in chapter 3 we had to exchange the neural networks classifier with an instance-based learning technique because the index information of the individual training images is lost when applying neural networks. Note, that we have replaced the neural network classifier only for the multi-image classification method. We decided to take a k-nearest neighbor approach in combination with Manhattan distance since it is simple and effective. However, in contrast to neural networks the classification time increases significantly with an increasing number of images and objects. Therefore, we implemented a vantage point tree (vp-tree) as stated in [196]: A vantage

point tree is generated by selecting initially one root element p from the set of elements S (in our case the feature vectors computed from the database images) in feature space. Then, a binary tree is recursively built by taking the median of the set of all distances $d_M = mediandist(p,s), s \in S$: All elements s such that $dist(p,s) \leq d_M$ are inserted into the left subtree and all elements s such that $dist(p,s) > d_M$ are inserted into the right subtree. A query q is performed in $O(logn)$ and the tree needs $O(n)$ space. In general vp-trees are applied in many different research fields for speeding up the classification process in high dimensional spaces [159, 161]. Other tree structures that are based on distance measurements can be used as well [129].
Since the original vp-tree algorithm only outputs the 1-nearest neighbor, we implemented a k-nearest neighbor search as described in [72]. Furthermore, in each node of the tree we added the index of the corresponding image for addressing the relation table.

Note, that the vp-tree has to be computed only once on the server and is uploaded to each phone together with the relation tables. During usage, no connection to the server is necessary and all computations are carried out directly on the mobile device.

4.3 On-line Relational Multi-Image Classification

If a video of an exhibit is captured, we extract keyframes on-the-fly as explained in section 4.2.1. After the recording is stopped, the final keyframes serve as basis to carry out the relational multi-image classification: Each keyframe is classified by the vp-tree to retrieve the k-nearest neighbors. Empirically, we chose k=7 as a trade-off between performance and possible misclassifications. For each of the k-nearest neighbors, we determine the near-far relations with all k-nearest neighbors of all previously determined keyframes. Note, that this is only possible if two nearest neighbor pairs belong to the same object (i.e., it can be indexed in the same relation table). For each matching near-far relation (i.e., the near-far relation retrieved from an object's relation table matches the near-far relation of the corresponding keyframes), the according object's vote count is increased. If multiple objects with equal vote counts exist the object with the best matching nearest neighbor is selected. This strategy is also applied if, for some reason, the user does not follow the required camera movement. In this case no valid relations are produced and no votes are counted.

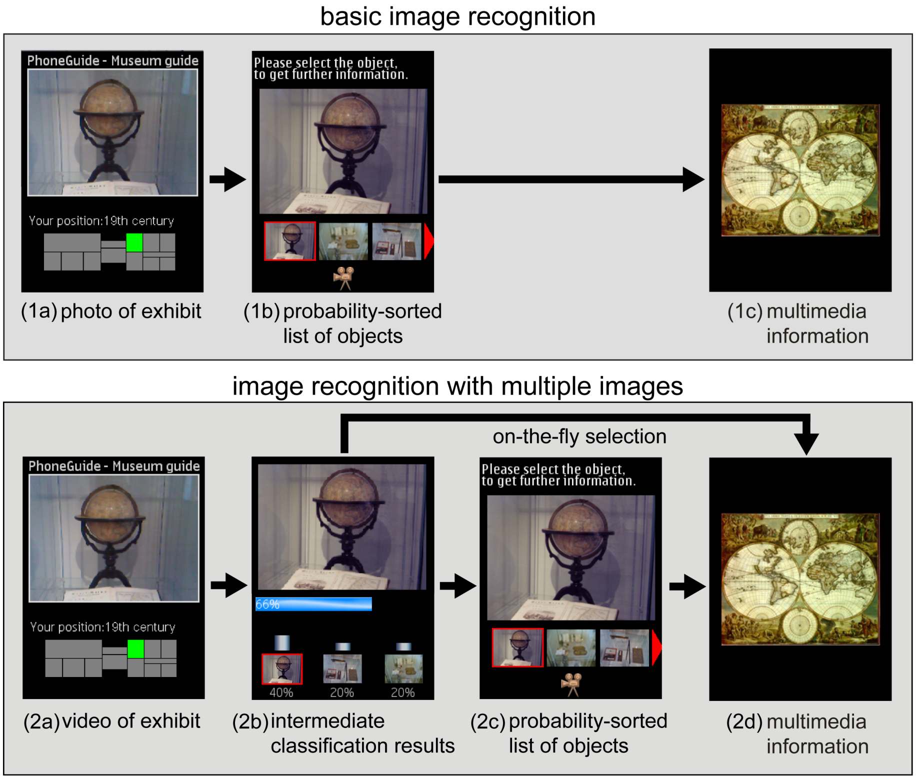

basic image recognition
PhoneGuide - Museum guide
Your position: 19th century
Please select the object, to get further information.
(1a) photo of exhibit
(1b) probability-sorted list of objects
(1c) multimedia information
image recognition with multiple images
on-the-fly selection
PhoneGuide - Museum guide
Your position: 19th century
66%
40%
20%
20%
Please select the object, to get further information.
(2a) video of exhibit
(2b) intermediate classification results
(2c) probability-sorted list of objects
(2d) multimedia information

Figure 4.3: Flow chart showing the user interface for single and multi-image classification. If the user takes a single photograph of an exhibit (1a), an image classification is carried out and the result is presented as a probability-sorted list of objects (1b). The correct object can be selected with a minimum number of clicks in order to retrieve the corresponding multimedia information (1c). If a set of images was recorded (2a) the relational multi-image classification is carried out for each keyframe. Intermediate results are visualized through object thumbnails in combination with the corresponding classification rate on the screen (2b). Users are allowed to select the correct result during the classification phase or by choosing from a list of objects that is presented at the end of the classification process to retrieve the multimedia information (2c-2d).

4.4 Graphical User Interface

The extensions of the graphical user interface compared to the interface of the single-image classification is illustrated in figure 4.3. Visitors who are facing an exhibit they wish to classify have to center it in the camera view of the phone and press the control button to start the recording (figure 4.3.2a). The button is released to finish the recording and to start the multi-image classification. For each determined keyframe, intermediate classification results are presented on the screen (figure 4.3.2b). The three most likely candidates are displayed as thumbnails with their corresponding probability represented by vertical bars. The probability denotes the quotient of the object's vote count and the sum of all vote counts and represents the algorithm's confidence about the results. In addition, a progress bar indicates the percentage of processed keyframes. There are two reasons why we present these on-the-fly results: First, it provides users with continuous feedback on the progress of the classification process. Second, it allows visitors to interrupt the classification process if the correct result is already displayed as a thumbnail. They can then either select it on-the-fly by moving a selection frame via the control button or by pressing 1, 2 or 3 on the keyboard. After the classification has been completed or aborted, the final probability-sorted list of objects is presented (figure 4.3c). This list contains all possible candidates, beginning with the most likely candidate on the left-hand side of the screen. The user can now select the correct object and the corresponding multimedia content is presented (figure 4.3d).

4.5 Evaluation

We evaluated our approach in the City Museum of Weimar, Germany. For this we carried out a qualitative user study with 11 inexperienced users between the ages of 23 and 37 years (average age: 28.5 years, 4 female, 7 male). We call the subjects ”inexperienced users” because they took pictures arbitrarily from different perspectives and scales and they did not have any knowledge of the classification process.
We recorded initial videos (240x180 pixels, 80 frames, showing different perspectives and scales) of 154 different exhibits in advance which were of different color, size and shape including black and white pictures, furnishings, clothes, etc. Of each frame we computed 95 global color features: 30-bin histogram in each color channel and mean in each color channel. To speed-up the feature extraction we omitted the variance features and the separation into image patches as mentioned in section 3.2.1. These videos were used for the off-line preprocessing stage as explained in section 4.2.

The resulting vp-tree needs ∼4.9 MB of storage and the relation tables approximately 1 MB. During preprocessing, relations could be assigned to 65.9% on average for all image pairs based on the SIFT motion vectors. Using our inference method this could be increased to 88.5%.
Each subject captured an arbitrary, freely chosen number of videos of different exhibits. They all used a Nokia N95 (cf. figure 3.4) during the field study. At the beginning of each user evaluation, the subjects were advised to center every object and to start from close up and move the phone away from the object. In total the subjects captured 493 videos (with a minimum of 23, a maxi-

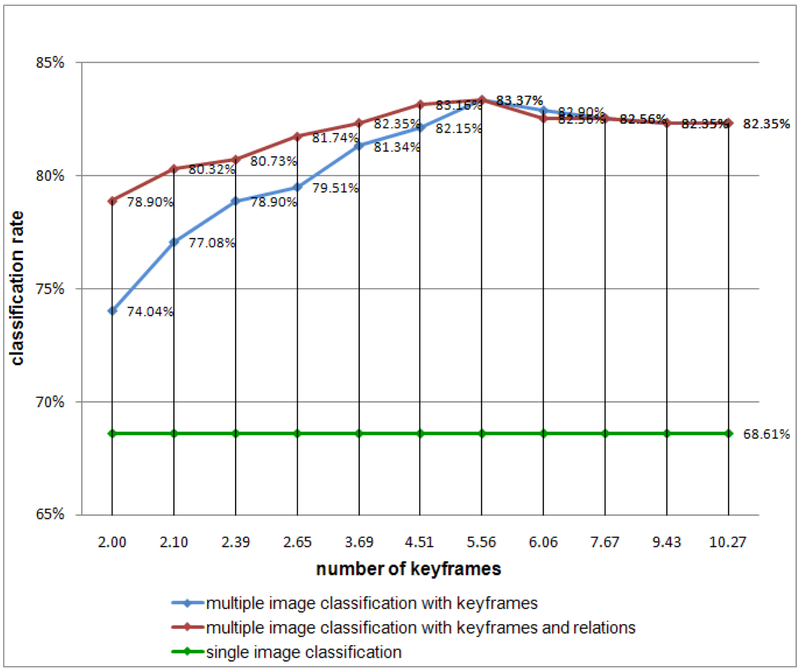


Figure 4.4: Result of a multi-image classification experiment, using a near-far camera movement. The multi-image classification was applied with and without relational reasoning and its performance is shown for different numbers of keyframes. The performance of single-image classification is independent of the number of keyframes and therefore constant.

mum of 80, and an average of 45 videos per subject) so each exhibit was captured at least 3 times (average: 3.2). The recording rate was approximately 4 fps and included on-the-fly feature computation. In total, the subjects captured 5060 images which resulted in ~10.27 images per object (with a minimum of 1 and a maximum of 19 images). The average duration for classifying a frame was approximately 1s and for determining its relation ~0.015s, i.e., ~1.015s in total. However, these timings strongly depend on our programming platform and on the mobile device (JavaME, Nokia N95).

The results of the user study are illustrated in figure 4.4. We show the multi-image classification techniques separately with and without relational reasoning as well as its single-image classification counterpart for different numbers of keyframes: The single image classification denotes the classification rate if each image of the video sequences is considered separately. It achieves a classification rate of 68.61% considering all 5060 images which is of course independent of keyframe extraction and thus constant.

The second graph shows the multi-image classification rate without relational reasoning. The graph shows that for an increasing number of keyframes, the classification rate also increases since more data is available for reasoning until it peaks at approximately 5-6 frames where the classification rate remains constant.

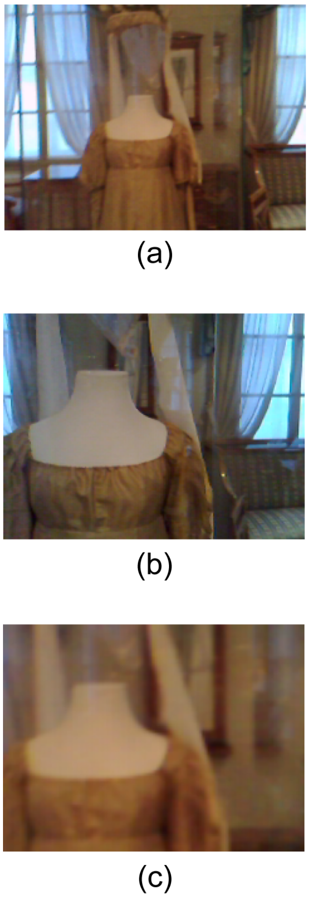


Figure 4.5: Comparison between near-far camera movement (a,b) and digital zoom from the far position (c).

The third graph shows the multi-image classification rate with relational reasoning. Its development is similar to that of the previous graph, however, especially if the average number of keyframes per object is low, the relational multi-image classification method outperforms the multi-image technique without relational reasoning by up to 5% and peaks after approximately 5-6 keyframes with a recognition rate of 83.37%. In addition, it needs on average 27.7% less keyframes to achieve the same classification rate than the multi-image classification without reasoning if the number of available keyframes is low (i.e., if the number of available keyframes is smaller than ∼6 on average). This increases the overall classification speed by up to 35%. For example, to achieve a classification rate of 81.74% the relational multi-image classification needs 2.65 keyframes (classification time: 2.67s) on average while the multi-image classification without relational reasoning requires 4.14 keyframes (classification time: 4.14s). Note that the classification times heavily depend on the applied classification technique.

Through the application of multiple images we can enhance the recognition rate of single image classification by up to ∼15% depending on the number of extracted keyframes. After approximately 5-6 frames the classification rates of both multi-image approaches are similar and no difference can be observed since the benefit of the relations is compensated for by the additional amount of keyframes. To ensure that each object benefits from the multi-image classification technique, we guarantee to consider at least 2 images per object by defining the last frame of the video sequence as additional keyframe if only one keyframe had been detected up to that point.

An alternative to recording a video while carrying out a near-far camera movement would be to take a single shot from a far position and then generate multiple nearer image instances using a

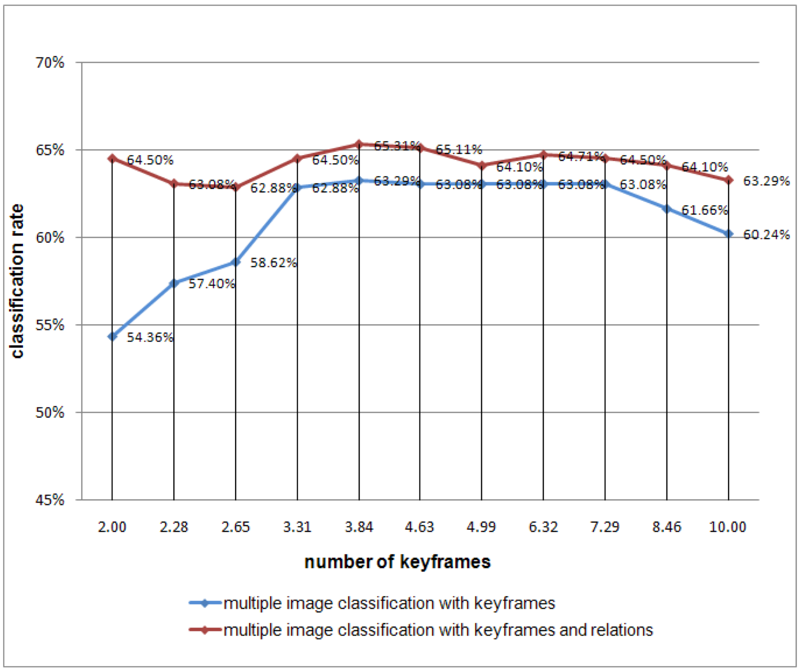


Figure 4.6: Result of a multi-image classification experiment, using a digital zoom of one shot. The multi-image classification was applied with and without relational reasoning and its performance is shown for different numbers of keyframes, comparable to figure 4.4.

digital zoom technique. In order to evaluate this digital zoom approach, we collected 16 images from different perspectives and scales and generated 4 additional digitally zoomed instances of each picture so that here too we had a total of 80 images for the preprocessing stage. For classification, we selected from each of the 493 recorded near-far video sequences the last image and created 9 digital zooming steps from them (10 images per object, 4930 images in total), in order to have a similar number of frames as for the near-far camera movement technique. Figure 4.6 shows the classification results of the digital zoom technique that we computed offline. As we can see, the general improvement of relational reasoning is similar to that of a near-far camera movement (cf. figure 4.4).

However, the overall classification rate is significantly lower than that of a near-far camera movement, and also lower than that of single image classification. We believe that the reasons for this are the following: The field-of-view is smaller if a high digital zoom level is selected compared to a related video frame representing the same perspective. This problem is illustrated in figure 4.5 where a set of objects is captured from a far distance (4.5a), from a near distance (4.5b) or generated by digital zooming of an image that is originally captured from a far distance (4.5c). The digitally zoomed scene does not only provide less details but also shows a narrower perspective

with less content. Consequently, useful image information is lost that is helpful for feature computation and classification. Since the classification of digitally zoomed images is generally lower, a higher number of false positives lead also to a lower rate of multi-image classification (with and without relational reasoning).
The impact of different elements such as keyframe extraction and relational reasoning of the near-far camera movement approach are illustrated in figure 4.7. For three selected numbers of keyframes the results reveal that relational reasoning achieves similar classification performance with little more than 2 frames (2.1) than the corresponding multi-image classification with $\sim$8 images (80.32% to 79.72%) if no relational reasoning or keyframe extraction is carried out. In addition, it illustrates that the more original frames are considered for classification per se, the smaller the impact of keyframe extraction since the classification performance saturates. This saturation results from the fact that the extracted image features are not discriminating enough for unambiguous classification.

Since the near-far camera movement leads to more promising results but is also somewhat cumbersome, we asked the subjects for their impression on the usability of our technique. On a scale of 1-5 they were asked to vote if the handling of moving the phone while capturing images is better (5) or worse (1) than capturing a single photograph (see appendix A.1). Five people voted that it makes no difference for them at all. Four people found it slightly worse. One person judged it much worse and one person slightly better. The average vote was 2.55. Finally, 10 of 11 subjects would prefer the near-far camera movement if it produces higher classification rates. Although the number of subjects is not representative, it does indicate that multi-image classification through an intuitive camera movement would be accepted in practice.

In section 8.3 we will show how this approach influences the overall classification rate of our entire system.

4.6 Discussion

In this chapter we have proposed a method that utilizes multiple images of a single object to improve the image classification performance. We have shown that this method can improve the classification rate by up to 15% compared to single image classification techniques. Furthermore, we have demonstrated that a digital zoom technique is not a viable alternative for near-far camera movement. If we consider the perspective relations among the database images which we have called relational reasoning, we can increase the classification speed by up to 35% and the approach

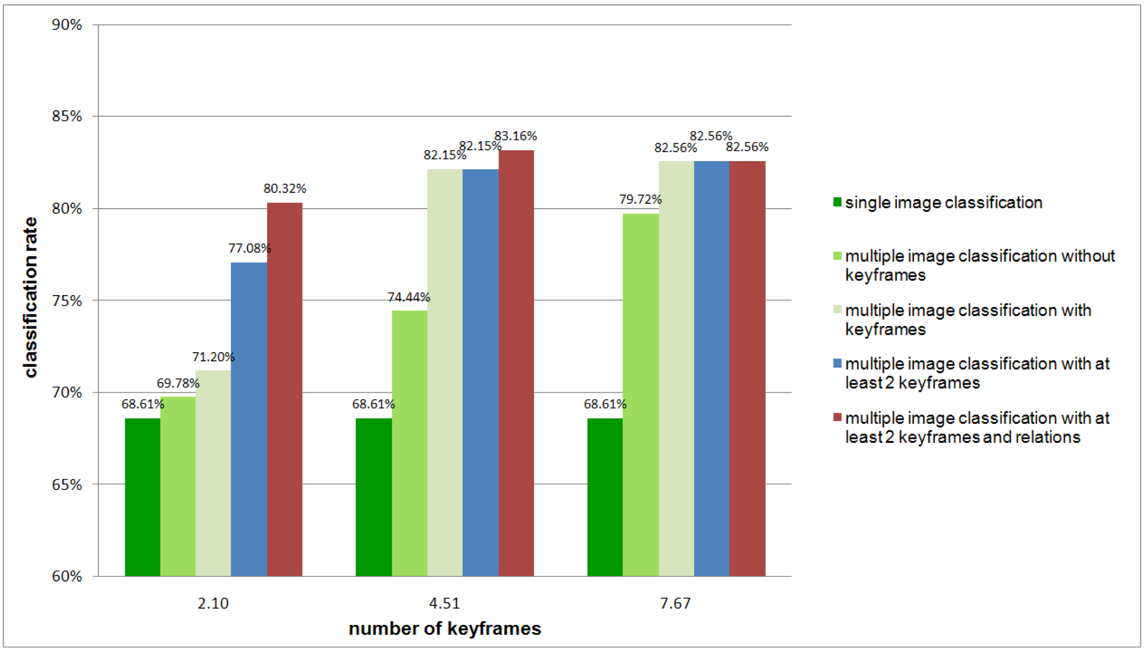


Figure 4.7: Single-image classification and the different elements of our relational multi-image classification and their impact on the classification rate depending on the number of extracted keyframes.

saves on average 27.7% keyframes if the number of available keyframes is low.

Although initial tests have demonstrated that the near-far camera movement outperforms related simple movements in terms of image description as explained in section 4.1, we have not clarified whether other camera movements exist that may perform better. In particular, more complex movements probably further enhance the classification rate, for example, when visitors are allowed to capture video sequences of an exhibit from arbitrary perspectives and distances. In such situations, unlike our simple predefined near-far camera movement, the image relations between the captured images would be unknown. Consequently, the complex camera movement have to be estimated through, for example, natural feature tracking [184].

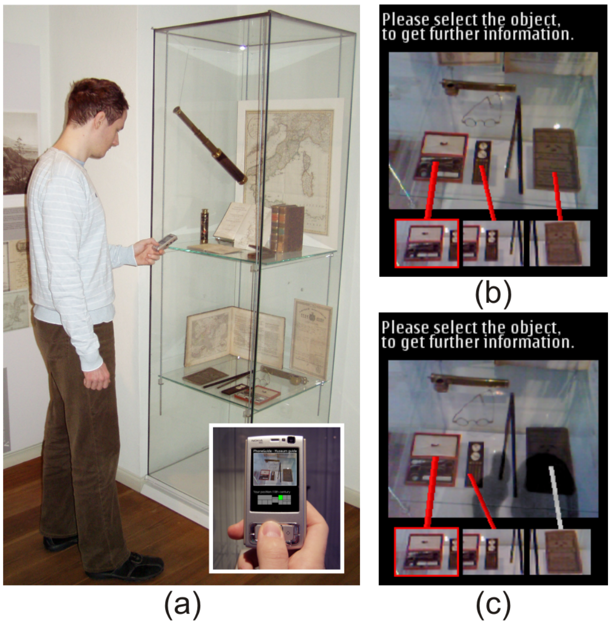
Furthermore, if multi-image classification is performed on-the-fly, instead of carrying out the classification afterwards, instructions on how to optimally move the camera to identify the object quickly could be shown on the camera display. Visitors could move their mobile device around the object until the confidence of the object recognition algorithm exceeds a predefined threshold.

5 Subobject Recognition

In many museums, objects are placed in showcases or behind other barriers to protect them against environmental influences and human curiosity. This means that if visitors want to target one of them for identification, they capture multiple exhibits in a single image. Consequently, the unambiguous classification of only one object is not possible.

On the other hand, sometimes people are not interested in a specific exhibit but want to investigate what kind of objects are located in front of them. They, therefore, take a picture of multiple exhibits and want to retrieve information about each object individually.

In both scenarios, an object recognition process is necessary that is capable of detecting and annotating each object individually that is visualized in a photo.



Please select the object, to get further information.

Please select the object, to get further information.

Figure 5.1: A visitor takes 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, the known spatial relationships among the subobjects still allow a correct identification.

In the previous chapters we have shown how a single object can be identified by either capturing one or multiple images of it. The image features that we apply for these classifications describe the whole image or image patches globally which means that each pixel in the image is considered equally. They are, therefore, unable to distinguish if only one or multiple objects were photographed.

To overcome this one could apply local image features which describe distinct keypoints in the image. However, applying such techniques exclusively is challenging, especially if objects are very small and of poor contrast: reliable local image keypoints might not be detected in the image areas of the corresponding photographed objects. As a result, no appropriate description would be available and the classification may fail.

We have, therefore, developed a technique to detect multiple objects, called subobjects, in a single image by considering the spatial relationships among the exhibits. This has several advantages: first, if at least one subobject was detected in the image through an object recognition task, it serves as a reference point to span search regions defined by the spatial relationships. These search regions are in general smaller than the entire image, so searching for additional subobjects reduces the search time in comparison to, for example, brute force methods. This is especially beneficial for performance-limited mobile phone applications. Second, the more subobjects are detected, the more precise the search regions and the more robust the detection with regard to misclassifications. Third, partially or even completely occluded subobjects (e.g., occluded by shadows or other exhibits) where image classifiers typically fail, can be detected with this approach.

Our method follows two basic steps: First, a set of objects is treated as one exhibit and it is identified via image classification. With this context information, the related subobjects are detected in the image through a combination of image classification and spatial relationships in the second step. After they are all detected, the subobjects are annotated as illustrated in figure 5.1b and the user can select the object of interest for retrieving corresponding multimedia information.

The research presented in this chapter is based on our publication ”Subobject Detection through Spatial Relationships on Mobile Phones”[26].

5.1 Off-line Preprocessing

As mentioned before, the classification process is separated into two steps: In the first step a regular image classification is carried out. It identifies the entire scene rather than individual subobjects in the image. The preprocessing steps that are necessary to ensure this classification as well as the classification itself are described in chapter 3.

In order to support the identification of subobjects during the second classification step, 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 that we have captured in advance. Then, we 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 our color features 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 subobject detection and recognition algorithm. They are initially computed on the server as part of the one-time preprocessing step. Once computed, they are used on the mobile phones for subobject classification during runtime. The following sections will explain how these two components are computed in more detail.

5.1.1 Registration and Tracking of Subobjects

The first step of the preprocessing is the detection of all subobjects in each frame of the captured video. As indicated above, the bounding boxes (subimages) of all subobjects are manually defined in the first frame of a scene video. They have to be automatically tracked throughout the subsequent video frames in order to compute global features for the subimages framed by the axis-aligned bounding boxes. In addition, the locations of the subobjects in each image have to be known for deriving the spatial relationships among the detected subobjects.

In general, many different algorithms are available to track an object through multiple images. We evaluated three different tracking techniques: Template matching with fast normalized cross-correlation [109], tracking based on SIFT features [115] and kernel-based mean shift tracking [47]. The fast normalized cross-correlation computes for each location a correlation factor in a predefined area. This factor is estimated by normalized pixel intensities of the image and the template. To speed-up the process the convolution is carried out in frequency domain. Tracking through SIFT is accomplished by computing SIFT features as explained in section 2.3.1 in a predefined area around the last location of the template. By matching the SIFT descriptors of the template with the features of the image patch, the location of the subobject is updated. Mean shift tracking computes for each pixel in the image a value that indicated the probability that it belongs to the template. The template's mean position is then derived by the first order image moments computed from the computed probability map.
By evaluating these techniques we found that mean shift tracking is the most robust method for low-resolution video recordings (160x120 pixels in this case). Local feature extraction techniques, such as SIFT, would perform better if the video resolution would be increased but to ensure reliable keypoints at very small objects would have been still challenging.

The algorithm of the preprocessing is illustrated in figure 5.2. The manually tagged subobjects (figure 5.2a) are clustered based on the size of their bounding boxes via a simple agglomerative clustering technique (5.2b). 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 then

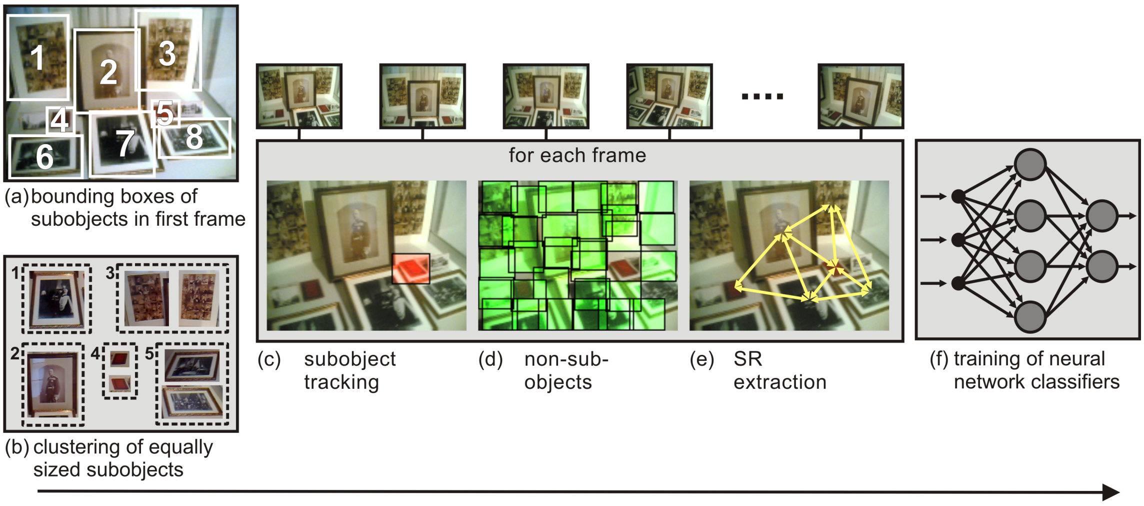

1
2
3
4
5
6
7
8
(a) bounding boxes of subobjects in first frame
1
2
3
4
5
(b) clustering of equally sized subobjects
....
for each frame
(c) subobject tracking
(d) non-sub-objects
(e) SR extraction
(f) training of neural network classifiers

Figure 5.2: Flow chart of the off-line preprocessing: After a video of a scene with subobjects is 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).

tracked throughout all frames via mean shift tracking (5.2c). The 2D pixel locations of each subobject's center on the image plane are used for deriving the spatial relationships to other subobjects within each frame (5.2e). In addition to the subimages that actually contain exhibits, additional subimages of the same size are also automatically collected in each frame (5.2d). We refer to them as *non-subobject* subimages. They are used later as negative samples for training the neural networks.

5.1.2 Generation of Subobject Classifiers

After tracking all subobjects throughout the training videos, a certain number of subimages for each subobject is stored and is available for training (figure 5.2f). 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 are extracted from each subimage and are 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 subobjects. 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., we have 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 [185].

5.1.3 Extraction of Spatial Relationships

If the detection of subobjects would be exclusively performed by image classification, the entire image has to be scanned and tested against different subobject classifiers. This is both computationally exhausting and unreliable. Spatial relationships describe how the subobjects are arranged in relation to each other (figure 5.2e). This has preliminary two advantages for the on-line 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 between two subobjects within the image. They are variant 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 subobjects 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. Other relations like a topological mapping (relations are defined in terms of the intersections of the boundaries and interiors of two sets) [56] or direct relations (cardinal directions) are not beneficial since they only provide relative but no absolute relationships among objects.

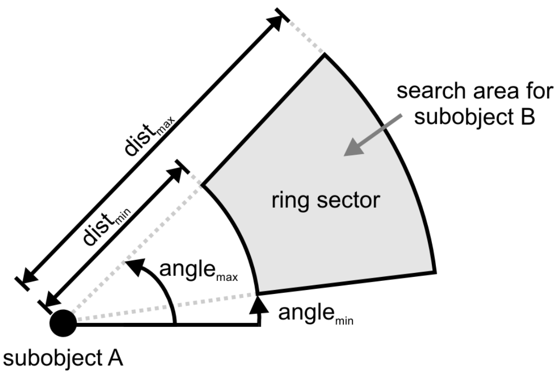


Figure 5.3: Predicted search area of an undetected subobject (B) relative to a detected subobject (A). The corresponding ring sector is defined by the minimum and maximum distances and angles that were extracted during subobject tracking in the off-line preprocessing.

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 5.3) that describes the location of one exhibit relative to another one. Each subobject is related to 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 is used for on-line classification during runtime.

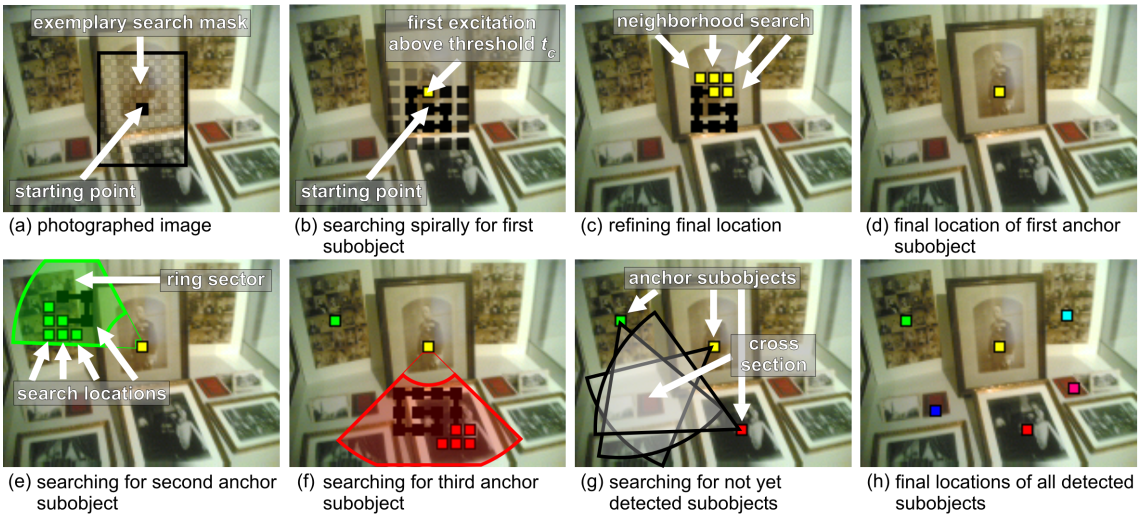
5.2 On-line Subobject Recognition

The on-line subobject detection algorithm can be separated into three main steps in order to identify N subobjects: In the first step, the algorithm searches for M ($M < N$) subobjects that serve as anchors for determining the current rotation and scale relationships among them reliably. Then the remaining $N - M$ subobjects can be detected faster while continuously refining the spatial relationships. Finally, undetected subobjects that are expected to appear in the image are located by prediction using the determined geometric dependencies. The following sections will explain this in more detail.

5.2.1 Detection of Anchor Subobjects

Since the correct scene context is given through the first classification step and the visitor's 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.
To find the first anchor subobject, no prior knowledge about geometric relationships or the actual quantity 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 5.4a) 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. With each step, the search mask's size is adjusted to all clustered subobject sizes that were generated during the off-line training. For each pixel region that is covered by a search mask, the global color features are computed from a precomputed integral image [181]. Integral images store in each pixel the sum of the intensity values in each color channel of all pixels that have a smaller x,y value than the current pixel. This has the advantage that for an arbitrary image region the mean in each color channel can be computed very fast and independent of the image size in a constant time.
The extracted features serve as input for the classifiers to identify the first anchor subobject. It is detected if the following conditions are met (cf. figure 5.4b): (1) the maximum excitation of C_{all} is above a predefined threshold t_c, (2) the size of the identified subobject equals 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 5.4c) 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 5.4d) 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 off-line preprocess (figure 5.4e). 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 vary either with



(a) photographed image

(b) searching spirally for first subobject

(c) refining final location

(d) final location of first anchor subobject

(e) searching for second anchor subobject

(f) searching for third anchor subobject

(g) searching for not yet detected subobjects

(h) final locations of all detected subobjects

Figure 5.4: After an image was captured, the subobject detection searches for the first anchor subobject by varying the size of the search mask (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 cross sections that define reliable search areas of the remaining subobjects (g). This is repeated until all subobjects are detected (h).

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 computed distance and expected distance (from the off-line preprocessing) between all possible detected subobject pairs at that time. The rotation correction angle is derived from the average quotient of the differences between the detected and the expected angle as described in [58]:

$$\theta_{RCA} = -\arctan \left[\frac{\sum\limits_{i=0}^{n-1} \sum\limits_{j=i+1}^{n} \sin \theta_{ij}}{\sum\limits_{i=0}^{n-1} \sum\limits_{j=i+1}^{n} \cos \theta_{ij}} \right] \tag{5.1}$$

where θ_{ij} denotes the difference between the detected angle and the expected angle between two subobjects i and j. The resulting rotation correction angle θ_{ij} is then applied during the detection

of further subobjects.
Another way of determining the rotation correction angle is by utilizing the built-in accelerometer of today's phones. They determine the relative pose of the device and can be applied before the subobject detection starts. Since several phones are not equipped with such sensors today we use the solution mentioned above that operates on each phone.

We also have to consider false positives (i.e., incorrectly detected subobjects) during the detection phase. False positives influence the successive development of spatial relationships and therefore lead to wrong search area predictions 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:

$$SIM_{cda} = \omega_1 \cdot P_c + \omega_2 \cdot SIM_d + \omega_2 \cdot SIM_a \tag{5.2}$$

$$P_c = P(A) \cdot P(B) \tag{5.3}$$

$$SIM_d = \left[1 - \frac{|D_{AB} - d_{AB}|}{\sqrt{W^2 + H^2}} \right] \tag{5.4}$$

$$SIM_a = \left[1 - \frac{|\alpha_{AB} - \beta_{AB}|}{180} \right] \tag{5.5}$$

Equation 5.2 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 5.3), 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 5.4), 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 off-line. The last component, SIM_a (equation 5.5), defines the normalized similarity of the currently computed angle β_{AB} between subobject A and B, and the expected (pre-computed) angle α_{AB}. All three components are weighted with ω_1, ω_2 and ω_3 ($\omega_1 + \omega_2 + \omega_3 = 1$). 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 found that a minimum number of three anchor subobjects is necessary for a reliable determination of the angle and distance of the phone relative to the real exhibits. From here, a faster detection technique that is mainly based on the spatial relationships can be used to find the remaining subobjects. This is explained in the following section.

If the SIM_{cda} of one subobject in relation to multiple other subobjects is low while the quality among the other subobjects is high, then this indicates that the particular subobject was probably misclassified and consequently its detection is discarded.

5.2.2 Detection of Remaining Subobjects

If a sufficient number of anchor subobjects are found, the remaining subobjects can be reliably detected by applying 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 5.4g). The cross sections 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 cross sections are not computed since the computational costs would be too expensive. Instead, the search locations (cf. figure 5.4e) are tested against each ring sector individually.

To detect the remaining subobjects, only C_{spec} of the subobject in question is applied. Remember, that we know which subobject is located in this search region based on its spatial relationships. Searching the subobjects 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, the detection of the remaining subobjects is 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 subobject in order to continuously refine the search areas.

However, if the output of C_{spec} for all tested locations is below the threshold t_c, no subobject is 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 object. Therefore, the locations of the missing subobjects are predicted exclusively through spatial relationships. Its location is defined to be the center of gravity of the corresponding cross section. An example of such a case is illustrated in figure 5.1c: Although the user casts a shadow on the book which leads to an image-based misclassification, the exhibit is still detected using spatial relationships.

Finding subobjects exclusively through their spatial relationships opens the opportunity to even

locate objects that are always completely occluded by other objects, or ones that are so small that image classifiers cannot detect them reliably. Such subobjects are tagged in the training video to extract the corresponding spatial relationships without training their C_{spec} classifiers and without considering them for the C_{all} classifier.
For instance, this is beneficial if the detected subobjects are individual parts of a large-sized object (e.g., a statue). Then the spatial relationships can be used to annotate missing parts based on the detected elements. For example, the location of a missing arm can be annotated to provide corresponding information about it.

After all subobjects have been detected (cf. figure 5.4h), the labeled list of subobjects is presented to the user, as illustrated in figures 5.1b,c.

5.3 Graphical User Interface

The extensions of the graphical user interface compared to the interface of the single-object classification as presented in section 3.2.1 is illustrated in figure 5.5: In the first step (1a, 2a), a scene, containing one or multiple exhibits, is photographed and identified. It classifies the entire scene rather than individual subobjects in the image. Following this process, a probability-sorted list of objects is displayed (1b, 2b). The list contains all possible candidates, beginning with the most likely candidate on the left-hand side of the screen. The user can now select the correct scene context with a small 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 as explained in section 3.2.1.

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

5.4 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,

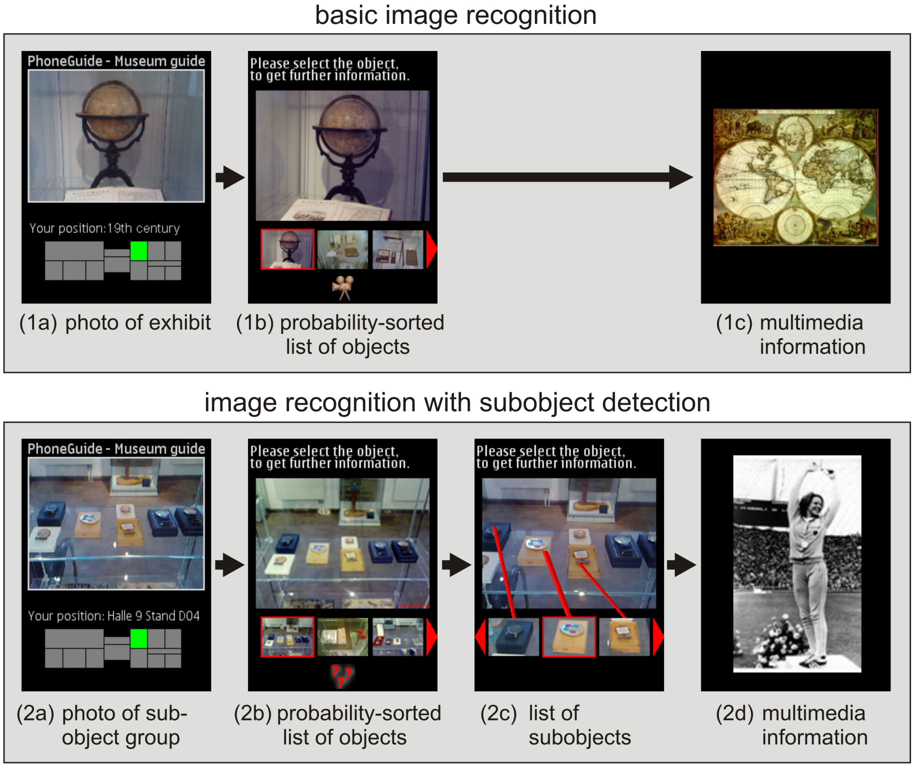

basic image recognition
PhoneGuide - Museum guide
Your position:19th century
Please select the object, to get further information.
(1a) photo of exhibit
(1b) probability-sorted list of objects
(1c) multimedia information
image recognition with subobject detection
PhoneGuide - Museum guide
Your position: Halle 9 Stand D04
Please select the object, to get further information.
Please select the object, to get further information.
(2a) photo of sub-object group
(2b) probability-sorted list of objects
(2c) list of subobjects
(2d) multimedia information

Figure 5.5: Flow chart showing 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 list of objects (1b). The correct object can be selected with a minimum number of clicks for receiving multimedia information (1c). If a group of subobjects is captured rather than a single object (2a), the user first has to acknowledge the scene classification (2b), before a consecutive subobject classification can be 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).

with inexperienced visitors)?

For the performance analysis, we have compared the subobject detection technique with a brute-force search method that scans the whole image for subobjects. Furthermore, we have compared our method with a brute-force search method with early stopping (ES) that cancels the search when 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. For both experiments (laboratory and field test) 12 subobject groups were selected (6 of them are displayed in figure 5.6). The number of subobjects per group ranged from 3 to 8 subobjects (average: 5.4). For each group, a video consisting of 90 frames (160x120 pixels) was recorded from various perspectives and distances. For the image description, we used 40 global color features consisting of three 10-bin color histograms as well as mean and variance in the color channels. In each video every third frame of each video was used for classification in the laboratory experiment so that in total 720 frames were applied for training and 360 frames were applied for simulating the recognition. The experiments were carried out on Nokia 6680 and Nokia N95 mobile phones (cf. figure 3.4).

5.4.1 Performance Analysis

In general, a subobject detection that applies spatial relationships should perform faster than methods that scan the entire image, since only predefined subregions are examined. In addition, the subobject detection 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 first compared our approach with a brute-force search method that scans the whole image for subobjects: Starting at the image center, the search mask is spirally moved to each possible location until it has reached each part of the image. At each location, global color features are extracted to perform the classification with the C_{all} and C_{spec} classifiers. In order to compare both methods properly, parameters such as search mask size and step size are the same as in our approach. 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 search method with early stopping is carried out in a similar way to the prior method. The only difference is, that it stops searching for a specific subobject, if the output of both classifiers, C_{all} and C_{spec}, are above t_c. 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 5.6. 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 is displayed.

For each subobject group, 30 randomly selected images from different perspectives and distances were used to determine the results. These images contained different quantities of subobjects,

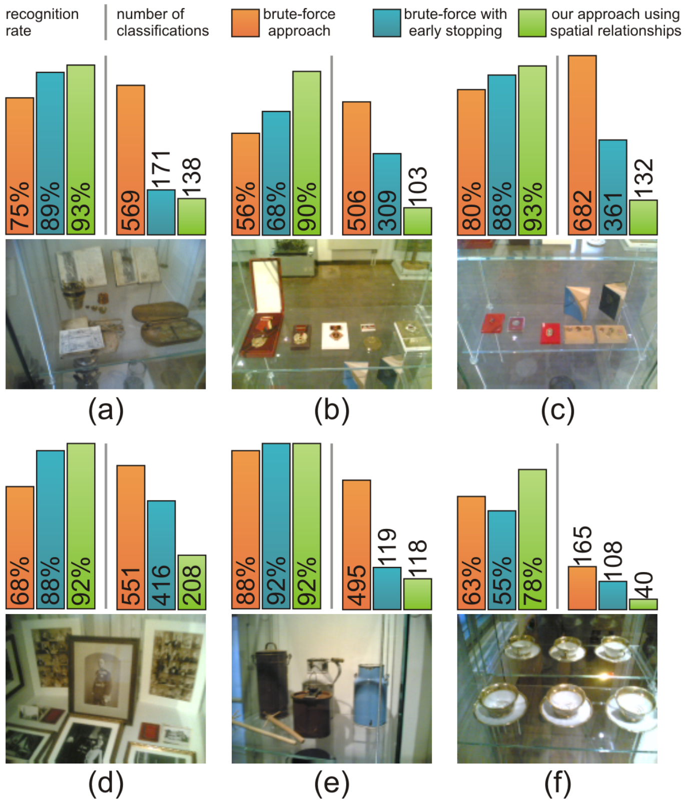


Figure 5.6: Average recognition rate and number of classifications for a brute-force search, a brute-force search with early stopping, and for our approach (6 out of 12 different subobject groups). Thirty images from different perspectives and distances were selected and classified for each group. The graphs show that our approach outperforms related approaches without spatial relationships, both in speed and recognition rate.

since they can be outside the captured frame or (partially) occluded. The brute-force search method reaches an average classification rate of 83.2% (for 12 subobject groups) with 13.4% false positives. The brute-force search 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 search approaches.

Beside an improved recognition rate, figure 5.6 illustrates that our recognition process also needs, on average, fewer classification steps, which correlates with lower classification times. Thus, our method is much faster than brute-force search methods and brute-force ES search methods.

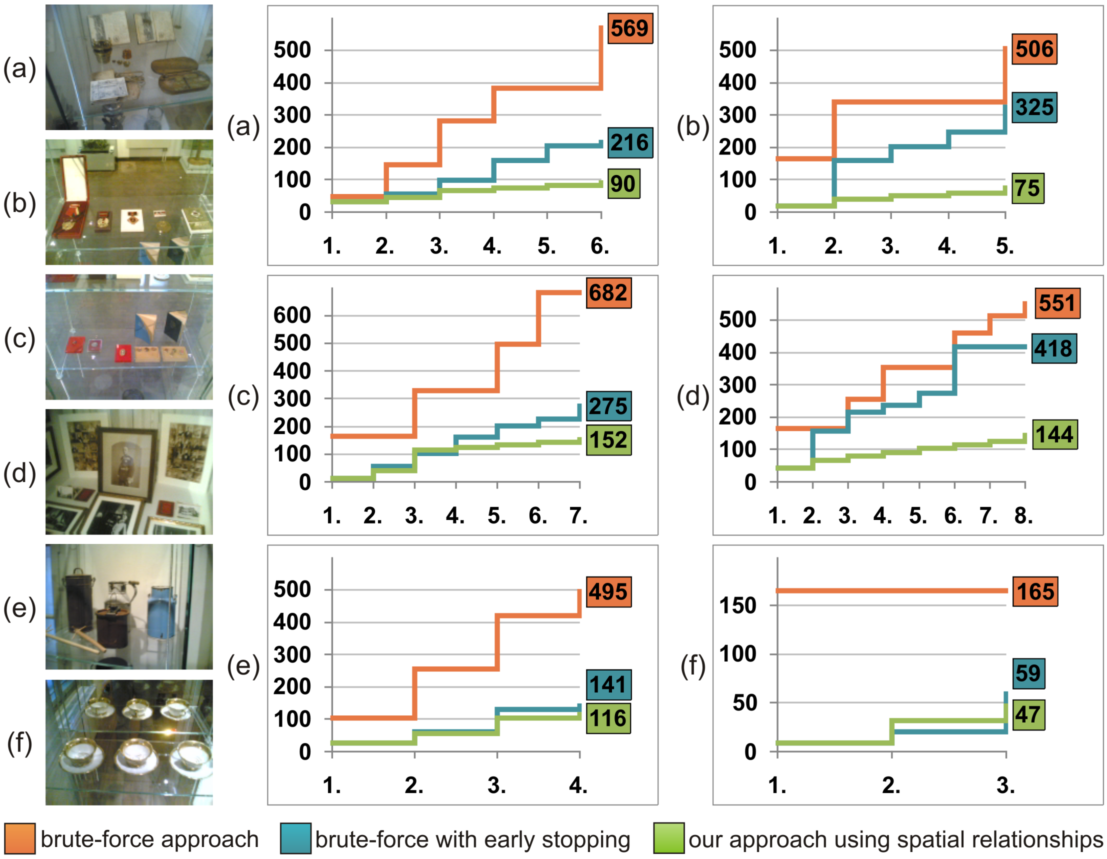


Figure 5.7: Number of classifications required for brute-force search, brute-force search with early stopping and for our approach. For each of the 6 subobject groups, one image was selected to determine the number of classifications for each subobject. It indicates that applying spatial relationships requires less classification steps.

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 5.7. 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 5.7a), 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 search approach and brute-force ES search 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 the subobjects are equally sized. Thus, if all subobjects have the same search mask size, the number of required classification steps is constant with the number of subobjects, as can be seen in figure 5.7f. However, even in such cases, the number of classification steps of the brute-force methods is still higher than in our approach.

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

5.4.2 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 a Nokia N95 mobile phone. The subobject groups, and consequently the spatial relationships and classifiers were identical to the ones that were used in the performance analysis. The size of the classification data for 12 subobject groups with a total of 64 subobjects was 237 kB.
The recognition rate 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 easily be visually separated, the classification performance was reliable. Thus, the worst recognition result (52.4%) occurred when a subobject set consisted of three almost identical cups in front of a mirror (cf. figure 5.6f). 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.

The time for subobject detection, including integral image computation, ranged between 1.25 seconds and 4.45 seconds, (average: 2.85 seconds on a Nokia N95), depending on the number of subobjects, the number of clusters and the number of required classifications. Since the first classification step (i.e., recognizing the scene context) takes less than 0.5 seconds 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 (see appendix A.2) and rate different aspects of our system with scores from 1 (worst) to 7 (best). With this, we wanted to receive feedback on the usability of the subobject detection as well as the user's acceptance of the required computation time and achieved classification rate. 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) were displayed. The duration of the first step took ∼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 be satisfied with a recognition duration of 2-4 seconds, and 46% would prefer a classification time under 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 for 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.

5.5 Discussion

In this chapter we have presented an approach that applies image classification in combination with spatial relationships to detect multiple objects in a single image. We have shown that the recognition of subobjects using spatial relationships is up to 68% faster than related approaches without spatial relationships. Results of a field experiment in a local museum have demonstrated that inexperienced users reach an average recognition rate for subobjects of 85.6% under realistic conditions.

However, there are also some drawbacks with our proposed method: One disadvantage is its sensitivity to scaling. If the distance between the visitors and the objects differs significantly from the distance used when recording the training images, the image classification might fail. This influences the application of the spatial relationships negatively, for example, because the anchor subobject could not be found. However, most people approach the same exhibits in a similar way and capture images from similar perspectives and distance, as experienced during the field study.

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 insufficient training and classification. Increasing the image resolution would solve this problem but also negatively affect the classification speed. However, as the processing power of mobile phones increases, this problem will decrease in future.

The proposed algorithm was designed and evaluated by considering a single photo for classification. However, as mentioned in chapter 4, we have shown how to recognize a single object by capturing a video of an exhibit instead of taking just one picture. This improves the classification rate significantly. Consequently, this technique could also be beneficial for detecting the subobjects more reliably. For example, one approach would be to classify the subobjects at multiple scales and combine the results, e.g., through average voting. However, up to now, we have not tested this.

6 ADAPTATION THROUGH PHONE-TO-PHONE COMMUNICATION

During the use of our mobile museum guidance system, we continuously collect user feedback by storing the captured images of the visitors on the mobile device as explained in section 3.2.2. At the end of a museum visit, these pictures are transferred to our server in order to adapt and improve the image classifiers over time.

One drawback to this approach is that the classifier cannot adapt to changes that happen during runtime. Once downloaded at entry, the classifiers remain unchanged until the visitor leaves the museum. The ability to continuously update the mobile phone data would have several advantages: First of all, the image classifier could compensate for varying lighting conditions. As explained in section 3.2.1 our basic object recognition approach is based on global color features. As the illumination changes, so too do the extracted image features in response to the level of illumination. The classification may then fail if the pre-captured training videos were recorded under different lighting conditions. It would be laborious to capture each exhibit under all possible illumination states, and this would result in unbalanced classifiers due to imprecise object descriptions. The application of more sophisticated image feature detection methods that are more robust with regard to varying levels of illumination would reduce the number of misclassifications. However, these too would be influenced negatively by strong illumination variations.

Another advantage would be that the system could adapt to changes that influence the object directly. For example, exhibits may be rearranged or exchanged, which cannot be compensated for by image classification techniques.

Finally, not only classification data could be updated but also the information content about exhibits. For instance, several approaches exist that allow visitors to annotate exhibits or leave digital messages associated with objects [133, 37, 194]. In such cases, the ability to rapidly update this data would be beneficial in order to retrieve up-to-date information.

One solution to accomplish this would be to set up a client-server infrastructure. The mobile phones would then be continuously connected to a remote server to exchange data. However, as mentioned in the introduction, the original aim was to avoid the usage of additional hardware in order to reduce the maintenance and hardware cost for indoor applications.
As a result, we propose an adaptation technique that exchanges data directly among the phones through ad-hoc network connections via Bluetooth. To cope with environmental changes such as varying lighting conditions or object manipulation, we use this technique to distribute user feedback

information during runtime. Through effective synchronization of multiple phones, we enforce cooperative classification improvements over the duration of the users' actual visit.

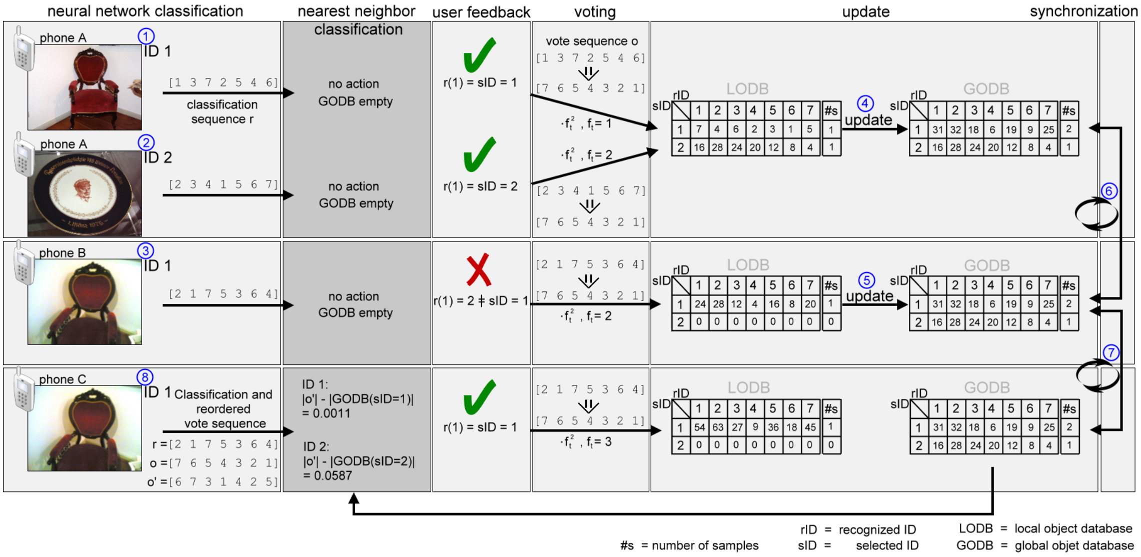
The research presented in this chapter is based on our publication ”Phone-to-Phone Communication for Adaptive Image Classification” [31].

6.1 Adaptation through Ad-hoc Networking

The basic idea of our phone-to-phone communication approach is that we collect information about the individual classification behavior of all visitors simultaneously on each user's local device. This information is derived from the user feedback that is recorded during individual recognition tasks. It is then shared with the phones of other visitors whenever possible.

As explained in section 3.2.1, we retrieve an ordered sequence (*r*) of potential result candidates as a response of the 3-layer neural networks after each classification. The first candidate in this sequence has the highest probability of being the correct object. In cases of false positives, however, it is likely that the correct object is still ranked under the top candidates - even though it is not the correct match. It is also likely, that the classification sequence is slightly different and a distinct description for each object. These two assumptions motivated us for a consecutive classification step: After classifying the image through a neural network, the corresponding classification sequence is reformulated as feature vector that describes the object. For each recognized exhibit, one feature vector is stored on the mobile device of each museum visitor. Through phone-to-phone communication these feature vectors are exchanged among the mobile phones and accumulated on each local device. The local collection of feature vectors then serve as database to carry out a nearest neighbor search to retrieve the final classification result. In the following we will describe this technique in detail.

With respect to figure 6.1, a vote is assigned to each ranked candidate in the classification sequence, after a recognition is carried out. The candidate with the highest probability (*r(1)*) gets the highest vote, the second element (*r(2)*) gets the second highest vote and so on. We assign votes instead of considering the output probabilities because it might happen that some probabilities become very small. This can lead to rounding errors after further processing. In addition, by employing votes it is straightforward to use the phone-to-phone communication in combination with the multi-image classification method, which also assigns votes to the result candidates as explained in chapter 4. The resulting vote sequence (*o*) is multiplied with the square of a time-dependent factor f_t^2 to amplify more recent classification results. This feature vector is then stored in a local object database



LODB (phone A)

sID \ rID	1	2	3	4	5	6	7	#s
1	7	4	6	2	3	1	5	1
2	16	28	24	20	12	8	4	1

GODB (phone A)

sID \ rID	1	2	3	4	5	6	7	#s
1	31	32	18	6	19	9	25	2
2	16	28	24	20	12	8	4	1

LODB (phone B)

sID \ rID	1	2	3	4	5	6	7	#s
1	24	28	12	4	16	8	20	1
2	0	0	0	0	0	0	0	0

GODB (phone B)

sID \ rID	1	2	3	4	5	6	7	#s
1	31	32	18	6	19	9	25	2
2	16	28	24	20	12	8	4	1

LODB (phone C)

sID \ rID	1	2	3	4	5	6	7	#s
1	54	63	27	9	36	18	45	1
2	0	0	0	0	0	0	0	0

GODB (phone C)

sID \ rID	1	2	3	4	5	6	7	#s
1	31	32	18	6	19	9	25	2
2	16	28	24	20	12	8	4	1

Figure 6.1: Adaptation example (sequence numbers encircled): Object 1 (chair) is first successfully recognized on phone A (1). The voted classification sequence is time-weighted with f_t^2 and stored in A's LODB. Then object 2 (plate) is successfully recognized on phone A, and the LODB is expanded accordingly (2). Phone B fails in recognizing object 1 (3), and stores the weighted vote sequence in its local LODB. Then phones A and B are able to establish a connection. They both update their GODBs first (4+5) before synchronizing them (6). Next, a connection between phones B and C is established and GODBs are synchronized (7). The classification of object 1 on phone C (8) would fail when relying solely on the feedback from the neural network, but it will succeed if the nearest neighbor classification is performed together with the GODB.

(*LODB*) table, which is individually managed on each phone. The size of the *LODB* table is *NxN*, where *N* is the number of objects that are trained in the neural network. The columns of the *LODB* indicate the recognized object IDs (*rID*) of *r*, as classified by the neural network. The rows indicate the selected object IDs (*sID*) that result from the users' feedback after each classification. Thus, the time-weighted vote sequence *o* computed from each classification sequence *r* is stored in the *LODB* at row *sID*. In the optimal case, *sID* and *r(1)* are always equal. In this situation, the diagonal of the *LODB* will store the highest votes. If, however, the neural network classification must frequently be corrected by user feedback, this will lead to higher votes on *LODB*'s off-diagonal entries. If the same *sID* results from the user feedback of multiple recognition tasks, the new vote sequences are weighted and added to the existing entries in the corresponding *LODB* row. A row-individual counter is then increased to indicate how many samples have been accumulated.

Our goal is to distribute and synchronize the information collected from the feedback of each user to all other users and apply this knowledge for adapting the classification of each individual visitor. If, for instance, the classification of one particular object will fail for many users due to the change of environmental lighting (e.g., sunlight passing through a window), this will be detected through a similar user feedback and votes of classification sequences for this object. The *LODB* of each user will then reveal a similar pattern at row *sID* that corresponds to the object. If this information can be shared with users that have not yet approached the object, it will help to adapt and improve the classification for this object before the users approach them. In the following, we will explain how this shared information can be used for adapting the local classification process, while section 6.2 goes into more detail about how the synchronization takes place.

For now, let us assume that an ad-hoc network connection can be established between phones that are located within signal range. In addition to the *LODB*, each phone stores a global object database (*GODB*) that contains the information gathered from other phones and from local classification trials. The *GODB* has the same structure as the *LODB*. Each phone's *GODB* will be updated with the local information stored in the *LODB*, but it will also be synchronized with information stored in the *GODB*s of other phones. How these updates and synchronization steps are realized will be explained in section 6.2. The synchronized *GODB*s that contain classification and feedback information from multiple phones and users allow adapting and improving the local classification process.

Again, a classification sequence *r* is the result after a new recognition attempt. Instead of relying exclusively on *r(1)*, as the result suggested by the neural network, the information stored in the *GODB* is also considered in a second classification step. We perform a nearest neighbor classification between the vote sequence that is derived from *r* and the row vectors stored in the *GODB*. Note, that before finding the nearest neighbor, the order of the entries in *o* has to be rearranged and normalized. This is necessary to match them with the order stored in the *GODB*. The rearranged sequence is denoted with *o'*, and is used for nearest neighbor classification instead of *o*. Finally, the ID that corresponds to the computed nearest neighbor represents the overall classification result. Figure 6.1 illustrates an example, where the initial classification through the neural network would fail. Only the additional comparison with the *GODB* in the second step leads to a correct result. As explained earlier, the unchanged voting sequence *o* is time-weighted with f_t^2 and added to the *LODB*. The next section describes how the *GODB*s are updated and synchronized.

6.2 Synchronization

As mentioned in section 6.1, the *GODB*s are synchronized among phones as soon as they are within signal range. The ad-hoc network is highly dynamic as visitors are moving continuously through the museum. Thus, we can only transfer data between two directly connected phones without routing. Indirect connections over multiple phones would be interrupted frequently and would therefore not be stable. In section 6.3, we explain how the transmission of the *GODB*s is realized in our framework.

If a connection is established, we carry out the following synchronization steps for each *GODB* row on both sides: As can be seen in figure 6.2, the number of samples that have been accumulated in all rows of all *GODB*s and *LODB*s are recorded in sample counters. Comparing each corresponding row in the *GODB*s of both connected phones, the row with the largest sample counter is selected. This row is first updated with the corresponding row of the local *LODB*. The *update* operation adds all entries of the *LODB* row to the entries of the *GODB* row –including the sample counter– and then resets all *LODB* row entries to 0 (including the sample counter). As explained in section 6.1, new samples can be added to the *LODB* row via new recognition tasks.
After updating the *GODB* row, it will be sent to the other phone, and having it received there it will *replace* completely the existing *GODB* row with the same *sID* (including the sample counter, as in the previous step). Then, a second *update* with the local *LODB* row on this side is carried out. The result is sent back and replaces the corresponding *GODB* row on the other side.
This synchronization sequence ensures that no classification and feedback information is considered more than once on the same phone. Otherwise, the data transmitted over multiple hops and at some point received by its originator would be incorrectly overemphasized, and lead to incorrectly weighted classification results. Therefore, loops are avoided in our ad-hoc network communication.

The synchronization steps are repeated row by row, until the full *GODB* is synchronized, or the connection has been lost. It is started again, as soon as a new connection can be established –either with the same phone or with another phone. Note, that rows are only synchronized if a higher sample count on either one side exists. This indicates that one side has more reliable results. If the sample counts are equal, no synchronization is triggered. This would also be the case if a new connection between two already synchronized phones will be established again. Time-weighting the vote sequences ensures that more recent classification approaches are up-weighted, while outdated information is down-weighted. Therefore, each *GODB* represents the most up-to-date classification state.

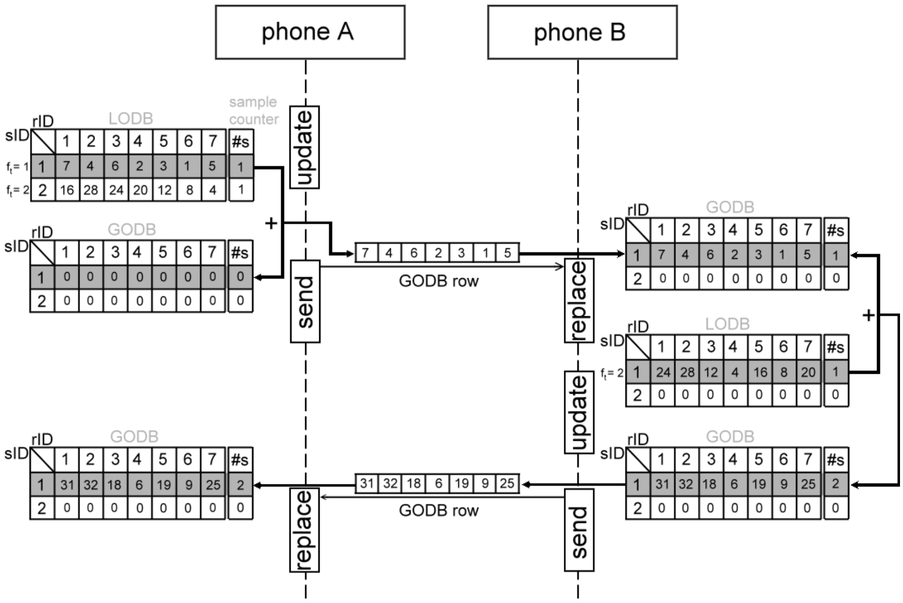


Figure 6.2: Steps for synchronizing one GODB row: If phone A synchronizes with phone B, A's GODB row has to be updated before being transmitted. After receiving it at B, this row replaces the existing row in B's GODB. Then, B has to perform an update and sends the result back to A, where it replaces the corresponding GODB row. After doing this for each row, the GODBs of A and B are synchronized and contain the most up-to-date information.

Another solution instead of merging the vote sequences of each object would be to store each vote sequence individually in order to provide a larger and more distinct database. However, this would not only continuously increase the classification time but also the data that has to be exchanged. With our technique, the *GODB*'s sizes remain constant on all phones.

Figure 6.3 shows an example of how locally connected information is propagated through ad-hoc network connections.

6.3 Implementation

In this section we will describe the practical challenges that arise when implementing an ad-hoc phone-to-phone network. In our implementation, we apply Bluetooth for wireless communication since it is widely established and integrated in most mobile phones.

In general, three steps are carried out to transfer data via Bluetooth between two devices: First, a scan for nearby devices has to be performed (*inquiry*). Second, for each of the detected devices, a

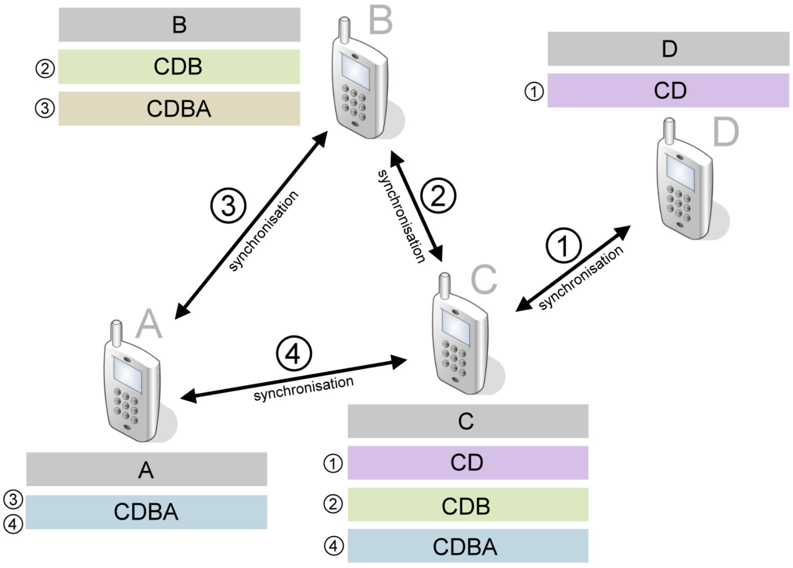


Figure 6.3: Propagation example of classification and feedback data in a mobile ad-hoc network (sequence numbers encircled): Initially, all phones store their local LODBs only. First, phone C synchronizes with phone D (1), then C with B (2), then A with B (3), and finally A with C (4). After these steps, A, B, C share the same and most up-to-date information that have been collected by all phones. Phone D would also be up-to-date if it would now synchronize with either A,B, or C.

service search must be executed in order to exchange connection and service parameters. Finally, the *connection* is established.

Although Bluetooth is very common, several practical drawbacks arise for ad-hoc networks. They need to be solved or bypassed for our purpose. For instance, during *inquiry*, the phone enters the internal *inquiry substate*. In this state, it continuously transmits an ID package at different hop frequencies and collects the Bluetooth device addresses of all devices that respond to the inquiry message. During this time, the phone cannot be detected by other phones until it leaves this substate and enters the *inquiry scan substate*. To ensure that not all devices enter the same substates synchronously, and therefore never detect each other, we introduce a random waiting time period t_w between both substates as proposed by [147]. Empirically we found that t_w=5-9 seconds is optimal for avoiding continuous deadlocks. With this additional waiting time, we can estimate the duration D that is required for establishing an ad-hoc connection between n_d phones:

$$D \approx n_i(n_d t_i) + (n_i - 1)t_w + n_d(t_s + t_r), \tag{6.1}$$

where t_i is the inquiry time that is proportional to the number of devices within proximity and their distances. If devices can be detected, t_i is approximately 1-5 seconds per device (for Java ME) but

is not lower than 10.24 seconds [131] – even if no devices can be detected. The parameter t_s is the times required for carrying out one service search (on average 6 seconds per device). Finally, t_r is the time required for transmitting the *GODB*s and service parameters in both directions. Should no devices be detected during the first inquiry (but are in signal range), we have to repeat this step on average n_i=1-4 times, with a delay of t_w seconds.
The transmission time t_r depends on the amount of synchronization data that has to be exchanged. In our case, the *GODB* is of the size *N* x *N* x 4 bytes (with *N* being the number of trained objects), and the list of sample counters is of the size *N* x 4 bytes. With a measured average transfer rate of $\sim$40 kByte/s for Bluetooth 1.1 (specified with 732,2 kbit/s) we synchronize two devices with $t_r < 1$ second, when assuming that $N = 50$ or less. These parameters were measured for a Nokia 6630 mobile phone with Bluetooth 1.1 (cf. figure 3.4). Phones equipped with Bluetooth 2.0 (specified with 2.1 Mbit/s) would allow much faster rates. In practice, we found that *D* ranges from 65-120 seconds for synchronizing three phones, for example.

This communication process runs in parallel with the localization procedure mentioned in section 3.2.3. Since the hardware addresses of the Bluetooth beacons are known, a preselection of detected Bluetooth devices is necessary. A phone-to-phone communication is only established for unknown devices.

6.4 Simulation

We have tested and validated our approach practically with three Nokia 6630 mobile phones. Although these experiments showed that image classification through synchronizations between the three devices improves and adapts to dynamic situations, the number of devices is far too low to represent realistic conditions. For proving the benefits of ad-hoc network adaptation for image classification, we have developed a simulator application. This tool simulates a stream of visitors in a museum over time. All parameters that are used for simulation (such as a floor plan and exhibits of a museum, visitor behavior, lighting conditions, and Bluetooth signal range) have been investigated and measured in advance in the City Museum of Weimar to guarantee realistic results. This simulator gives us the opportunity to evaluate different configurations and user scenarios, which are currently not possible to achieve in this extent with real users.

Figure 6.4 illustrates a screenshot of a museum that we have modeled with our simulation tool. The floor plan and the locations of the exhibited objects have been measured in the museum. For each of the 50 objects (from a total of 116 objects being displayed on the entire floor) that are located

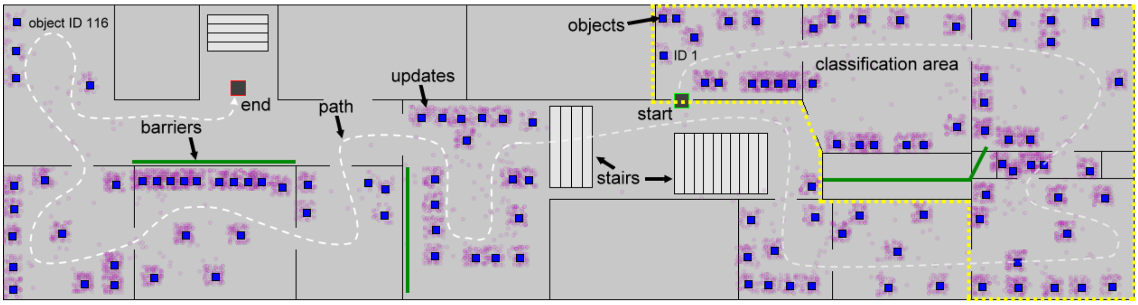
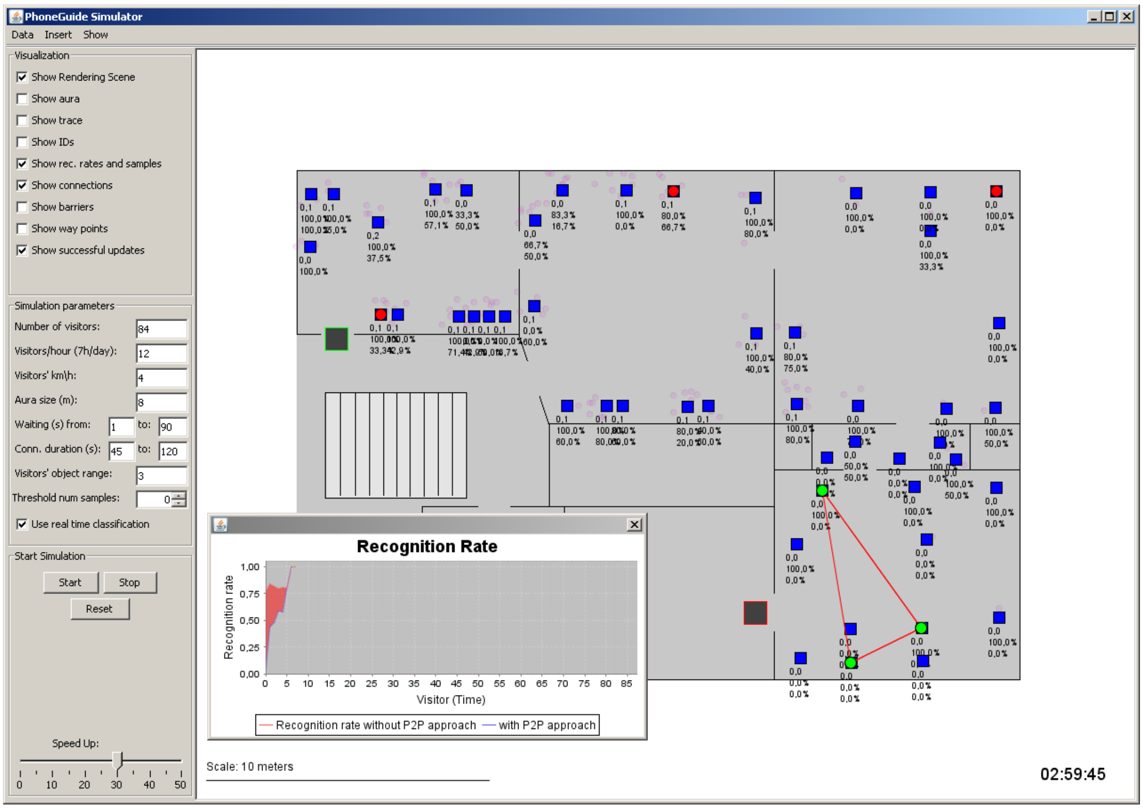


Figure 6.4: Simulation screenshot: The floor plan of the City Museum of Weimar with the locations of the exhibited objects (blue boxes). For each successful ad-hoc synchronization during simulation, the locations of both phones have been tagged (red dots). The majority of successful data exchanges are in the proximities of closely located objects, since visitors gather together in these areas for a longer time. The solid green lines define barriers that do not transmit the Bluetooth radio signal. The dotted yellow frame outlines the area in which we tested our approach with real image data and under realistic illumination conditions.

in our test area (framed with dotted yellow lines) that we consider for our simulation, we captured a pool of 200 images per object (160x120 pixels) from multiple perspectives and distances at two different time of days. Half of these images were used for an initial training of the neural networks while the other half was used for the simulation itself. To approximate the recognition process properly, the simulation tool is connected with the Java ME wireless toolkit emulator through a local TCP/IP connection. This emulator executes the PhoneGuide front end application in exactly the same way as it would be executed on the mobile device. Taking photographs of an object for recognition is simulated by randomly selecting one entity of the corresponding object's image pool that was not used for initial training. The user's feedback is also simulated by always picking the correct object ID after classification. As mentioned in section 3.2.3 the signal range of a class 2 Bluetooth emitter (that is used in mobile phones) is defined to be 10 m. In order to account for reflections or absorptions of the signal we set it to 8 m. Furthermore, we measured the transmission of radio signals through walls in the museum. These measurements are also taken into account in the simulation: Thick walls that did not transmit the signal are marked with signal barriers. Consequently, blocking and transmitting room elements are considered during the simulation, but complex reflections of the radio signals are not. The user interface of the simulator in combination with an on-going simulation is illustrated in figure 6.5.

The museum that we have modeled consists of many small to mid-sized rooms. Each room shows different exhibits which thematically depend on the objects presented in previous and following



☑ Show Rendering Scene
□ Show aura
□ Show trace
□ Show IDs
☑ Show rec. rates and samples
☑ Show connections
□ Show barriers
□ Show way points
☑ Show successful updates

Simulation parameters

Number of visitors: 84
Visitors/hour (7h/day): 12
Visitors' km\h: 4
Aura size (m): 8
Waiting (s) from: 1 to: 90
Conn. duration (s): 45 to: 120
Visitors' object range: 3
Threshold num samples: 0
☑ Use real time classification

Figure 6.5: User interface of our simulation application. Left: the necessary parameters to configure the simulation. Right: On-going simulation. Bottom: real-time recording of the classification rate.

rooms (the museum shows the evolution of a city in time). Although they were able to move freely, we have observed that visitors follow a main path that leads them through different exhibition contexts. Consequently we simulated to stream of visitors comparable to this path. On this rough path, however, the simulated visitors are free to move randomly to arbitrary exhibits (e.g., that are located in the same room) and can skip or move back to objects. In the simulation, visitors enter the museum at random times in between the opening hours. We did not find particular peak hours in case of the City Museum of Weimar. The speed with which visitors walk from exhibit to exhibit, as well as the examination time for each object is also selected randomly within the range of our observations.

The synchronization between two phones is simulated by a scheduler that continuously triggers an inquiry and service search for each simulated visitor, as explained in section 6.3. Whether or not a connection can be established and a synchronization is successful depends on the visitors' movements and on the time that is required for the synchronization which we derived from equation

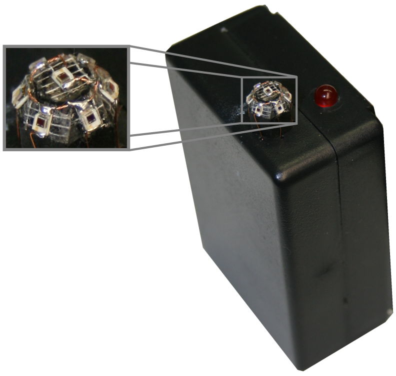


Figure 6.6: Seven hemispherically aligned light sensors with internal memory to record the illumination state over multiple days.

6.1 and the measured parameters explained in section 6.3.

We have simulated dynamic environmental changes with measured light probes that have been recorded within our real test area during the opening hours of one day. For this, we have developed sensor boxes with small, battery powered omnidirectional light sensors and internal memory as illustrated in figure 6.6. The recorded changes of environment lighting within the 7 opening hours are used to generate a smoothed luminance function as plotted in figure 6.7. For each simulated recognition task, the luminance of the pre-captured test images is scaled accordingly to the measured luminance at the time of classification. To achieve this, we roughly matched the transfer function of the phone's integrated camera with the transfer function of the light sensor through a calibration phase.

6.5 Evaluation

With our simulation, we have evaluated two different scenarios to prove that an ad-hoc synchronization adapts to dynamic changes and therefore improves the image classification.

The first scenario shows the development of the recognition rate over time for a rapidly changing illumination. As mentioned above, the lighting changes are caused by the increasing sunlight that has been measured at these times within our test area of the museum. The result is plotted in figure 6.7. We simulated 84 and 252 visitors (12 and 36 visitors per hour, entering the museum randomly) over the opening hours of one day. Note, that 12 visitors per hour equals the average number of visitors per day for a German museum, as it is reported in [1]. The average walking speed was assumed to be 4 km\h, and the average examination time was set to a range between

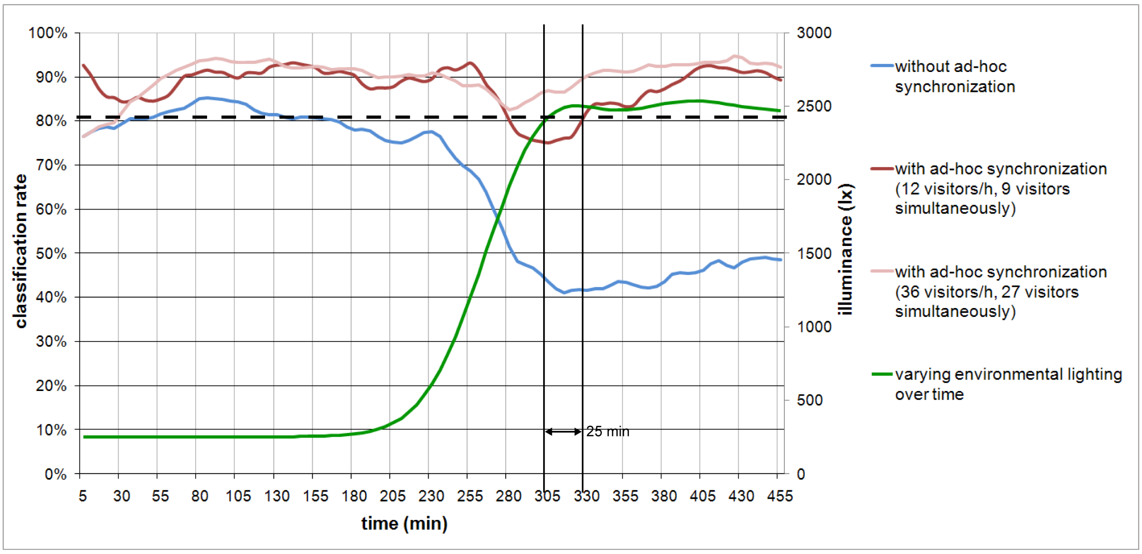


Figure 6.7: Average recognition rates for an unadapted (blue) image classification, and an image classification that is adapted through ad-hoc synchronization (dark red: 12 visitors/h and light red: 36 visitors/h), simulated over the opening hours of the museum under measured lighting changes (green). The adapted classification always performs better than the non-adapted classification, is less fragile to dynamic environmental changes, and recovers quickly to sudden changes.

1 second (visitor just moves on immediately after classification) and 90 seconds (visitor listens to the multimedia information after classification).

Figure 6.7 presents the average recognition rates of all visitors at a particular time that can be achieved with (light/dark red) or without (blue) adapting the image classification through our ad-hoc synchronization approach. As it can be seen, the illumination changes rapidly (due to sunlight) after 210 minutes past the museum opening, and stabilizes after 305 minutes. In this case, the recognition rate of the non-adapted classification process drops from ∼80% to ∼45%. The ad-hoc synchronization, however, leads to a higher overall recognition rate and to a quick adaptation to environmental changes, such as the lighting conditions in this example. The adapted classification drops relatively little, and can recover after a short time (in this example: after roughly 25 minutes the recognition rate for 12 visitors/h reaches the original level of the non-adapted classification, and even improves further). The classification performance after the completed adaptation is around ∼45% higher than the performance of non-adapted classification under the same condition. In addition figure 6.7 shows that the more visitors are available the fatser is the adaptation process.

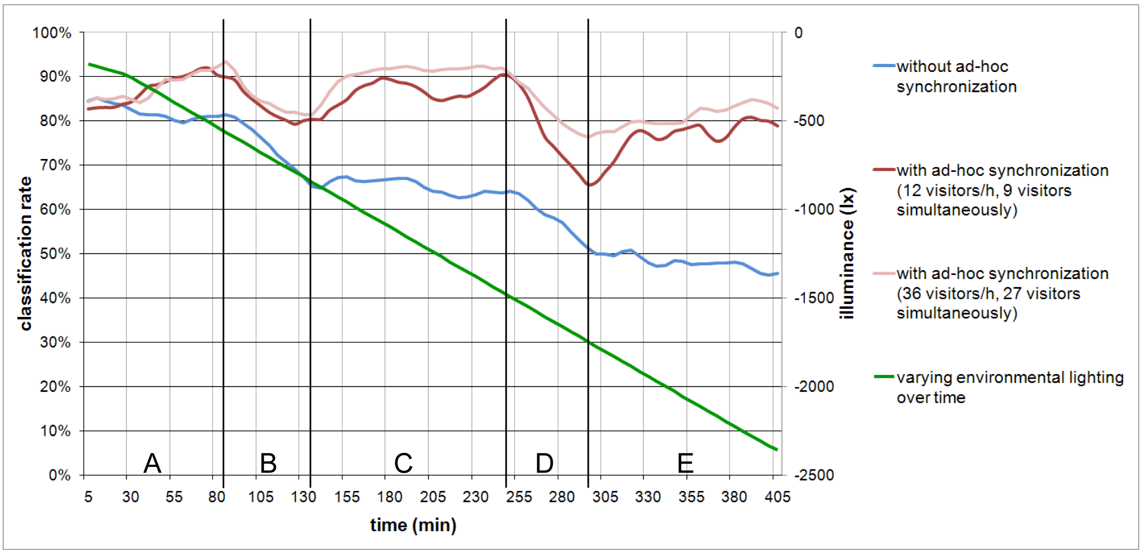


Figure 6.8: Same as in figure 6.7, but with constantly decreasing illumination. While the recognition rate of the non-adapted classification decreases, the recognition rate of the adapted classification drops too – but is always better. If the performance of the non-adapted classification stagnates, the recognition rate of the adapted classification improves through ad-hoc synchronization.

In the second scenario, we investigate how our approach performs for non-abrupt, but slight and continuous changes, such as a constant decrease in illumination. This is shown in figure 6.8. For this experiment, we apply a synthetic illumination curve rather than true measurements. With linearly decreasing illumination, the recognition rate of the non-adapted classification decreases in intervals. The reason for this seems to be the specialization of the neural networks weights which display different sensitivity to varying inputs. The behavior of the adapted classification is correlated to the behavior of the neural network. If the recognition rate of the neural network decreases (sections B and D in figure 6.8), the recognition rate of the adapted classification decreases too. However, the adapted classification is always better than the non-adapted classification. If the recognition rate of the neural network stagnates (sections A, C, and E in figure 6.8), the adapted classification improves through ad-hoc synchronization.

One other observation that can be made when investigating the distribution of successful synchronizations, as visualized in figure 6.4 is that most of them appear at locations with a high object density. The reason for this is that at these places, visitors will remain for a longer time within signal range, and synchronizations become more likely. But on the other hand, this also implies that objects which are located more isolated from others will cause lower recognition rates. This is the case if all visitors follow a similar path through the museum. In this situation, synchronizations will only be beneficial for quick adaptations if they are performed before the actual recognition task is carried out. This is more likely for areas with a dense object distribution, than for areas with a sparse one.

In section 8.3 we will show how this approach influences the overall classification rate of our entire system.

6.6 Discussion

In this chapter we have proposed an adaptation technique that exchanges classification data among phones through ad-hoc phone-to-phone communication. We have shown that this method improves the overall recognition rate of local image classification algorithms and adapts them to dynamic environmental changes, such as varying lighting conditions. Furthermore, we have explained how adaptation and synchronization can be realized and have analyzed how they perform on Bluetooth-equipped mobile phones. Since it is difficult to validate our method with a large number of real devices and users we have proven our approach in a simulation under as realistic conditions as possible. Concrete experiments in a smaller scale (with three devices) have confirmed the simulation results.

The most limiting factor of our current implementation is the relatively long response time. As can be seen in figures 6.7 and 6.8, the network requires approximately 25-45 minutes for full recovery –depending on how far it is decreased and how many visitors are available. This is mainly due to Bluetooth-related performance limitations (cf. section 6.3). If the link between two devices could be established more quickly, the update rate would be higher and consequently the adaptation faster. Future versions of the Bluetooth protocol will allow connections to be established in a matter of milliseconds which would solve this problem [78].

A drawback of the evaluation is the lack of an extensive field study. Although we have tried to determine as many parameters as possible from real-world measurements for our simulation, the result is not a substitute for practical field testing. The main reason for our decision to simulate our approach, besides the lack of a large number of phones for a field study, is the unpredictable and erroneous behavior of Bluetooth. Since Bluetooth is designed to be a data transfer protocol it behaves erratically when ”misused” as a localization technique. For example, if mobile phones continuously scan for nearby devices and/or services, performance dropouts start to occur regularly, they stop inquiring for a certain period of time or crash completely. Comparable issues were reported by [185, 146]. Again, newer versions of the Bluetooth protocol will hopefully resolve these issues and we believe that the observations and outcome of a user study would be comparable to our simulation results.

7 Adaptation through Pathway Analysis

As already pointed out in the related work chapter, the use of location information significantly improves the classification performance of mobile vision-based information systems. Whether indoors or outdoors, RF-technologies such as WLAN, RFID, Bluetooth or GPS and GSM-localization can be applied to approximate the location of users. Based on their position, only the objects in the vicinity are considered as potential result candidates for an object recognition task. In section 3.2.3 we have shown how Bluetooth emitters distributed in a museum have improved the classification performance of PhoneGuide in the past.

However, there are several practical challenges and drawbacks that occur when using localization technology as part of a mobile guidance system in museums. First of all, museums have to invest in the hardware to equip extensive areas of their buildings with RF-technology. Additionally, it requires maintenance by appropriately trained staff.
Furthermore, as we have experienced during the evaluation of our Bluetooth tracking technique, continuous scanning for RF-transmitters significantly increases the mobile phone's energy consumption. In combination with an activated camera, the battery of an ordinary mobile phone runs low after 1-2 hours of use. Furthermore, Bluetooth in particular seems not to be designed for continuous location determination since its scanning mechanism is time-consuming and error-prone as mentioned in section 6.6. All of these drawbacks motivated us to find a more unobtrusive approach that supports image classification through location information but overcomes the limitations presented above.

In the 1980s, Veron and Levasseur [179] conducted a field study to examine user behavior in artistic environments. As a result they identified four different visitor categories that were based on their paths, movements, and observations during the time spent in museums. For example, one category is the ”ant”-visitor who observes every exhibit in the museum and follows the path proposed by the curator, while, for instance, the ”grasshopper” visitor only looks at the artworks they are interested in – without following a predefined route.
Based on these findings, we wanted to develop a technique that takes advantage of the fact that the pathways of different museum visitors are similar. Consequently, if we observe the movement pattern of a certain visitor during his stay in the museum, we can compare this with previous pathway observations in order to predict the visitor's future locations. The model of four different pathway categories, however, seems to be too simple to cover all the different types of visitor movement patterns. As pointed out by Gabrielli et al. [73] the visitors' ”activity is not structured;

they move in the physical space following their interests and preferences and exploiting the real and perceived affordance of the environment". Consequently, user modeling needs to be more sophisticated.

In the following we will describe two different adaptation techniques for improving classification performance of our object recognition algorithm by interpreting spatio-temporal data of museum visitors.
The first technique applies gathered pathway data –that is the sequence of exhibits previously photographed by individual visitors– to predict the sequence of objects captured by subsequent visitors. This enables us to combine the result of our image-based classification with the pathway prediction analysis, after taking a photograph.
The second technique examines spatio-temporal data in order to determine areas in the museum which users do not return to after they have left them. These areas can be disregarded for subsequent image classifications after the visitor has progressed past them. Consequently, the number of possible candidates decreases continuously as the visitor moves through the museum.

The research presented in this chapter is based on our publication "Localization and Classification through Adaptive Pathway Analysis" [33].

7.1 Data Acquisition

To analyze pathways of museum visitors for localization and classification, we first needed to collect the necessary spatio-temporal data. For this we carried out a user evaluation in the Natural History Museum Erfurt in Germany by examining three floors of the museum. Each floor has a size of approximately 10x20 m, and has no separating rooms or walls. The visitors were able to walk freely among the exhibits and were not limited by architectural constraints. This was the main criterion for selecting this museum instead of the City Museum in Weimar where we conducted our previous user studies and where mobility is more limited. Furthermore, museum attendance in Erfurt was higher which was beneficial for data collection. The floors are connected by stairways and each floor has only one entrance and one exit. The exhibits are mainly located in closed showcases and are of different sizes, ranging from small dissected birds to wild boars and entire trees (cf. figure 7.1).

To collect data for our experiment, we asked regular visitors to carry around a mobile phone (running a photo capture application) during their visit. They were requested to photograph each exhibit they found interesting and wanted to know more about (see appendix A.3 for the description of the

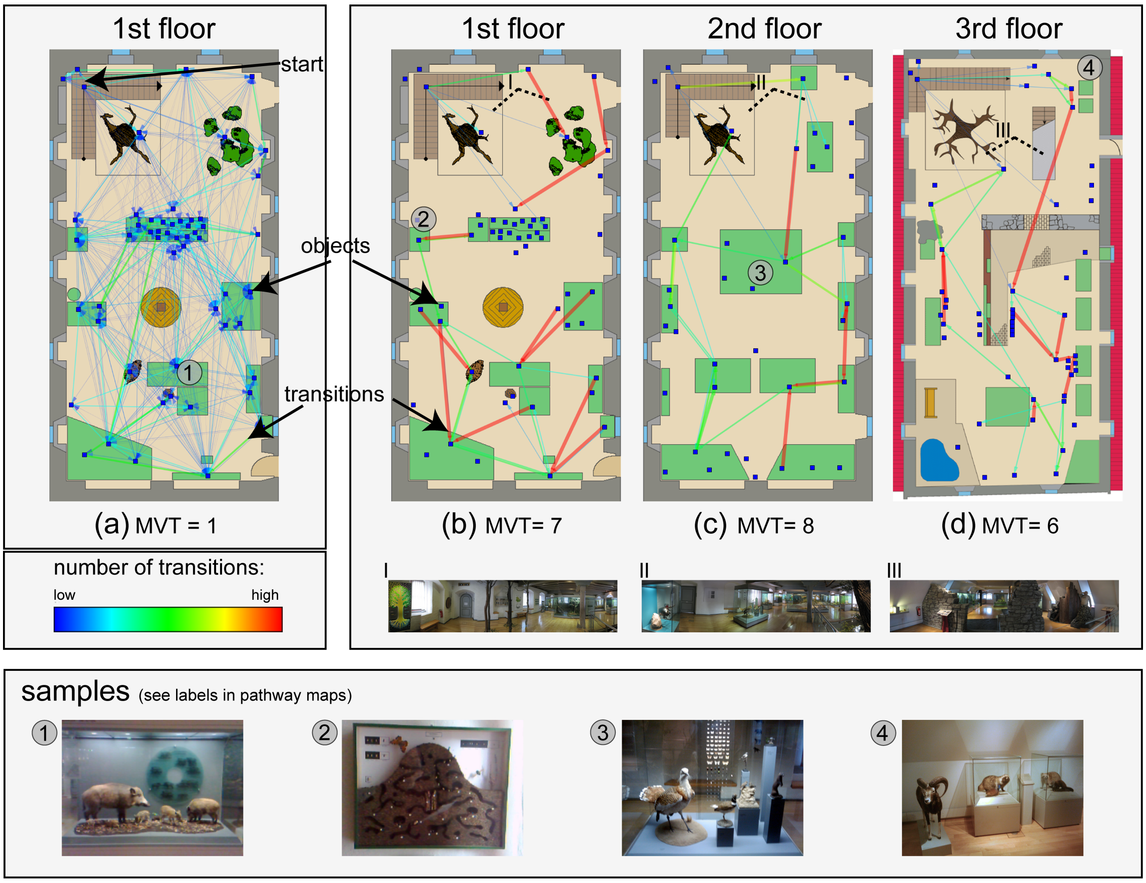


Figure 7.1: Color-coded pathway visualization of 132 museum visitors in the Natural History Museum Erfurt, Germany. Each transition between two objects is visualized for the first floor (a). The main routes for all floors are visualized in b-d. MVT denotes the minimum visitor transitions. Edges are visualized if the number of visitor transitions between the corresponding objects is larger or equal than MVT. Several sample exhibits are illustrated on the right.

experiment). Each captured image was stored locally on the device in combination with a timestamp. At the end of their visit, they returned the phones and the images were transferred to a server for identifying the photographed exhibits manually. We collected a total of 2,926 images over a period of 4 months.

In order to evaluate the results under controlled conditions we classified the images manually during the data collection phase of this experiment. In practice this data would normally be collected during the application of PhoneGuide as explained in section 3.2.2. Besides the definition of the pathways, this field study also revealed the exhibits of interest to the visitors. Since we did not influence or continuously observe the visitors during their stay we retrieved a natural selection of meaningful objects that we could use for our evaluation. From our point of view, this is the most realistic way of choosing the reference data for an object recognition task.

In addition, we extended our simulation application introduced in section 6.4. Using the application we modeled the different floors of the museum as well as the location of the corresponding exhibits true to scale. By extracting and comparing the locations of the individual exhibits, we generated a distance map of the museum that stores the distances between each object pair. Furthermore we visualized the collected pathway data and identified the main visitor routes as shown in figure 7.1.

7.2 Generation of the Pathway Classifier

To accomplish the prediction of future exhibits, we generate a classifier based on the collected pathway data. In the following sections, we will call this *pathway classifier* which complements the *image classifier* we introduced in section 3.2.1.

The generation of the pathway classifier can be separated into two parts: The creation of a distinct description of the pathways and the training of the classifier itself.

In general, approaches that apply models for predicting locations consider only a small number of different location possibilities. For example, many approaches try to predict user movements in buildings [180] for which the number of locations (rooms) is small. This means, that each possible first-order transition (i.e., a pathway with a history length of 1) between each pair of locations can be obtained through user evaluation in an adequate amount of time. Consequently, one easy way to predict the next visited object is to count the visitors that are moving from one exhibit to any other exhibit and compute the corresponding probability e.g., with a Markov predictor.

However, in museum environments with hundreds of exhibits, the collection of every transition would be time-consuming and would require user evaluations spanning several years to ensure a stable prediction for each transition. For instance, our user evaluation revealed that $N = 146$ different object locations were approached. This would result in $N^2 = 21,316$ different first-order transitions for a brute-force method. To ensure a comprehensive basis for prediction, each of these transitions therefore needs to be approached by multiple visitors during the training phase. Without this, transitions that have rarely been approached by visitors or not at all, cannot be predicted properly.

Fortunately, it is not mandatory to determine the probabilities of each object transition individually as long as the corresponding locations of the objects are described sufficiently (which we achieve using distance maps as explained earlier). For example, if two exhibits A and B are located in the

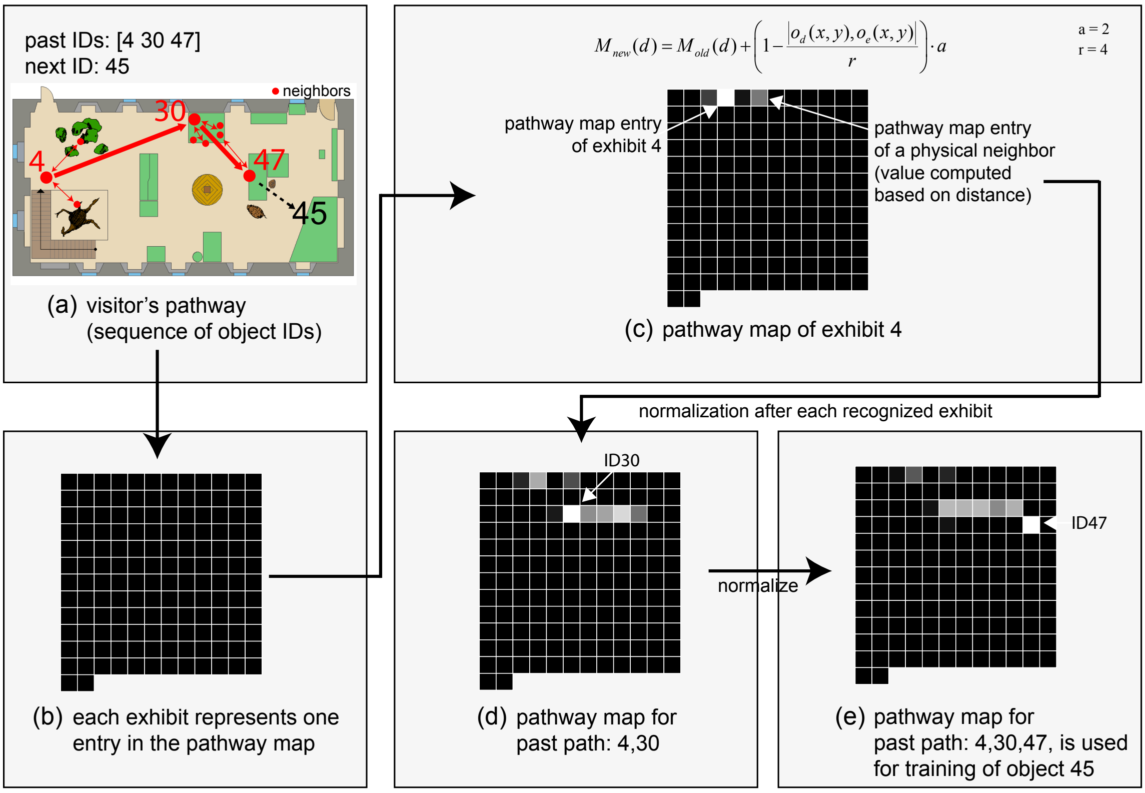

past IDs: [4 30 47]
next ID: 45
neighbors
30
4
47
45
(a) visitor's pathway (sequence of object IDs)
$M_{new}(d) = M_{old}(d) + \left(1 - \frac{|o_d(x,y), o_e(x,y)|}{r}\right) \cdot a$
a = 2
r = 4
pathway map entry of exhibit 4
pathway map entry of a physical neighbor (value computed based on distance)
(c) pathway map of exhibit 4
normalization after each recognized exhibit
(b) each exhibit represents one entry in the pathway map
ID30
normalize
(d) pathway map for past path: 4,30
ID47
(e) pathway map for past path: 4,30,47, is used for training of object 45

Figure 7.2: Algorithm for translating pathways (a) into a 2D map representation (b). Each entry of the pathway map represents one exhibit. Their weights reflect the current segment of a visitor's path. The highest weight is assigned to the current (captured) object, and gradually decreased values indicate neighbor objects within range (c). To decode the order of visited exhibits, the pathway maps are normalized in such a way that the weights of previously captured objects are attenuated over time (d-e).

same showcase and exhibit C is in another showcase, then it can be assumed that there is only a minor difference in the pathway if a visitor has photographed A *or* B before she moves to C. Hence we describe pathways that include co-located objects similarly. This makes it possible to predict the correct exhibit even if the corresponding past pathway has never been traced before.

Figure 7.2 shows an example of how the pathway representations are generated. For the case illustrated in figure 7.2a a visitor moves from exhibit 4 to 30 and finally to exhibit 47. Each exhibit is represented as an entry in a table that we call pathway map (cf. figure 7.2b). At the beginning of a museum visit this map is empty and initialized with 0. After the visitor has photographed the first object (in this case exhibit 4) the entries $M(d)$ of the map M are updated by means of equation 7.1:

$$M_{new}(d) = M_{old}(d) + m(d) \quad ,with$$
$$m(d) = \begin{cases} 0 & if\,|o_d, o_e| > r \\ \left(1 - \frac{|o_d(x,y), o_e(x,y)|}{r}\right) \cdot a & otherwise \end{cases} \tag{7.1}$$

$M_{new}(d)$ denotes the new entry representing exhibit d, with $d = 1..N$. It is derived from the physical distance between the photographed object o_d and all other objects o_e with $e = 1..N$ divided by a predefined range r. The range denotes the maximum area of influence of the photographed exhibit to its neighbors and is determined empirically. The higher r, the more entries are affected and the more similar the maps of different pathways appear. The distances are retrieved through the distance map as stated in section 7.1. The distance-based weighting is required to ensure that objects which are located in close proximity lead to similar pathway maps.

The pathway map that results after object 4 was photographed is illustrated in figure 7.2c. Entry 4 has value 1 since the distance from object 4 to itself is zero. Entry 3 and 6 have non-zero values because they lie within range r. All remaining objects were not affected since they are located too far away.

Before the pathway map for the next object on the path is updated, the values are normalized. This is evident since it decodes the sequence of visited exhibits as well as the number of passed exhibits that are stored in the pathway map. It is comparable to the definition of history length in related approaches [180]. This property can be controlled by an attenuation factor a. The higher this factor (e.g., $a = 2$ in figure 7.2), the faster the entries in the pathway maps converge to zero with an increasing path length.

This procedure is carried out for all pathways collected through our user evaluation. For each object that has been photographed, the corresponding pathway map based on previously visited exhibits is stored (e.g., the pathway map in figure 7.2c is assigned to object 30; the pathway map in figure 7.2d is assigned to ID 47 and so on). These maps are re-arranged in N-dimensional vectors that serve as training samples for a 3-layer neural network that we call the pathway classifier.

Especially for mobile devices, the application of neural networks as classifiers is beneficial since they are small in size and do not increase with an increasing number of training samples. In addition, the neural networks classification process is not time-consuming.

7.3 Ensemble Classification

The pathway classifier as well as the image classifier are trained on our server. As explained before, photographs of exhibits taken by visitors serve as input for the image classifier. For the pathway classifier, the pathway map computed based on the visitors previous path, is used as input. Both classifiers output a matching probability for each object. These probabilities are combined through average voting [185] with:

$$P(d) = \frac{p_i(d) + w \cdot (p_p(d) + p_a(d))}{1 + 2w} \tag{7.2}$$

The probability $P(d)$ that exhibit d was photographed is computed by summing the probabilities of the image classifier $p_i(d)$, the pathway classifier $p_p(d)$ and the a priori probability $p_a(d)$. In our context the a priori probability ensures that frequently photographed exhibits are more likely to be under the final candidates than less frequently photographed ones. The pathway elements are weighted empirically by w ($= 0.5$, in our experiments) since in general the image classifier is more reliable than the pathway classifier. We determined this weight empirically by conducting a test series to find a value for w that maximizes the classification rate of a preselected test set.

7.4 Segmentation through Pathways

Besides using the pathways for predicting which exhibits museum visitors will approach next, we have developed a second technique that further reduces the number of potential result candidates. Here we assume that there are areas in a museum that users do not return to after they have left them. Possible reasons include, for example, that they are not interested in looking at exhibits several times or that the exhibition has to be visited on a predefined path.

To determine these areas, we translate the collected pathway data of N nodes into a weighted directed graph $G(V,E)$, comparable to [111]. Our goal is to find weakly connected subgraphs $G'(V,E)$ (we refer to them as *pathway clusters*) where the edges that interconnect with these subgraphs are directed either inwards *or* outwards. To detect these pathway clusters we propose an approach that first carries out a depth-first search, starting at each node V (object) of $G(V,E)$. Only successive nodes of V with a corresponding edge weight $w > t_w$ ($w = 1..U$, where U denotes the maximum number of visitors that move between two objects) are considered. To avoid outliers and consequently prevent cycles, with t_w ($t_w = 1..U$), we specify that a minimum number of visitors must have approached the same transition to be valid. We decided to choose $t_w = 1$.

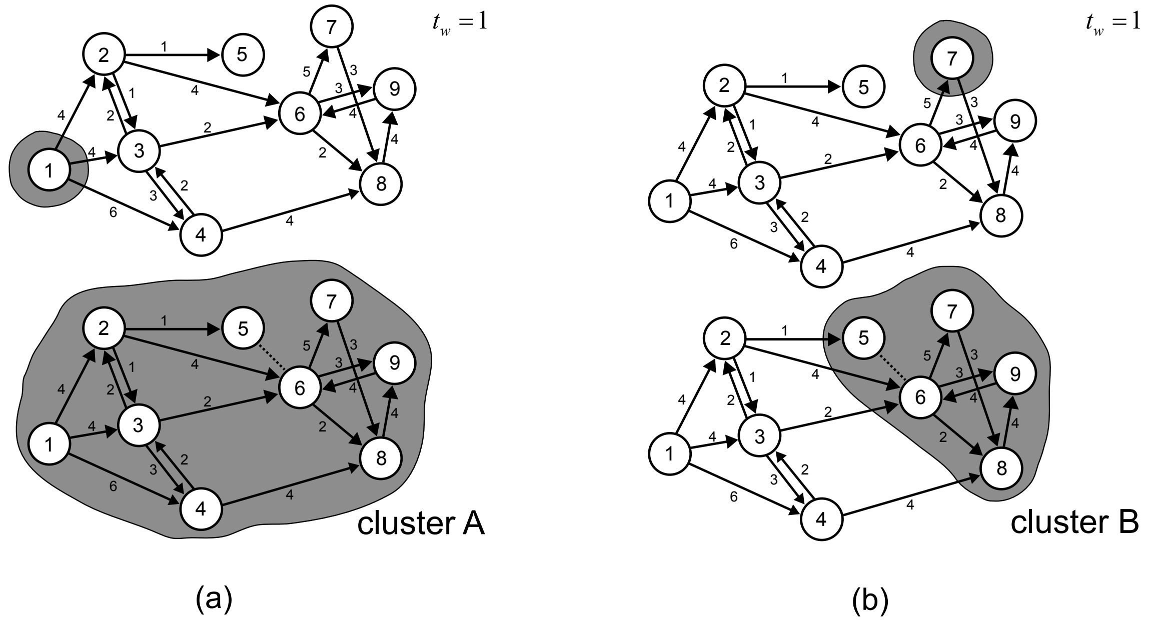


Figure 7.3: Illustration of pathway clustering by applying a depth-first search for two different starting nodes (a, b).

All depth-first searches that start from each node form one single subgraph $G'(V,E)$. Many of these N subgraphs are redundant. If the number of identical entities of the same subgraph is above a threshold t_c ($t_c = 1..N$), this subgraph is strong enough to define one pathway cluster. It can be determined automatically by minimizing the number of almost identical clusters for an increasing threshold. For the data that was acquired in the context of the field study explained in section 7.5, we found that $t_c = 10$. In this case the pathway cluster has been created from 11 different starting nodes. This was the smallest number of nodes where the algorithm did not generate clusters that differ by only one or two nodes. If $t_w > 0$ the unclustered nodes are assigned to all pathway clusters that have an equal physical distance to the corresponding node.

An example of the pathway clustering is illustrated in figure 7.3a: When starting with nodes that represent the beginning of a path (e.g., object 1), the resulting subgraphs consist of all the exhibits, since the corresponding successors cover the entire exhibition (cluster A). Since $t_w = 1$, the edge between node 2 and 5 is not considered during clustering and is added to the closest clusters at the end of the clustering algorithm.

If the algorithm selects objects as starting points that are placed at more central locations of the exhibition, such as object 7 in figure 7.3b, the number of successors decreases because some nodes can no longer be reached. For instance, there is no direct edge from nodes 6, 7, 8, 9 to 1, 2, 3 or 4. Hence, cluster B is limited to five nodes, since in addition node 5 is assigned to the cluster as mentioned earlier. Consequently, after a visitor photographs either object 5, 6, 7, 8 or 9 the algorithm detects this cluster and selects the corresponding, cluster-individual image classifier that identifies only objects 5-9 for future object recognition tasks.

A more restrictive solution for this could be to consider only those objects for image classification that are direct successors of the last photographed exhibit. For example, if a visitor captured object 1, then for the next shot, only exhibits 2, 3 and 4 are considered as possible result candidates. We omitted this idea because the collected amount of pathway data do not allow a precise prediction of all future exhibits. If a possible transition between two objects has never been traced, the corresponding image classification would fail. However, as explained in section 3.2.2, we continuously collect pathway data during the use of our PhoneGuide system. If over a longer period of time no new transitions are detected, this approach could be beneficial.

7.5 Evaluation

Over a period of four months, we collected the pathway data of 132 regular museum visitors at the Natural History Museum Erfurt, Germany. The visitors (84 male and 48 female) were between the ages of 8 and 79 years (average: 36 years) with various educational backgrounds and interests. As mentioned earlier, they were requested to photograph all exhibits they found interesting as they walked completely unattended through the museum. In total, they captured 2,926 images (with a minimum of 2, a maximum of 58, and an average of 22 images per visitor) of 146 different exhibits.

We separated this data into a training and evaluation set. The training set (images taken by 117 users) was used to train the pathway classifier and to compute the pathway clusters, while the evaluation set (images taken by the remaining 15 visitors; as mentioned in the context of previous field studies, we refer to them as inexperienced users because they took pictures arbitrarily from different perspectives and scales) was used to determine the classification performance offline. Note, that the entire classification process also runs adequately on mobile phones with an execution time of ∼1.3s on a Nokia N95 in Java (cf. figure 3.4). Due to a lack of multimedia data, we decided to evaluate the classification results offline –but using real user data. The image classifier was trained by recording short video sequences (160 frames, 240x180 pixels) of every exhibit from different perspectives and scales and by computing 100 global color features (three 30-bin histograms for each color channel, mean and variances of each frame as described in section 3.2.1). The size of the image classifiers was 415 kB and of the pathway classifier 100 kB.

approach	image classification only	image classification with BT localization	image classification with pathway segmentation
inexperienced users	50.67	73.99	60.59
experienced user	65.68	80.16	72.39
inexperienced user (best of 3)	66.22	84.18	73.73

approach	image classification and pathway classification	image+pathway classification with segmentation	image+pathway classification with BT localization
inexperienced users	64.34	71.31	78.82
experienced user	74.53	78.02	87.67
inexperienced user (best of 3)	79.36	83.38	87.67

Figure 7.4: Classification rates for different approaches: image classification only, image classification and Bluetooth (BT) localization, image classification and localization through pathway clusters, image classification and pathway classification, image and pathway classification with localization via pathway clusters, image and pathway classification with Bluetooth localization.

Three pathway clusters were detected in our experiment by applying the clustering algorithm to the training set. The boundaries of these clusters correlate with the transitions of two museum floors. Visitors who entered the next floor generally did not return to already visited levels. Consequently, we have automatically trained three image classifiers: one containing 146 objects of all floors, one containing 90 objects of the second and the third floor, and one containing 51 objects of the third floor. If the algorithm detects that a visitor has left that particular floor, the corresponding neural network is selected that is specialized for recognizing the objects of the remaining floors.

It is likely that museums with several smaller rooms will have more clusters. In such cases multiple neural networks can be trained that probably improve the classification rate even further. Although, our museum does not provide optimal conditions for our algorithms, we were still able to show that they perform well.

The 15 inexperienced users captured a total of 373 images (with a minimum of 9, a maximum of 49, and an average of 25 images per visitor).

The resulting classification rates are illustrated in figure 7.4. We have compared six different approaches: the first approach uses one image classifier that has been trained to the entire set of 146 objects without any localization or pathway information. The second approach uses Bluetooth lo-

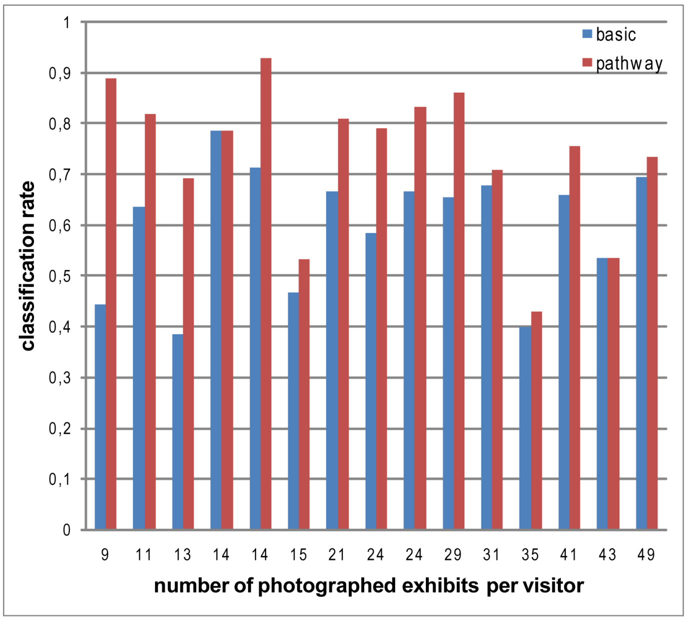


Figure 7.5: Illustration of the classification rate of the image classifier with and without the pathway classifier and pathway clustering for each visitor. It proves the pathway classifier's independence of the path length (number of successively photographed exhibits) as well as its consistent benefit for image classification.

calization and evaluates one specific image classifier for each floor that is tagged and identified with a Bluetooth emitter. The third approach evaluates the determined pathway clusters, while the fourth approach applies the pathway classifier. The last two approaches (row six and seven in 7.4) compare the combination of image classification and pathway classification for the two different localization techniques (Bluetooth and pathway clusters). All of these approaches have been evaluated with the images taken by the inexperienced users that are assigned to our evaluation set. As mentioned above, they captured the exhibits unattended from arbitrary perspectives and scales and thus provide realistic evaluation data. Furthermore, we evaluated the approaches by one additional user (we refer to him as an ”experienced user”, because he captured the pictures more carefully) who repeated these pathways but photographed the exhibits from similar perspectives that had been trained in advance. We will refer to the data of the experienced user later in section 8.3. For the first case, we also computed the probability of the correct exhibit being under the top three classification candidates. The reason for this is that the user interface displays the three most likely candidates as described in section 3.2.2. Being under the top three means, that a correct object is displayed to the user although not necessarily as the first choice.

Our results show that the image classification in combination with pathway classification and pathway clustering achieves a classification rate that compares well with image classification in conjunction with Bluetooth localization. Furthermore it outperforms the image classification alone by up to 20%. Eleven pictures out of the 107 misclassified images were not recognized correctly in this case because the corresponding visitors violated the predefined pathway clusters. However, through applying the pathway segmentation we could improve the classification rate by approximately 10%. By combining the Bluetooth localization with the pathway predictor, the overall classification rate can be further improved by 7-9%. Figure 7.5 compares the recognition rates for the image classifier alone, as well as the image classifier in combination with the pathway classifier and pathway clustering, with respect to the number of captured images for each of the 15 evaluation subjects. It reveals that a combination of image and pathway classification is, in most cases (13 out of 15), better (and never worse) than image classification alone regardless of the path lengths.

The prediction rate of the pathway classifier itself based on the data of the inexperienced users is displayed in figure 7.6. It illustrates, for instance, that with a probability of 30.5%, the first 10 prediction candidates contain the object that the visitor will approach next. If the first 30 prediction candidates are considered, the probability is 65.9% compared to a probability of $30/146 = 20.5\%$ if the next potential exhibit is randomly selected. Even if the pathway classifier assigns a low probability to the correct next object, the overall classification can still be successful due to the presence of the image classifier and the a priori probability.

In the course of the field study we also handed out a questionnaire to the 132 visitors at the end of their visit (see appendix A.3). We asked them if they would prefer to type numbers or take pictures to identify an object. Although they did not receive any multimedia content after photographing an exhibit we were convinced that they got a first feeling of this pointing interaction technique. Approximately 60% of the visitors answered that they would prefer to take pictures instead of typing reference numbers. Older people in particular stated that taking pictures would be beneficial because reading small numbers on the tags was difficult. This would be by-passed by the pointing technique. Furthermore, the main argument for using reference numbers was its simplicity. However, this was also the main argument for taking images.

In section 8.3 we will provide further details and show how this approach influences the overall classification rate of our system.

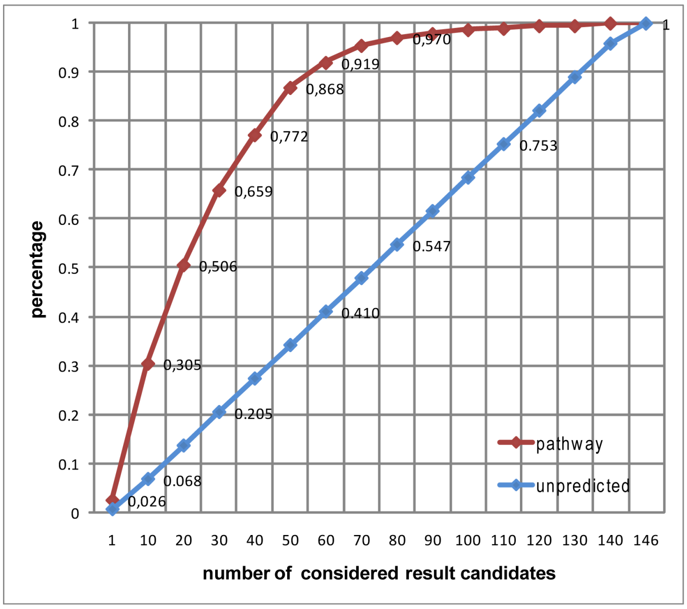


Figure 7.6: The probability of selecting the correct next object is significantly higher using prediction (red line) as opposed to a random selection (blue line).

7.6 Discussion

In this chapter we have demonstrated how the analysis of pathway patterns can improve mobile image classification. We developed two main adaptation techniques to determine the location of museum visitors and apply them in order to increase the classification rate of object recognition: The first technique applies collected pathway data to predict the sequence of objects captured by future visitors. The second technique examines spatio-temporal data in order to determine areas in the museum which visitors no longer return to after visiting them. These areas can be disregarded for subsequent image classifications after the visitor has progressed past them. Consequently, the number of possible candidates decreases continuously as the visitor moves through the museum. In a user study we showed that the recognition rate of image-based classification can be improved by up to 20% when combined with pathway prediction. In addition, we demonstrated that by applying pathway prediction, we can achieve similar classification rates to approaches that combine image-based classification with Bluetooth localization –but without the use of auxiliary hardware.

One potential drawback of this approach is the initial effort that is necessary to model a detailed map of the locations of all exhibits in the museum. To allow a time-efficient and intuitive modeling, we have developed an editor that imports digital floor plans. Objects can be positioned on the map via drag&drop and intuitive object manipulations. For example, it required approximately 20 minutes to model all three floors for our experiment.

In chapter 4 we presented an approach that utilizes multiple images for classification. Although we have not tested this in practice, the adaptation of this approach would be similar to the process of single image classification: at the end of the multi-image classification the nearest-neighbor distances of the voted result candidates can be averaged and reformulated as probabilities as proposed by Ting et al. [170]. The probabilities are then combined with the pathway predictor to improve classification performance.

In addition to predicting future exhibits based on those already visited, the prediction could be further refined by considering the time between exhibits. If a visitor takes a picture of an object, we can track the duration until he captures the next exhibit. Based on the elapsed time and the spatial configuration of the exhibits in the museum (that we have defined through our distance map, see section 7.1) we can determine the objects that can be physically reached by the visitor. This can be either computed statically by assuming a constant speed of travel for all visitors or dynamically by adaptively training the maximum speed of the users.
We have tested both approaches in our experimental setup. Unfortunately, the results were not satisfying. The main reason for this was the architecturally compact design of the museum. The museum consists of three floors and each floor had approximately 200 m^2. So visitors could move either horizontally or vertically to all regions in the museum in a short period of time. Consequently, the time-based predictor commonly returned all exhibits of the museum as possible future locations. This technique could, however, be beneficial for larger museums or in outdoor applications.

Although we have presented and evaluated our approach in the course of our PhoneGuide system, these techniques are generally speaking not limited to museum guidance tasks. For location-based outdoor information systems in particular, such as city guides, where the rough locations of users can be retrieved through GPS, additional pathway predictions can be beneficial. For example, hardware-based localization techniques then only need to be used temporarily to save battery consumption.

For indoor or outdoor guidance systems, the adaptively generated pathway information can also serve as input for recommendation systems. Through collaborative filtering [99], for instance, the mobile application offers suggestions based on common preferences that have been derived from the behavior of previous users. In section 9.2 we will discuss this idea in more detail.

8 EVALUATION

In the last four chapters we have presented different approaches and techniques for realizing adaptive image classification on mobile phones. We have demonstrated how to identify one or multiple objects in a single image and how the classification rate can be increased by combining the results of multiple pictures. Through adaptation techniques based on ad-hoc networking and person movement prediction we have shown how the classification performance can be improved even further. In the context of a mobile museum guidance system we have explained the benefits of each approach and evaluated each method.

In the following we provide a technical overview of the system architecture of our entire PhoneGuide system that comprises the approaches presented in this thesis as well as the methods developed prior to this work. We will explain the main parts necessary to realize our system and provide details on the data that is collected on each visitor's phone during use. Afterwards, we present a flow diagram of our user interface that provides an overview of all the extensions that we have presented in the previous chapters. Finally, we will show how the different adaptation techniques that we have developed influence the overall classification rate, discuss their pros and cons and outline their transferability.

8.1 System Architecture

The approaches we have presented in this thesis as well as the previously developed techniques used in our PhoneGuide system result in a series of requirements for the system architecture [35, 32]. For example, on the server side, user interfaces have to be provided to input training or multimedia data, while the mobile phone application needs proper data handling to use the different classifiers and adaptation parameters.

An overview of the basic system architecture of our entire PhoneGuide system is illustrated in figure 8.1. It consists of a server as well as one or multiple clients (mobile phones). The visualized processing elements represent the main software units of our system.

The server carries out all preprocessing steps that are necessary to facilitate object recognition on the mobile phones. It offers two interfaces. Through the user interface an expert provides the initial image and preprocessing data that is necessary to create the different elements of the mobile image classification process. To begin with these are initial videos that were recorded of each exhibit in a museum for the basic image classification task as explained in section 3.2.1. If a group of exhibits was recorded in a video, the expert manually identifies each visible subobject in the first video frame to assist the training of the subobject classifier as described in chapter 5. Furthermore, a distance map of the museum and pathway data is provided for the pathway analysis (cf. chapter 7) as well as rules for the cell-based Bluetooth tracking (cf. section 3.2.3). Finally, multimedia content about the exhibits is provided such as images, videos or audio files. We call this entire set of data (excluding the multimedia content) ”adaptation parameters” because these parameters are continuously updated through data that is collected by the individual visitors' phones as describe later. This data is stored on the server.

After preprocessing the adaptation parameters, for example, by extracting keyframes from the videos, the related data is routed to the different classifier creators.
One group, the image classifiers, is created by the *image classifier creator*. It extracts image features from the video frames and generates 3-layer neural networks for the object recognition process as stated in section 3.2.1. In combination with the *spatial relationships creator* it provides the classifiers for the subobject recognition as explained in chapter 5. To accomplish this, the selected subobjects are tracked throughout the videos and the spatial relationships among these subobjects are computed. The *image relations creator* determines the near-far relations among the video frames and stores them in relation tables as explained in chapter 4. In combination with the image classifier it generates a vp-tree for a nearest-neighbor search with relational reasoning. Finally, the *pathway classifier creator* analyzes the pathways of different museum visitors and generates the corresponding classifier for pathway prediction. For this it uses the distance map to compute the pathway features.

When visitors enter the museum, this entire set of classification data (image and pathway classifiers, spatial relationships, relation tables and rules) together with the individual multimedia set and front-end application is transferred to the visitors' devices through the client/server interface as explained later. Note that communication between server and client takes place only twice: at the beginning and at the end of a museum visit as we will explain later.

The mobile application, in turn, holds three additional software interfaces. They are used for receiving important adaptation parameters and for presenting the classification result: the sensor interface allows the mobile phones to search for distributed sensor boxes in the museum via Bluetooth. Their identifications are used by the *classifier selector* to determine the current location and to select the signal cell's individual image classifier. The phone-to-phone interface enables the devices to automatically exchange classification data to cope with environmental changes as described in chapter

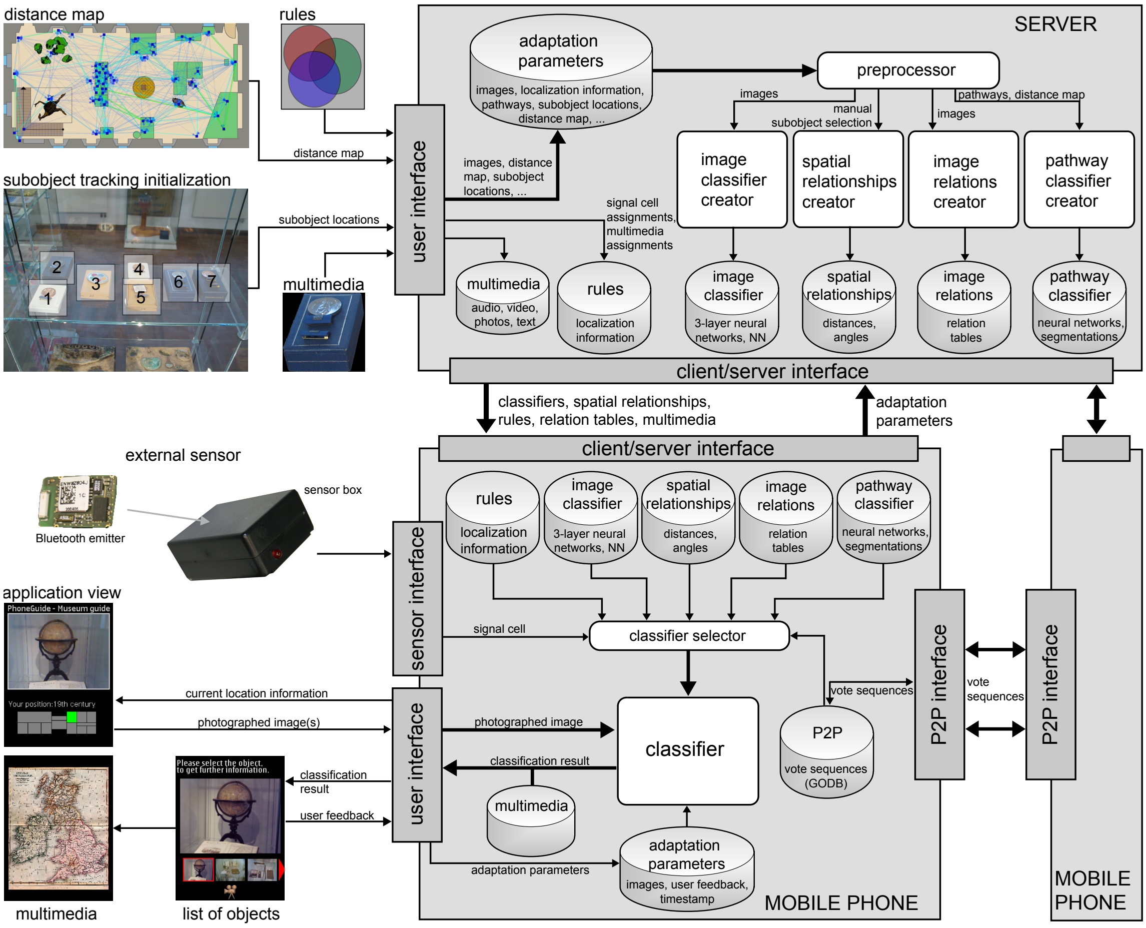


Figure 8.1: Overview of our adaptive mobile museum guidance system: Classification data (image and pathway classifiers, spatial relationships, relation tables, rules) and multimedia in combination with the front-end application are initially retrieved by the phones from a server when entering the museum. During runtime, no server connection is necessary. The visitors interact with the classification system through a user interface by taking photographs and providing interaction feedback. Through the sensor interface, phones can search for external sensors to determine their approximate location while the phone-to-phone interface makes it possible to exchange classification data in order to cope with sudden environmental changes. Instead of sensors, the pathway classifier can be applied to predict the location of users. When leaving the museum, adaptation parameters are transmitted to the server to help continuously improve the system over time.

6. Finally, the user interface visualizes the current location of the museum visitors on a map and the probability-sorted list of objects and subobjects showing the (intermediate) recognition results. At

the end of each classification the multimedia content is displayed. In our current implementation this comprises audio- and video files as well as images.

The object recognition is carried out by the *classifier*. This central element of the mobile application has access to the adaptation parameters (e.g., for creating the pathway map) and receives all classifiers that are necessary to identify the photographed object from the *classifier selector*. This unit selects the appropriate classifiers based on the available underlying information such as the location of the user (e.g., through BT-tracking or pathway segmentation). The final result is then displayed on the mobile phone's screen and the client application continuously tracks the users' selection of the correct exhibit. This provides an exact mapping between photographed image and real object ID. In combination with the images, these mappings are stored on the mobile device as adaptation parameters. Furthermore, the timestamp of each picture is recorded to reconstruct the users' pathways afterwards.

At the end of a museum visit, the mobile devices automatically transfer their collected adaptation parameters to the server to enable the image and pathway classifiers to continually learn and improve over time as explained in section 3.2.2.

In our scenario, the exchange of data between server and client can either be accomplished through a USB cable, a wireless connection (e.g., Bluetooth or WLAN) or using a memory card. In the latter case, visitors receive the front-end application of PhoneGuide on a memory card at the beginning of their museum visit. The adaptation parameters collected during their visit can be stored on the card and it is then given back at the end of the visit. Currently, we have implemented data exchange through USB. The size of data to be transferred depends on the number of different objects and the amount of multimedia data available. In our current implementation the size of the core application including the data for the user interface is approximately 3.8 MB. Furthermore, additional memory would be necessary for the image classifier (neural network: $\sim$415 kB or vp-tree: $\sim$4.9 MB, relation tables: $\sim$1 MB), the pathway classifier ($\sim$100 kB) and the subobject data ($\sim$237 kB) as explained in the corresponding chapters.

The server application was implemented in Matlab 7.1 [120] to utilize mainly the Image Processing and Neural Network Toolbox. It also consists of a Java 6 interface that converts the data from the Matlab data format to the mobile phones' platform. The mobile front-end application of PhoneGuide was implemented in Java ME 2.5.2, MIDP 2.0, CLDC 1.1.

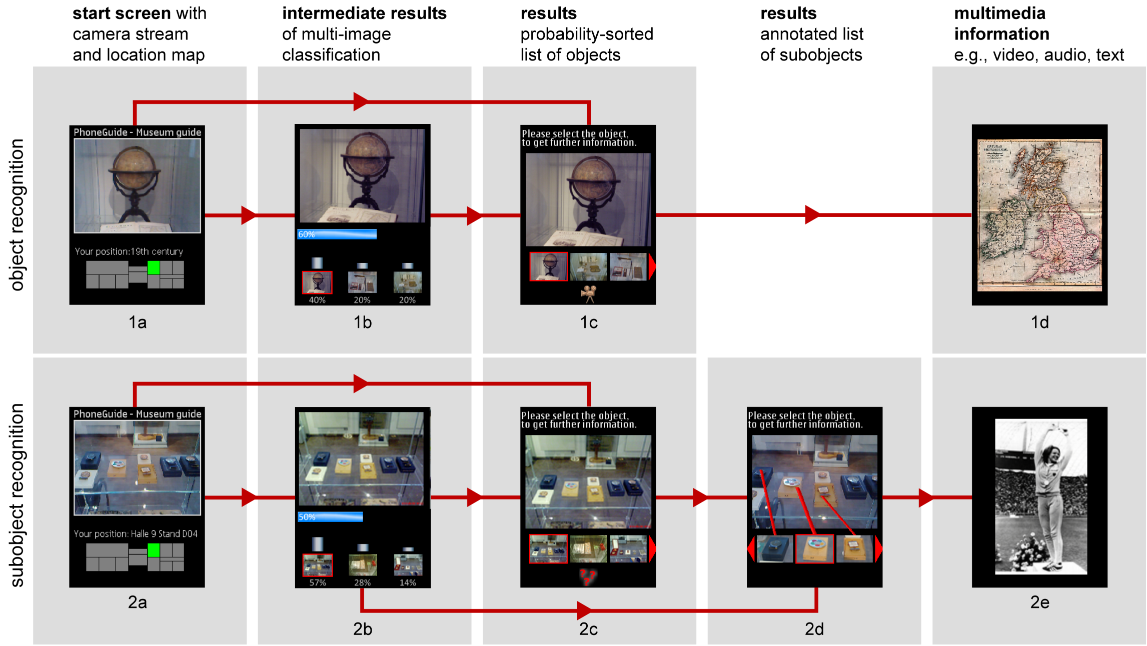

start screen with
camera stream
and location map
intermediate results
of multi-image
classification
results
probability-sorted
list of objects
results
annotated list
of subobjects
multimedia information
e.g., video, audio, text
object recognition
subobject recognition
PhoneGuide - Museum guide
Your position:19th century
60%
40%
20%
20%
Please select the object,
to get further information.
1a
1b
1c
1d
PhoneGuide - Museum guide
Your position: Halle 9 Stand D04
50%
57%
28%
14%
Please select the object,
to get further information.
Please select the object,
to get further information.
2a
2b
2c
2d
2e

Figure 8.2: Flow diagram showing the different elements of our user interface if either a single exhibit is photographed or if multiple objects in a single image are captured.

8.2 User Interface

In this section we will illustrate the flow diagram of the entire mobile user interface including all extensions that have been presented in the thesis. As the details of these extensions have been described in the corresponding chapters on the different object classification algorithms (cf. section 3.2.1, section 4.4 and section 5.3), we will only provide an overview here.

The flow diagram in figure 8.2 is separated into two cases: The displayed user interface if a single exhibit is photographed as well as the case if multiple objects in a single image are captured.
If a visitor takes a picture of a single exhibit (8.2.1a) the image is recognized and a probability-sorted list of objects is presented (8.2.1c). After selecting the correct object, multimedia content is displayed (8.2.1d). If the visitor takes a video of an object through a near-far camera movement as explained in chapter 4 instead of just one image, intermediate results are presented (figure 8.2.1b).

If a visitor takes a picture of a group of objects, the image is recognized and a probability-sorted objects list is shown comparable to the previous case (8.2.2c). However, instead of retrieving multimedia content after selecting the correct object, the subobject detection algorithm is carried out. The result is a list of annotated subobjects (8.2.2d). By selecting the subobject of interest, the corresponding multimedia data is displayed (8.2.2e). If the visitor takes a video of a group of objects as explained in the previous case, intermediate classification results are displayed before the list of objects is presented as shown in figure 8.2.2b.

8.3 Classification Performance

In this thesis we have presented different approaches that improve the overall classification rate of our adaptive mobile museum guidance system. In each chapter on the corresponding method, we have compared our technique with an unadapted image classification and pointed out its advantages and benefits. To validate this information we carried out multiple user studies in two different museums.

In this section we will demonstrate how the various techniques influence the classification rate when they are applied in conjunction. We discuss the different results and give explanations for occurrent variations.
As a basis, we selected the measured classification rates from the field study that we conducted during the evaluation of our pathway analysis approach (cf. chapter 7). The benefits of the remaining techniques are estimated as explained later.

Figure 8.3 illustrate the final results. It shows the classification rates of the object recognition approach (red), of the object recognition approach when the top 3 result candidates are considered (purple) and of the subobject recognition technique (green) for the various adaptation methods that we have presented in this thesis. These approaches are applied successively so that the last element on the x-axis reveals the classification rate when all adaptation techniques are used together.

The basic object recognition approach (100 global color features, 3-layer neural network, 373 images, 240x180 pixels) achieved a recognition rate of 50.67% for 146 objects if the top ranked result candidate is considered. This relatively low classification rate has several reasons. First of all, multiple images in the evaluation set were of poor quality since the users were relatively inexperienced in photographing exhibits using a mobile phone. Some people had problems ”finding” the object in the display of the device to capture it properly. Others were careless in how they took pictures, probably because they knew they would not receive any feedback. Consequently, the images were motion blurred, objects were only partly photographed or off-center. These findings are confirmed by the fact that an experienced user (a user that captured the images more carefully) achieved a classification rate of 65.68% as illustrated in section 7.5.

In the context of this evaluation we considered the phone-to-phone communication approach as a

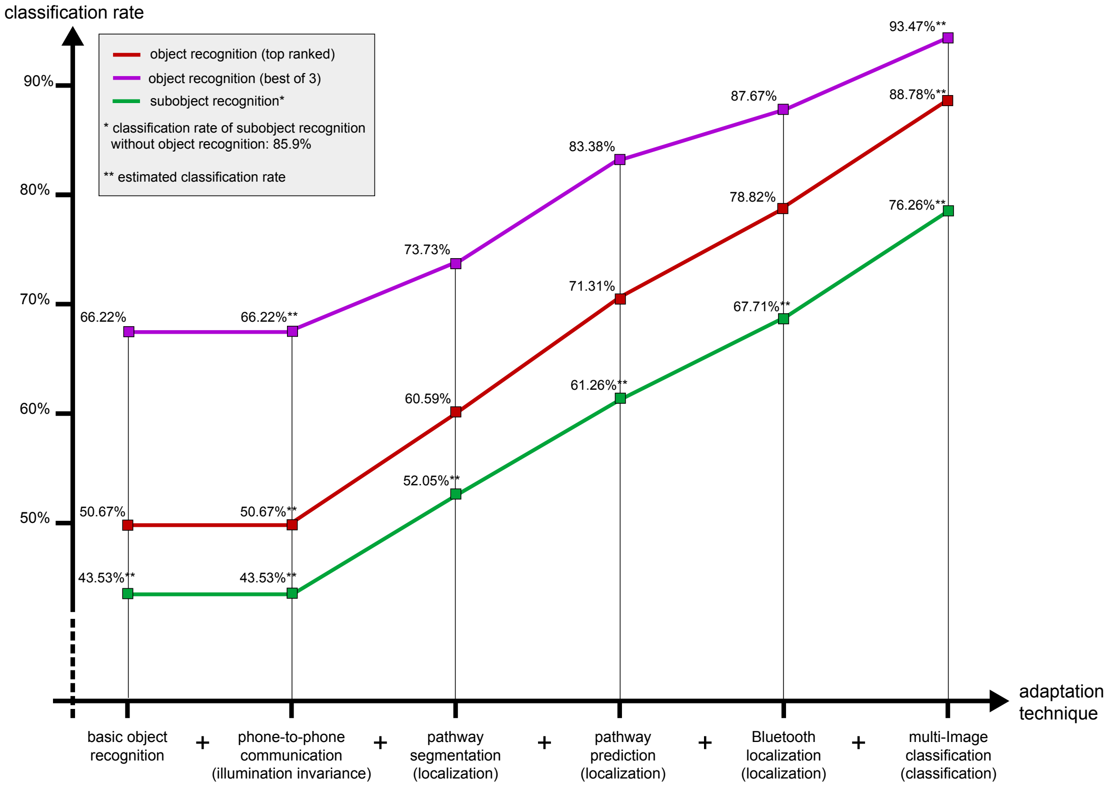


Figure 8.3: The benefits of our adaptation approaches for the classification rate of the object recognition approach (red), of the object recognition approach when the top 3 results are considered (purple) and of the subobject recognition technique (green).

technique for making image classification less susceptible to changes in illumination levels. Consequently the graph shows no changes in the classification rate. Our decision to undertake this evaluation is the lack of an extensive field study that shows the exact improvement of the recognition rate in a realistic use case. However, its ability to adapt to environmental changes was confirmed through simulations as explained in section 6.5.

If we combine the basic object recognition technique with the pathway segmentation approach we achieve a classification rate of 60.59%. If we add the pathway prediction method we achieve a classification rate of 71.31%.

The classification rate rises to 78.82% if the Bluetooth tracking approach is added. In this case the pathway segmentation is redundant because it holds the same object subsets as the localization approach. However, this is only valid for our particular museum setup and is not universally the case. In our system, the pathway segmentations are considered as additional location cells that are combined through conjunction with the cells of the Bluetooth tracking approach as explained in

section 3.2.3.
We selected a relatively conservative solution for the Bluetooth tracking technique by defining each floor as one location/signal cell. Of course, a higher classification rate can be achieved, if more Bluetooth emitters are distributed throughout the museum with a lower signal range. This would result in a finer grid of overlapping signal cells with smaller sets of possible result candidates. However, this would also increase the investment and maintenance costs for the museum and result in longer scanning durations to identify the Bluetooth beacons in the museum.

Finally, if we add the classification using multiple images[7] we achieve a classification rate of 88.78% when considering only the top ranked result of the object classification. This result is estimated because we evaluated this technique in the course of a different user study where we could not collect any pathway data. Consequently, we approximated the recognition rate by considering the relative gain of this approach in comparison to the basic object recognition achieved in the same user study as presented in chapter 4. A mathematical explanation of the estimation can be found in appendix A.4.

The purple graph in figure 8.3 shows the corresponding classification rates of the different approaches if we consider the top 3 ranked classification results. Here we achieve a final classification rate of 93.47%. The top 3 results are relevant because our user interface displays an at-a-glance overview of the top 3 candidates as the initial final results screen on the visitor phone as illustrated in the previous section. The visitor can then select the correct item of interest with a few clicks to retrieve the corresponding multimedia content.

The subobject recognition approach achieves lower classification rates than the object recognition technique because it is a consecutive classification step. Consequently, if the first image classification fails, then we count this as a misclassification for the subobject detection as well. However, if we consider the subobject recognition exclusively, we achieve a classification rate of 85.9% as shown in section 5.4 without applying an adaptation technique.

These results reveal that even if the basic image classifier performs weakly, we can achieve classification rates of almost 90% in a realistic museum scenario with ordinary museum visitors.

To provide a baseline to classify these results, we applied an object recognition technique based on local image features. We utilized the SIFT implementation of [178] for the classification task. SIFT is known to be invariant to rotation and scale and robust with respect to illumination changes and occlusion and is widely used for many different object recognition tasks as explained in chapter 2.

[7] In this case we consider our multi-image classification approach as additional adaptation technique.

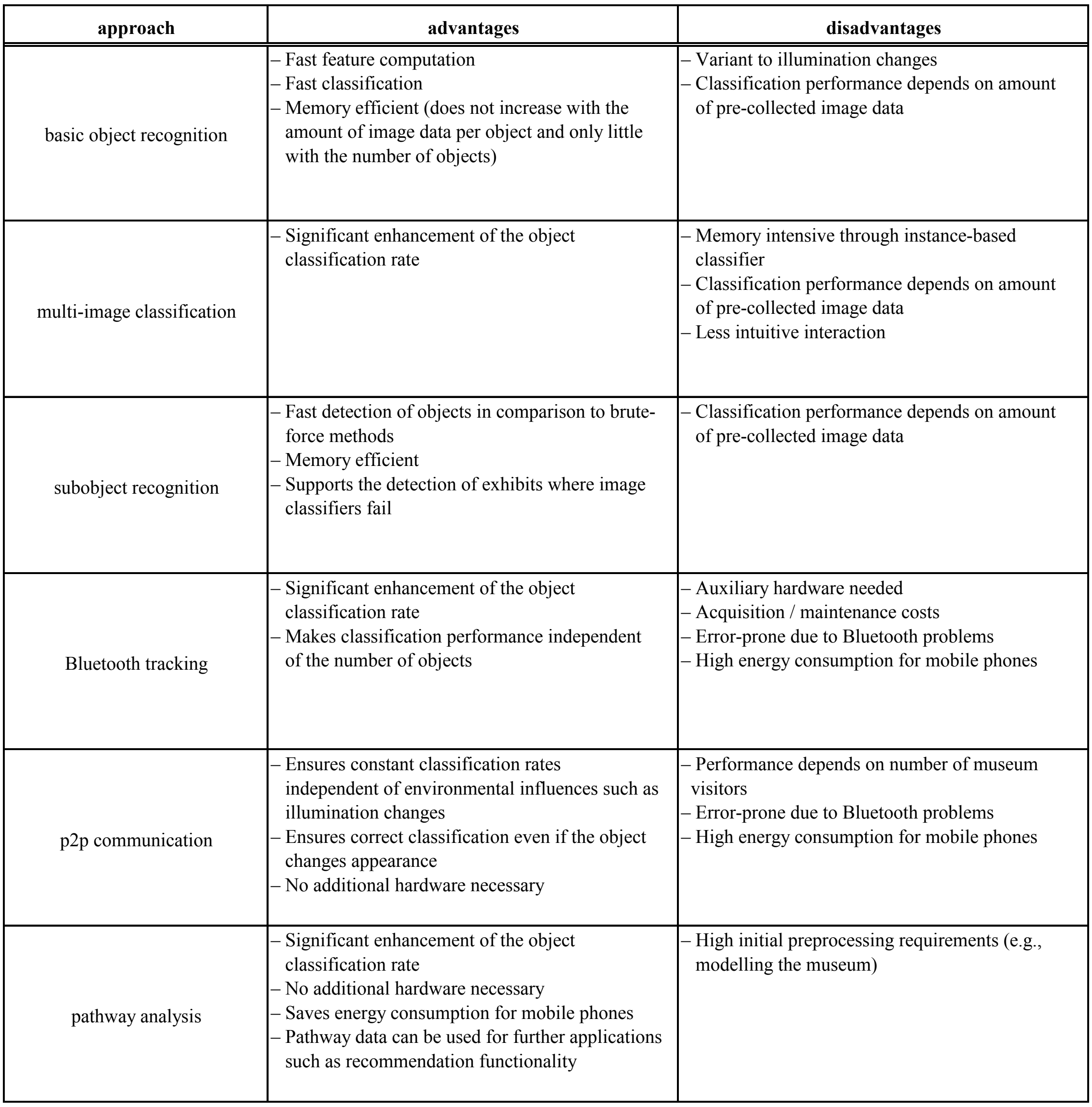
We randomly selected 16 images per object as a reference from the image set that we used to train our neural networks. We have applied the standard settings for the algorithm as proposed by Lowe et al. [115] and utilized a nearest-neighbor search for matching. After classifying the pictures of the subjects we used for testing our approach, we achieved a classification rate of 68.18%. This result is better than our basic image classification and could probably be improved by additional techniques such as geometry consistency checking. For example, related approaches that rely exclusively on SIFT/SURF techniques (in combination with localization) achieve classification rates of 70%-85% [52] or 77% [169] for different object sets. However, the result show that a more sophisticated object recognition algorithm has room for improvement and is outperformed by our adapted image classification system.

8.4 Discussion

At the end of the individual chapters of the approaches presented in this thesis, we discussed different aspects of our solutions and outlined improvements and extensions. In this section we take a look at the methods from a broader point of view and offer an overview of their advantages and disadvantages as illustrated in figure 8.4. The intention is to provide a rough estimation of their practicability and assist the reader in selecting one or multiple techniques for possible future applications. Furthermore, we outline in this chapter how the approaches we have presented in this thesis can be applied if the basic object recognition algorithm were to be exchanged.

8.4.1 Advantages and Disadvantages

The basic object recognition algorithm that we apply for identifying objects is simple but practical and is characterized by fast feature computation and classification. Through the application of neural networks it is memory efficient and does not increase in size with the number of training samples per object and only little with the amount of objects. The use of global color features facilitates the robust recognition of multiple objects. However, the application of color features also makes it vulnerable to changes in illumination levels. Furthermore, since the features are not invariant to scale or rotation, the classification performance depends on the discrepancy between the photos taken by the visitors and the pre-captured images used for training. We try to overcome this problem in practical use by continuously collecting image data through user feedback as explained in section 3.2.2.



approach	advantages	disadvantages
basic object recognition	– Fast feature computation – Fast classification – Memory efficient (does not increase with the amount of image data per object and only little with the number of objects)	– Variant to illumination changes – Classification performance depends on amount of pre-collected image data
multi-image classification	– Significant enhancement of the object classification rate	– Memory intensive through instance-based classifier – Classification performance depends on amount of pre-collected image data – Less intuitive interaction
subobject recognition	– Fast detection of objects in comparison to brute-force methods – Memory efficient – Supports the detection of exhibits where image classifiers fail	– Classification performance depends on amount of pre-collected image data
Bluetooth tracking	– Significant enhancement of the object classification rate – Makes classification performance independent of the number of objects	– Auxiliary hardware needed – Acquisition / maintenance costs – Error-prone due to Bluetooth problems – High energy consumption for mobile phones
p2p communication	– Ensures constant classification rates independent of environmental influences such as illumination changes – Ensures correct classification even if the object changes appearance – No additional hardware necessary	– Performance depends on number of museum visitors – Error-prone due to Bluetooth problems – High energy consumption for mobile phones
pathway analysis	– Significant enhancement of the object classification rate – No additional hardware necessary – Saves energy consumption for mobile phones – Pathway data can be used for further applications such as recommendation functionality	– High initial preprocessing requirements (e.g., modelling the museum)

Figure 8.4: Overview of the advantages and disadvantages of the different approaches presented in this thesis.

The multi-image classification approach has the same disadvantages as the previous technique but increases the classification performance significantly through combined image classification. However, since we apply relational reasoning to speed up classification we have to apply a nearest-neighbor approach instead of neural networks. This increases memory consumption and is best applied where memory use is less critical. Furthermore, the need to employ a prescribed kind of camera movement is slightly cumbersome for the visitors.

The application of our subobject detection approach is a fast and memory efficient way of detecting and recognizing multiple objects in a single image. By means of spatial relationships among the subobjects we can enhance the classification speed significantly in comparison to related brute-force methods and improve the classification rate by defining search regions. Furthermore, spatial relationships allow the detection of objects where image classifiers fail. Again, the classification performance depends on the amount and the quality of pre-collected image data.

The localization of museum visitors by distributing Bluetooth emitters throughout the museum has the highest impact on the improvement of the classification rate. This is especially effective when many Bluetooth emitters are distributed throughout the museum, resulting in smaller signal cells and a correspondingly smaller number of possible result candidates per cell. This significantly enhances the classification rate especially when many different objects have to be recognized. A disadvantage is that the acquisition and maintenance of Bluetooth emitters is costly. In addition, mobile phones need to continuously scan for the Bluetooth emitters, which is energy consuming and can be error-prone.

The phone-to-phone communication approach does not increase the classification rate per se but ensures a constant classification rate independent of the current illumination state and can cope with revised object replacement without the need for additional hardware. As with Bluetooth localization, this method also consumes battery power and interphone communication is somewhat error-prone. Furthermore, to be effective a minimum number of museum visitors with mobile phones is needed.

Pathway analysis leads to a comparable level of classification improvement to the Bluetooth localization method. Its advantage is that it does not require additional hardware or RF-communication which reduces the battery consumption of the phones and obviates the need for hardware and maintenance costs. Furthermore, the pathway data can also be used for other purposes such as recommendation tasks. A disadvantage of this approach is that it requires more time and effort for the preprocessing steps, such as modeling the locations of the exhibits.

8.4.2 Transferability

As mentioned earlier, the basic object recognition algorithm is fast and memory-efficient but also error-prone, for example, where scale or illumination levels change. Replacing this algorithm with more sophisticated algorithms such as SIFT or SURF (cf. section 2.3.1) would increase the basic classification performance as explained in the previous section. However, as pointed out in the introduction, we believe that our approaches would still be beneficial even if the image features or the entire algorithm were to be replaced. In the following we describe how in theory the different techniques presented in this paper can be combined with other object recognition algorithms.

The multi-image classification method, for example, improves the recognition rate of object recognition algorithms by effectively combining results of multiple image classifications. Although the selection of multiple images through a near-far camera movement would be less useful for object recognition algorithms that are based on local image features, because they are scale-invariant, multiple pictures could be still valuable if the camera movements showed the exhibits from different angles. This would increase the number of extracted keypoints and would describe the object more properly. The relational reasoning technique could then serve as additional geometric consistency check for the different keypoints.

The application of spatial relationships among objects in a single image, as proposed by our subobject detection algorithm, can also be used as a geometric consistency check. Furthermore, through the definition of search regions, the object recognition algorithm only has to search for specific objects in predefined image areas. This would reduce the time required for computationally expensive feature extraction, especially where high-resolution images are concerned. Furthermore, only the corresponding database keypoints of this image area need to be considered during classification. In addition, the ability to detect fully occluded or invisible objects is an additional benefit for each object recognition process.

The phone-to-phone communication approach would still be useful in cases where local light variations are significant or where the appearance of the object changes in some other way, e.g. through revised object placement. Since the classification sequence is exchanged among the phones rather than the image features, the applied object recognition algorithm does not influence our proposed synchronization process.

Finally, we can expect that our pathway analysis approach will improve every mobile image classification system since it provides additional information that is combined with the object recognition algorithm. Only minor changes are necessary if, for example, local image features are applied that are based on an instance-based classification. In this case, the corresponding classifier does not assign any probabilities to the potential result candidates which would be necessary for combining them with pathway prediction. A solution for this is presented by Ting et al. [170] who propose a mathematical conversion of nearest-neighbor distances to matching probabilities.

9 CONCLUSION AND FUTURE WORK

9.1 Conclusion

The advent of high-performance mobile phones in combination with built-in sensors such as cameras, localization techniques such as GPS or wireless communication technology such as Bluetooth or WLAN, has opened up the opportunity to develop new context-aware applications for everyday life. In particular, applications for context-aware information retrieval in conjunction with image-based object recognition have become a focal area of recent research.

In this thesis we have presented object recognition algorithms for mobile phones that recognize one or multiple objects in a single image. To enhance the classification performance of these approaches, we designed different adaptation techniques such as ad-hoc networking or person movement prediction that exploit the unique characteristics of mobile devices. In the context of a mobile museum guidance system we have shown how these approaches can be applied and how they perform in realistic use cases.

The main achievements of this work can be summarized as follows:

- Overview and classification of existing techniques to design and develop context-aware mobile information systems.
- Development of an object recognition algorithm that classifies a single exhibit by utilizing multiple photos captured during a short camera movement.
- Development of an object recognition algorithm that detects multiple objects in a single image by utilizing the spatial relationships among the exhibits in the captured photo.
- Development of an adaptation technique that allows mobile phones to exchange classification data among each other to ensure robustness against environmental changes.
- Development of an adaptation technique that utilizes the spatio-temporal pathway data of museum visitors to unobtrusively improve the classification performance.

The remainder of this section will explain these achievements in detail.

At the beginning of this thesis we provided an overview of the basic techniques that can be used to realize context-aware mobile information systems. We have identified two main areas that use either location to provide context-aware information or utilize vision-based techniques to identify specific entities in the users' environment. In indoor environments, the localization is mainly based on communication technologies that are used for positioning. Consequently, we considered the signal range of the technology used to further subdivide this area into applications that employ either short-range or long-range emitters.

The vision-based techniques were separated into assisted-based as well as into natural object recognition approaches depending on whether additional reference tags are used. The corresponding related work of these areas revealed that for improving the classification performance, only the location and the orientation of the corresponding mobile phone are considered as additional adaptation parameters. Since the classification rates of many approaches showed room for improvement, they provided the motivation to develop further adaptation techniques as presented in this thesis.

The remaining achievements of this thesis can be divided into two main parts.

First, we designed different object recognition techniques that make it possible to identify one or multiple photographed exhibits. For recognizing a single object we implemented a simple yet fast and memory-efficient object recognition method in the past that extracts global color features from a captured image and utilizes a 3-layer neural network for classification.

Based on this, we have shown how the object recognition rate can be improved by combining the results of multiple images that were captured of a single exhibit. By applying a nearest neighbor search instead of neural networks for classification we could consider the relations among the database images retrieved. By proposing a simple near-far camera movement that visitors carry out when facing an exhibit, we can reject images of possible result candidates that do not adhere to this near-far relation. This improves the recognition rate and classification speed if the number of captured images is low.

To detect multiple objects (subobjects) within a single image, we developed a consecutive classification step that considers the spatial relationships among the subobjects to improve the classification rate as well as classification duration. For this a search mask spirally searches in the center of the image for known subobjects. If an exhibit is found, the spatial relationships span search areas relative to the detected subobject so that the search for any remaining subobjects is carried out only in the predefined image areas. This reduces the number of classification steps and increases the classification rate since it narrows down the amount of possible locations of the exhibit.

For the second part, we developed a variety of different adaptation techniques to enhance the classification performance of the object recognition task.

The most straightforward means of improving mobile image classification systems is to consider the location of the corresponding visitor. Based on his position, only the objects in the vicinity serve as possible result candidates during the object classification task.

In this thesis, we have proposed two different adaptation techniques to accomplish this. The first solution, which was developed prior to this thesis, is based on localization using RF-technology. Bluetooth beacons are distributed throughout the museum which span a grid of overlapping signal cells. The visitors' mobile phones continuously scan for these emitters. If one or multiple Bluetooth beacons are detected the location is derived based on the physical area that is covered by the signal ranges of the emitters.

The second solution is based on an analysis of visitors' pathway data. During a field study we investigated how visitors move through a museum. Based on this knowledge we designed a classifier that takes into account the history of already visited objects to predict the future locations of users. In contrast to related approaches, we not only consider the location of visited exhibits but also the position of co-located objects to describe the pathways thoroughly. This facilitates the prediction of future exhibits, even if the past pathway has never been traced before. The predictions for each exhibit are then combined with the object recognition process to retrieve a final classification result. In addition, we developed an algorithm that detects physical areas in the museum that visitors do not return to after passing them. This makes it possible to exclude the corresponding exhibits in these areas as possible result candidates for future classifications.

Our evaluation revealed that the pathway analysis approach results in performance improvements similar to traditional RF-based localization techniques but without their drawbacks such as the need for additional hardware and maintenance costs or the high energy consumption.

In order to cope with environmental changes such as varying lighting conditions or revised object placement, we developed an adaptation technique to seamlessly exchange classification data among the visitors' mobile phones. Through the effective synchronization of the devices, we enforce cooperative classification improvements during the users' actual visit in the museum. To accomplish this, the mobile phones collect the classifier's individual classification sequence of each photographed exhibit. This classification sequence is unique for each object and serves as a feature vector that describes the current appearance of the object. By distributing these sequences, they serve as an additional database for carrying out a second classification step on phones that subsequently approach this object. The synchronization is facilitated by establishing ad-hoc networks among co-located mobile phones via Bluetooth. This ensures that each phone holds the most up-to-date information. This makes it possible to achieve high classification rates even if the appearance of objects changes significantly.

The preprocessing data that is necessary to ensure the functionality of our system, such as image data or pathways to train the corresponding classifiers, is collected in advance by an expert.

However, using our user interface the system is also able to collect this data during usage. A probability-sorted list of result candidates is displayed on the visitor's mobile phone after image classification. This allows the users to select the correct object and retrieve multimedia content even if the corresponding classification has failed. Furthermore, this gives us a ground-truth concerning the current object of interest and allows us to store the captured pictures in combination with additional information such as a timestamp on the mobile devices. This data can then be uploaded at the end of a museum visit to a server where it can be used to continuously improve the classifiers over time.

We have demonstrated that our basic object recognition rate achieves a classification rate of $\sim$50% for 146 different exhibits. These exhibits were defined by ordinary and inexperienced visitors in a museum. Through field studies we revealed that the application of all adaptation techniques could increase the classification rate of a simple object recognition algorithm by about 40% from $\sim$50% to $\sim$89%. If we consider the top 3 result candidates we achieved an overall classification rate of 93.47% for a real-world scenario.

9.2 Future Work

While specific aspects of future work and improvements have been discussed individually for each of the presented approaches in the corresponding chapters, this section offers an outlook on future work from a broader point of view. Some of the proposed methods are ideas or basic approaches that we have discarded or not pursued conclusively over the past few years due to time limits.

For example, we started to design a localization technique based on Bluetooth that determines the position of visitors in museums exclusively through their mobile devices. First, the relative location of visitors is discovered by scanning for nearby phones. If a phone is detected, the corresponding visitor has to be located in the signal range of the other phone. If multiple phones are detected, the visitor is located in the intersection of the circular areas defined by the different signal ranges. The absolute position of the user is derived by the objects that were last photographed by the visitors in his vicinity. The radii of each circular area is then consequently defined by the signal range of Bluetooth as well as the maximal covered distance since the last object was photographed based on an average walking speed.
For outdoor applications, a comparable technique based on GSM and GPS was introduced by [177]. We discarded this approach primarily due to the same challenges that we have faced during the development of our phone-to-phone communication technique. For instance, we had performance

problems during the continuous scanning for Bluetooth devices. However, we believe that this technique could be beneficial as a cost-effective localization technique that does not require any auxiliary hardware in indoor environments.

In chapter 6 we introduced sensor boxes that are equipped with omnidirectional light sensors to measure the illumination state in a room over a certain period of time. This sensor data can also be used to adapt the image classification on mobile phones. To accomplish this, we equipped the sensor boxes with Bluetooth modules to allow them to communicate with mobile devices. The visitors' mobile phones scan for these sensor boxes and receive the current illumination state based on the measured light samples. The object recognition system can then select the image classifier from a predefined set that best suites the current lighting condition.

Although we have tested our system in several field studies, which have revealed the benefits of our techniques, a long-term practical test of our system in a museum still needs to be undertaken. The main reason for this is the lack of a comprehensive database of multimedia content that would be necessary to ensure a realistic application of our system. If visitors retrieve valuable information about the exhibits, we could evaluate the users' acceptance of the application and the benefit of additional multimedia content such as pictures or videos.

A first insight into the acceptance of multimedia content was revealed by asking the participants for their opinion in the course of our pathway analysis approach (cf. section 7.5). On a scale of 1 (worst) to 7 (best), 132 visitors were asked to rate their interest in multimedia information about exhibits. The mean result was 5.14 indicating a high interest in multimedia content.

Furthermore, the object recognition process could be enhanced by a rejection technique. Currently, the object classification algorithm always provides a list of possible result candidates even if the photographed object was never trained. In a realistic scenario, visitors should be informed if no information about an exhibit is available to prevent the user searching for the correct object in the displayed list of result candidates.

From an application point of view we started to design a recommendation system that enhances the visitors' museum experience. To accomplish this, we began by implementing a recommendation system based on the collected pathways that consists of two parts: on the one hand, visitors receive recommendations of exhibits during their actual stay in the museum by continuously comparing the pathway of the current visitor with pathways of previous visitors. The recommended exhibits or areas in the museum are then highlighted on a map displayed on the mobile phone.

On the other hand visitors can retrieve tailored information before commencing their museum visit.

Therefore the visitor indicates his interests and preferences by answering a short questionnaire on the mobile phone, for example, whether he is interested in statues rather than vases or in paintings of Picasso. The pathways of previous users who have filled out the questionnaire and whose choices correspond closely to the current visitor's preferences serve as recommendations for the visitor's path through the museum. The corresponding route is then displayed on a map.

Another idea is based on an analysis of how the exhibits were captured. For example, if the exhibits are sorted based on the frequency of how often they were photographed, a ranking can be compiled. This would allow visitors to compose a path through the museum that navigates them to the most popular exhibits. After an exhibit on this list is photographed, the visitor is automatically navigated to the next object.

The collection of such information would also open up the possibility of providing statistical data for the museum directors which they could use to adapt their exhibitions. For example, by visualizing the pathways of all visitors, the museums could identify areas that are approached by visitors only sporadically. Consequently, they could exchange certain exhibits or change the guidance system in order to navigate more visitors to these places.

As mentioned earlier, the accumulation of all the necessary data for the different recommendations would be facilitated by the application of our system and the continuous collection of the adaptation parameters.

A APPENDIX

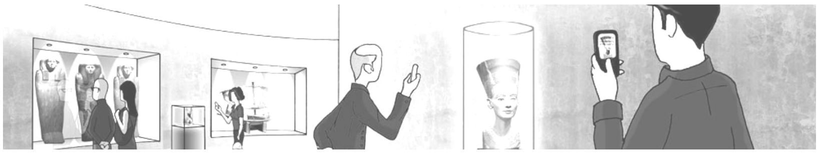
A.1 Questionnaire for Evaluating Multi-Image Classification

Alter:				
Du hast unterschiedliche Videos von Ausstellungsstücken im Museum aufgenommen, indem du das Handy von vorne nach hinten bewegt hast. Im Vergleich zum Aufnehmen eines einzelnen Fotos, wie viel besser oder schlechter war die Handhabung beim Aufnehmen eines Videos?				
viel schlechter □	□	gleich □	□	viel besser □
Würdest du das Aufnehmen eines Videos, dem Aufnehmen eines Fotos vorziehen, wenn dadurch die Erkennungsrate (mehr Objekte werden richtig erkannt) merklich steigen würde?				
ja: □ nein: □				
Hier kannst du Kommentare oder Anmerkungen hinterlassen.				

A.2 Questionnaire for Evaluating Subobject Recognition

PhoneGuide:

Mobile Phone Enabled Museum Guidance



Evaluierungsbogen für die Subobjekterkennung

Alter			Geschlecht	m	w
Handybesitzer	ja	nein			

Sie haben gerade die Subobjekt-Erkennung von PhoneGuide getestet, einem mobilen Museumsführer für Handys. Bitte bewerten Sie die Applikation anhand folgender Fragen (1=sehr schlecht, 7= sehr gut).

Wie gut kamen Sie mit der Erstellung eines Fotos zur Erkennung zurecht?	1	2	3	4	5	6	7
Wie empfanden Sie die Erkennungsrate (Wurden die gewünschten Subobjekte gefunden)?	1	2	3	4	5	6	7
Wie genau entsprach die gefundene Position der tatsächlichen Position der Subobjekte?	1	2	3	4	5	6	7
Wie fanden Sie die Übersichtlichkeit der dargestellten Ergebnisse?	1	2	3	4	5	6	7

Die Suche in PhoneGuide besteht aus zwei Phasen: Der Objektgruppenerkennung und der Subobjektdetektion und –erkennung. Der komplette Vorgang findet direkt auf dem Gerät statt, ohne zusätzliche Wartezeit und Kosten durch Verbindungen zum Provider. Die nächsten Fragen beschäftigen sich mit dem zeitlichen Aufwand.

Wie empfanden Sie die Wartezeit in der ersten Phase (1=zu lang, 7=sehr kurz)?	1	2	3	4	5	6	7

Wie lange wären Sie bereit, auf ein Ergebnis zu warten?	Unter 1s	1s – 2s	2s – 4s	4s – 6s			
Wie empfanden Sie die Wartezeit für die Subobjektdetektion (1=zu lang, 7=sehr kurz)?	1	2	3	4	5	6	7
Wie lange wären Sie bereit, für die Suche nach Subobjekten zu warten?	Unter 1s	1s – 2s	2s – 4s	4s – 6s			

Das System braucht keine Internetverbindung, d.h. es fallen keine weiteren Verbindungskosten für Sie an. Um die Erkennung zu verbessern oder um eine Vielzahl an Multimediadateien anzubieten, könnte eine Internetverbindung aufgebaut werden. Wären Sie bereit, zusätzliche Wartezeit und Kosten (abhängig vom Provider) in Kauf zu nehmen?

Welche zusätzlichen nutzungsabhängigen Kosten fänden Sie pro Museumsbesuch akzeptabel?	Keine	Unter 1€	1€-2€	2€-3€	
Würden Sie die eben präsentierte Applikation interessant finden?	Ja	Vielleicht	Nein		

Falls Sie noch weitere Anmerkungen, Kritiken oder vielleicht Ideen haben, tragen Sie diese bitte im folgenden Feld ein.

Anmerkungen

Viel Dank für Ihre Teilnahme!

A.3 Questionnaire for Analyzing Pathways

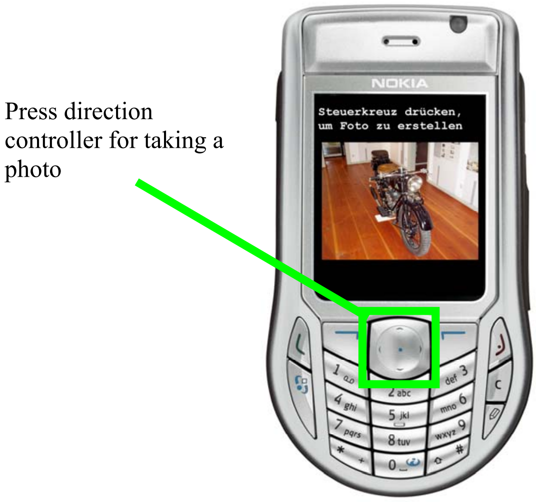
Experiment – Description

Dear participant,

In the context of a scientific study of the department of Media Systems at the Bauhaus-University Weimar, we would like to investigate the visitors' behavior in a museum. Especially pathways of the visitors can be useful to improve digital museum guidance systems.
To determine these pathways we hand out a mobile phone to you. Each time you are facing an exhibit (e.g. models, artifacts, pictures, paintings) that has caught your interest, we ask you to take a photo of the exhibit with the mobile phone (you can take as much photos as you want). This photo is stored automatically on the device and allows us to track the pathway afterwards. At the end of your visit we will ask you for answering five questions.
The collected data is only used for research purposes at the Bauhaus-University and is not distributed to others. The participation is voluntary.

Explanation:

1. Target the object
2. Press direction controller for taking a photo



Date:	
Time:	

Dear participant,

You have taken multiple photos with a mobile phone during your visit in the museum. We would now like to ask you to fill out the following questionnaire based on your experiences.
The collected data is only used for research purposes at the Bauhaus-University and is not distributed to others.

Please, fill in your personal experiences and rate them on a scale from 1 to 7 (only one cross per question):

Age:		Gender:	□ m □ f
Country:			

How satisfied have you been with the kind of information presentation (text panels) about the exhibits in this museum?

very			average			not at all
□	□	□	□	□	□	□

Estimate to what extend you would like to have alternative information presentation techniques like audio, images or videos in museums?

very			average			not at all
□	□	□	□	□	□	□

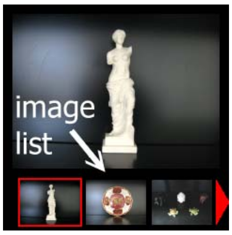
Assume that the mobile phone, you have just used, would have presented (optionally) individual acoustic (audio) and visual (text, images, videos) information to each exhibit that you photographed.
Under this assumption: Estimate to what extend you are willing to use a/your phone as museum guide (comparable to an audioguide) with the functionality mentioned above?

very			average			not at all
□	□	□	□	□	□	□

There are two possibilities to identify an exhibit with a digital device: Either you type in a reference number, which is located close to an object with your keypad (cp. audioguides), or you take a picture of the exhibit.
What kind of identification (reference number, photo) would you prefer (e.g. with your own phone)?

type in number □ take a picture □

Misclassifications might occur if object recognition is used (approx. 1 object out of 10 is misclassified). Thus, a photo can be matched to a wrong exhibit. In such cases, you have to select the correct exhibit from a short list of images (see figure) to receive the corresponding multimedia content.
Under this assumption: Estimate to what extend you are willing to use your/a mobile phone as museum guide (comparable to an audioguide) with the functionality and limitations mentioned above.



very			average			not at all
□	□	□	□	□	□	□

A.4 Estimation of Classification Rate

In section 8.3 we estimated the benefit of our multi-image classification approach for the overall classification rate because it was evaluated in the course of a different user study. To approximate the benefit, we consider the relative gain of this approach in comparison to the basic object recognition result as presented in chapter 4.

Let p_u denote the classification rate of the basic object recognition (application of a single image) of user study A and let p_a be the classification rate of the multi-image classification of user study A. The classification gain of the multi-image classification is then defined as

$$\frac{p_a - p_u}{1 - p_u}$$

Let p'_u denote the classification rate of the basic object recognition of user study B and let p'_a be the classification rate of the multi-image classification of user study B that has to be estimated. Then, p'_a can be computed by

$$\frac{p'_a - p'_u}{1 - p'_u} = \frac{p_a - p_u}{1 - p_u}$$

$$\Leftrightarrow p'_a - p'_u = (p_a - p_u)\frac{1 - p'_u}{1 - p_u}$$

$$\Leftrightarrow p'_a = p'_u + (p_a - p_u)\frac{1 - p'_u}{1 - p_u}$$

If $p_u = 0.6861$, $p_a = 0.8337$, $p'_u = 0.7882$, then

$$p'_a = 0.7882 + (0.8337 - 0.6861)\frac{1 - 0.7882}{1 - 0.6861} \sim 0.8878$$

REFERENCES

[1] Institute for museum research. *Statistische Gesamterhebung an den Museen der Bundesrepublik Deutschland für das Jahr 2006*, 2006. 87

[2] Lauri Aalto, Nicklas Göthlin, Jani Korhonen, and Timo Ojala. Bluetooth and wap push based location-aware mobile advertising system. In *proceedings of the 2nd international conference on Mobile systems, applications, and services*, pages 49–58, 2004. 34

[3] Gregory D Abowd, Christopher G Atkeson, Jason Hong, Sue Long, Rob Kooper, and Mike Pinkerton. Cyberguide: a mobile context-aware tour guide. *Wireless Networks*, 3(5):421–433, 1997. 17

[4] Shivani Agarwal, Aatif Awan, and Dan Roth. Learning to detect objects in images via a sparse, part-based representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(11):1475–1490, 2004. 32

[5] Sherif Akoush and Ahmed Sameh. Mobile user movement prediction using bayesian learning for neural networks. In *proceedings of the 2007 international conference on Wireless communications and mobile computing*, pages 191–196, New York, NY, USA, 2007. ACM. 36

[6] Adriano Albertini, Roberto Brunelli, Oliviero Stock, and Massimo Zancanaro. Communicating user's focus of attention by image processing as input for a mobile museum guide. In *proceedings of the 10th international conference on Intelligent user interfaces*, pages 299–301, New York, NY, USA, 2005. ACM. 25

[7] Katrin Amlacher, Patrick Morris Luley, Gerald Fritz, Alexander Almer, and Lucas Paletta. Mobile object recognition using multi-sensor information fusion in urban environments. In *proceedings of ICIP*, pages 2384–2387. IEEE, 2008. 28, 29

[8] Yaw Anokwa, Gaetano Borriello, Trevor Pering, and Roy Want. A user interaction model for nfc enabled applications. In *proceedings of the Fifth IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 357–361, Washington, DC, USA, 2007. IEEE Computer Society. 18

[9] Clemens Arth, Daniel Wagner, Manfred Klopschitz, Arnold Irschara, and Dieter Schmalstieg. Wide area localization on mobile phones. In *proceedings of the 2009 8th IEEE*

International Symposium on Mixed and Augmented Reality, pages 73–82, Washington, DC, USA, 2009. 22

[10] Infrared Data Association. http://www.irda.org/. 15

[11] Ronald Azuma. A survey of augmented reality. *Presence*, 6(4):355–385, 1997. 23

[12] Paramvir Bahl and Venkata N. Padmanabhan. Radar: An in-building rf-based user location and tracking system. In *proceedings of INFOCOM*, pages 775–784, 2000. 20

[13] Rafael Ballagas, Michael Rohs, and Jennifer G. Sheridan. Sweep and point and shoot: phonecam-based interactions for large public displays. In *CHI '05 extended abstracts on Human factors in computing systems*, pages 1200–1203, New York, NY, USA, 2005. ACM. 22

[14] Jakob E. Bardram. Applications of context-aware computing in hospital work: examples and design principles. In *proceedings of the 2004 ACM symposium on Applied computing*, pages 1574–1579, New York, NY, USA, 2004. ACM. 2, 1

[15] Mortaza S. Bargh and Robert de Groote. Indoor localization based on response rate of bluetooth inquiries. In Ying Zhang and Yinyu Ye, editors, *proceedings of MELT*, pages 49–54. ACM, 2008. 21

[16] John R. Barry. *Wireless Infrared Communications*. Kluwer International Series in Engineering & Computer Science, 1994. 17

[17] J. Basak, K. Bhattacharya, and S. Chaudhury. Multiple exemplar-based facial image retrieval using independent component analysis. *IEEE Transactions on Image Processing*, 15(12):3773–3783, 2006. 31

[18] X. Basogain, S. Renteria, M.A. Olabe, A. Campo, and A. Torrens. Mobile locator: Helping the context awareness. In Hans-Jörg Bullinger and Jürgen Ziegler, editors, *proceedings of Broadband IEEE International Symposium on Multimedia Systems and Broadcasting, 2009. BMSB '09.*, pages 1–5, 2009. 20

[19] Jörg Baus, Antonio Krüger, and Wolfgang Wahlster. A resource-adaptive mobile navigation system. In *proceedings of the 7th international conference on Intelligent user interfaces*, pages 15–22, New York, NY, USA, 2002. ACM. 22

[20] Jörg Baus, Keith Cheverst, and Christian Kray. *A Survey of Map-based Mobile Guides Map-based mobile services - Theories, Methods and Implementations*, chapter 13. Meng/Zipf, Springer, 2005. 15

[21] Herbert Bay, Beat Fasel, and Luc Van Gool. Interactive museum guide. *proceedings of the Seventh International Conference on Ubiquitous Computing UBICOMP, Workshop on Smart Environments and Their Applications to Cultural Heritage*, September 2005. 5, 3, 20, 28, 29

[22] Herbert Bay, Tinne Tuytelaars, and Luc J. Van Gool. Surf: Speeded up robust features. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *ECCV (1)*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer, 2006. 26, 27, 28, 31

[23] Benjamin B Bederson. Audio augmented reality: a prototype automated tour guide. In *proceedings of Conference companion on Human factors in computing systems*, pages 210–211, New York, NY, USA, 1995. ACM. 17

[24] Jeffrey S. Beis and David G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *proceedings of CVPR)*, page 1000, Washington, DC, USA, 1997. IEEE Computer Society. 30

[25] Francesco Bellotti, Riccardo Berta, Alessandro De Gloria, and Massimiliano Margarone. User testing a hypermedia tour guide. *Pervasive Computing*, 1(2):33–41, 2002. 14

[26] Benjamin Brombach, Erich Bruns, and Oliver Bimber. Subobject detection through spatial relationships on mobile phones. In Cristina Conati, Mathias Bauer, Nuria Oliver, and Daniel S. Weld, editors, *proceedings of Intelligent User Interfaces*, pages 267–276. ACM, 2009. 60

[27] Mauro Brunato and Csaba Kiss Kalló. Transparent location fingerprinting for wireless services. In *proceedings of Med-Hoc-Net 2002*, 2002. 20

[28] Roberto Brunelli. *Template Matching Techniques in Computer Vision: Theory and Practice*. John Wiley & Sons, 2009. 30

[29] Raffaele Bruno and Franca Delmastro. Design and analysis of a bluetooth-based indoor localization system. In Marco Conti, Silvia Giordano, Enrico Gregori, and Stephan Olariu, editors, *PWC*, volume 2775 of *Lecture Notes in Computer Science*, pages 711–725. Springer, 2003. 20

[30] Erich Bruns and Oliver Bimber. Adaptive training of video sets for image recognition on mobile phones. *Journal of Personal and Ubiquitous Computing*, 13(2):165–178, 2008. 9, 31, 40, 42

[31] Erich Bruns and Oliver Bimber. Phone-to-phone communication for adaptive image classification. In *proceedings of Advances in Mobile Computing and Multimedia*, pages 276–281, 2008. 78

[32] Erich Bruns and Oliver Bimber. Adaptation techniques for mobile image classification. In *proceedings of Wireless Communication and Information*, pages 235–247, 2009. 106

[33] Erich Bruns and Oliver Bimber. Localization and classification through adaptive pathway analysis. *Journal of Pervasive Computing, minor revision*, 2010. 93

[34] Erich Bruns and Oliver Bimber. Mobile museum guidance through relational multi-image classification. *Accepted at Conference of Multimedia and Ubiquitous Engineering*, 2010. 47

[35] Erich Bruns, Benjamin Brombach, and Oliver Bimber. Mobile phone enabled museum guidance with adaptive classification. *Journal of Computer Graphics and Applications*, 28(4):98–102, 2008. 106

[36] Erich Bruns, Benjamin Brombach, Thomas Zeidler, and Oliver Bimber. Enabling mobile phones to support large-scale museum guidance. *Journal of MultiMedia*, 14(2):16–25, 2007. 8, 43

[37] Jenna Burrell and Geri Gay. E-graffiti: evaluating real-world use of a context-aware system. *Interacting with Computers*, 14(4):301–312, 2002. 77

[38] C. C. Chang, P. C. Lou, and H. Y. Chen. Designing and implementing a rfid-based indoor guidance system. volume 7, pages 27–34, 2008. 17

[39] Sudarshan S. Chawathe. Beacon placement for indoor localization using bluetooth. *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems*, 2008. 21

[40] Guanling Chen and David Kotz. A survey of context-aware mobile computing research. Technical report, Dartmouth College, 2000. 13, 15, 22

[41] Tao Chen, Kui Wu, Kim-Hui Yap, Zhen Li, and Flora S. Tsai. A survey on mobile landmark recognition for information retrieval. In *proceedings of the 2009 Tenth International Con-*

ference on Mobile Data Management: Systems, Services and Middleware, pages 625–630, Washington, DC, USA, 2009. IEEE Computer Society. 26

[42] Keith Cheverst, Nigel Davies, Keith Mitchell, Adrian Friday, and Christos Efstratiou. Developing a context-aware electronic tourist guide: some issues and experiences. *proceedings of CHI*, pages 17–24, 2000. 19

[43] Hae Don Chon, Sibum Jun, Heejae Jung, and Sang Won An. Using rfid for accurate positioning, 2004. 17

[44] Omar Choudary, Vincent Charvillat, Romulus Grigoras, and Pierre Gurdjos. March: mobile augmented reality for cultural heritage. In *proceedings of the seventeen ACM international conference on Multimedia*, pages 1023–1024, New York, NY, USA, 2009. ACM. 24

[45] Cisco. Wifi location based services- design and deployment considerations. https://cisco.hosted.jivesoftware.com/docs/DOC-3418, 2008. 16, 20, 22

[46] Toll Collect. http://www.toll-collect.de/. 2, 1

[47] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *Transactions Pattern Analysis Machine Intelligence*, 24(5):603–619, 2002. 61

[48] Sajal K Das, Diane J Cook, Amiya Bhattacharya, Edwin O. Heierman III, and Tze-Yun Lin. The role of prediction algorithms in the mavhome smart home architecture. *Wireless Communications*, 9(6):77–84, 2002. 35

[49] PDA Database. http://www.pdadb.net/. iv, 44

[50] Thomas Deselaers, Daniel Keysers, and Hermann Ney. Features for image retrieval: an experimental comparison. *Information Retrieval*, 11(2):77–107, 2008. 31

[51] A.K. Dey and G.D. Abowd. Towards a better understanding of context and context-awareness. *proceedings of CHI 2000 Workshop on the What, Who, Where, When, and How of Context-Awareness*, 2000. 2, 1, 13

[52] Mario Döller, Günther Kockerandl, Simone Jans, and Lyes Limam. Moidex: Location-based mtourism system on mobile devices. In *proceedings of Multimedia Computing and Systems*, pages 199–204. IEEE, 2009. 4, 3, 27, 29, 114

[53] Richard O. Duda and Peter E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Communication ACM*, 15(1):11–15, 1972. 26

[54] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006. 36

[55] Max Egenhofer and Robert Franzosa. Point-set topological spatial relations. *Journal of Geographical Information Systems*, 5 (2):161–174, 1991. 32

[56] Max J. Egenhofer and John Herring. Categorizing binary topological relationships between regions, lines, and points in geographic databases. Technical report, University of Maine, Department of Surveying Engineering, 1991. 63

[57] Stephan Eichler, Benedikt Ostermaier, Christoph Schroth, and Timo Kosch. Simulation of car-to-car messaging: Analyzing the impact on road traffic. In *proceedings of the 13th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, pages 507–510, 2005. 34

[58] Essam A. El-Kwae and Mansur R. Kabuka. A robust framework for content-based retrieval by spatial similarity in image databases. *Transactions on Information Systems*, 17(2):174–198, 1999. 33, 66

[59] Albrecht Schmidt Enrico Rukzio, Sergej Wetzstein. A framework for mobile interactions with the physical world. In *Invited paper special session "Simplification of user access to ubiquitous ICT services" in Wireless Personal Multimedia Communication (WPMC'05)*, Aalborg, Denmark, September 2005. 18

[60] Nicolas Eude, Bertrand Ducourthial, and Mohamed Shawsky. Simulation of car-to-car messaging: Analyzing the impact on road traffic. In *proceedings of the 7th IFIP International Conference on Mobile and Wireless Communications Networks*, 2005. 34

[61] Beat Fasel and Luc Van Gool. Interactive museum guide: Accurate retrieval of object descriptions. *Adaptive Multimedia Retrieval: User, Context, and Feedback*, 4398:179–191, 2007. 27

[62] Mirko Fetter and Tom Gross. Locarhythms: Real-time data mining for continuous detection and prediction of stays. In *proceedings of the 2009 35th Euromicro Conference on Software Engineering and Advanced Applications*, pages 64–71, Washington, DC, USA, 2009. IEEE Computer Society. 36

[63] Julio Oliveira Filho, Ana Bunoza, Jürgen Sommer, and Wolfgang Rosenstiel. Self-localization in a low cost bluetooth environment. In *proceedings of the 5th international*

conference on Ubiquitous Intelligence and Computing, pages 258–270, Berlin, Heidelberg, 2008. Springer-Verlag. 20

[64] Klaus Finkenzeller. *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards and Identification*. Wiley, 2003. 15, 16

[65] George W. Fitzmaurice. Situated information spaces and spatially aware palmtop computers. *Communication ACM*, 36(7):39–49, 1993. 1

[66] Rob Flickenger, Corinna ”Elektra” Aichele, Carlo Fonda, Jim Forster, Ian Howard, Tomas Krag, and Marco Zennaro. *Wireless Networking in the Developing World*. Hacker Friendly LLC, 2006. 19

[67] Paul Föckler, Thomas Zeidler, Benjamin Brombach, Erich Bruns, and Oliver Bimber. Phoneguide: museum guidance supported by on-device object recognition on mobile phones. In *proceedings of Mobile and ubiquitous multimedia*, pages 3–10, 2005. 8, 40

[68] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centers of circular features. *ISPRS Intercommission Workshop on Fast Processing of Photogrammetric Data*, pages 149–155, 1987. 32

[69] NFC Forum. http://www.nfc-forum.org/. 15

[70] E. Frias-Martinez, G. Magoulas, S. Chen, and R. Macredie. Modeling human behavior in user-adaptive systems: Recent advances using soft computing techniques. *Expert Systems with Applications*, 29(2):320–329, August 2005. 36

[71] Gerald Fritz, Christin Seifert, and Lucas Paletta. A mobile vision system for urban detection with informative local descriptors. In *proceedings of the Fourth IEEE International Conference on Computer Vision Systems*, page 30, Washington, DC, USA, 2006. IEEE Computer Society. 3, 5, 2, 24, 25, 28, 29

[72] Ada Wai-chee Fu, Polly Mei-shuen Chan, Yin-Ling Cheung, and Yiu Sang Moon. Dynamic vp-tree indexing for n-nearest neighbor search given pair-wise distances. *The VLDB Journal*, 9(2):154–173, 2000. 51

[73] F. Gabrielli, P. Marti, and L. Petroni. The environment as interface. *Proceedings of the i3 Annual Conference: Community of the Future*, pages 44–47, 1999. 92

[74] Gartner. Dataquest insight: The top 10 consumer mobile applications in 2012. *http://www.gartner.com/it/page.jsp?id=1230413*, 2009. 3, 2

[75] David Gavilan, Hiroki Takahashi, Suguru Saito, and Masayuki Nakajima. Mobile image retrieval using morphological color segmentation. In *International Conference on Mobile Computing and Ubiquitous Networking*, 2006. 5, 3, 25, 29

[76] GS1 Germany GmbH. http://www.gs1-germany.de/. 23

[77] Google Goggles. http://www.google.com/mobile/goggles/. 27

[78] Bluetooth Special Interest Group. http://www.bluetooth.com/english/press/pages/pressreleases-detail.aspx?id=4. 90

[79] Anwar M. Haneef and Aura Ganz. Mobile agent based network access for mobile electronic guidebooks. In *proceedings of the International Workshop on Mobility and Wireless Access*, page 22, Washington, DC, USA, 2002. IEEE Computer Society. 14

[80] Rene Hansen, Rico Wind, Christian S Jensen, and Bent Thomsen. Seamless indoor/outdoor positioning handover for location-based services in streamspin. In *proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, pages 267–272, Washington, DC, USA, 2009. IEEE Computer Society. 22

[81] Jonathan S Hare and Paul H Lewis. Content-based image retrieval using a mobile device as a novel interface. In *proceedings of Storage and Retrieval Methods and Applications for Multimedia*, pages 64–75, 2004. 4, 2, 25, 29

[82] Andy Harter, Andy Hopper, Pete Steggles, Andy Ward, and Paul Webster. The anatomy of a context-aware application. In *proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*, pages 59–68, New York, NY, USA, 1999. ACM. 21

[83] X. C. He and N. H. C. Yung. Curvature scale space corner detector with adaptive threshold and dynamic region of support. In *proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2*, pages 791–794, Washington, DC, USA, 2004. IEEE Computer Society. 26

[84] Niels Henze, Torben Schinke, and Susanne Boll. What is that? object recognition from natural features on a mobile phone. *proceedings of Workshop on Mobile Interaction with the Real World*, 2009. 4, 3, 28

[85] Jeffrey Hightower and Gaetano Borriello. Location systems for ubiquitous computing. *Computer*, 34(8):57–66, 2001. 15

[86] Atsushi Hiyama, Jun Yamashita, Hideaki Kuzuoka, Koichi Hirota, and Michitaka Hirose. Position tracking using infra-red signals for museum guiding system. In Hitomi Murakami, Hideyuki Nakashima, Hideyuki Tokuda, and Michiaki Yasumura, editors, *proceedings of Ubiquitous Computing Systems*, volume 3598 of *Lecture Notes in Computer Science*, pages 49–61. Springer, 2004. 17

[87] Thomas Hofer, Wieland Schwinger, Mario Pichler, Gerhard Leonhartsberger, Josef Altmann, and Werner Retschitzegger. Context-awareness on mobile devices - the hydrogen approach. In *proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 9*, page 292.1, Washington, DC, USA, 2003. IEEE Computer Society. 4, 3

[88] Albert S. Huang and Larry Rudolph. *Bluetooth Essentials for Programmers*. Cambridge University Press, 2007. 20

[89] Yo-Ping Huang, Tsun-Wei Chang, and Frode Eika Sandnes. A ubiquitous interactive museum guide. In *proceedings of the 5th international conference on Ubiquitous Intelligence and Computing*, pages 720–731, Berlin, Heidelberg, 2008. Springer-Verlag. 16

[90] Jonathan J. Hull, Xu Liu, Berna Erol, Jamey Graham, and Jorge Moraleda. Mobile image recognition: architectures and tradeoffs. In *proceedings of the Eleventh Workshop on Mobile Computing Systems &#38; Applications*, pages 84–88, New York, NY, USA, 2010. ACM. 28

[91] IDTechEx. Rfid forecasts, players & opportunities 2009-2019. *http://www.idtechex.com/research/articles/rfid_market_forecasts_2009_2019_00001377.asp*, 2009. 2, 1

[92] InfoTrends. Worldwide camera phone forecast: 2007-2012. *http://www.capv.com/public/Content/Press/2008/07.29.2008.html*, 2008. 2

[93] D. N. F. Awang Iskandar, James A. Thom, and Seyed M. M. Tahaghoghi. Content-based image retrieval using image regions as query examples. In Alan Fekete and Xuemin Lin, editors, *proceedings of ADC*, volume 75 of *CRPIT*, pages 39–47. Australian Computer Society, 2008. 32

[94] J.L. Izkara, X. Basogain, and D. Borro. Wearable personal assistants for the management of historical centers. *International Journal of Architectural Computing*, 7(1):139–156, 2009. 24

[95] Xiangyu Jin and James C. French. Improving image retrieval effectiveness via multiple queries. In *proceedings of of the 1st ACM international workshop on Multimedia databases*, pages 86–93, New York, NY, USA, 2003. ACM Press. 31, 46

[96] Sewook Jung, Uichin Lee, Alexander Chang, Dae-Ki Cho, and Mario Gerla. Bluetorrent: Cooperative content sharing for bluetooth users. *Pervasive and Mobile Computing*, 3(6):609–634, 2007. 34

[97] Kamol Kaemarungsi and Prashant Krishnamurthy. Modeling of indoor positioning systems based on location fingerprinting. In *INFOCOM*, 2004. 20

[98] Donnie H. Kim, Jeffrey Hightower, Ramesh Govindan, and Deborah Estrin. Discovering semantically meaningful places from pervasive rf-beacons. In Sumi Helal, Hans Gellersen, and Sunny Consolvo, editors, *proceedings of UbiComp*, ACM International Conference Proceeding Series, pages 21–30. ACM, 2009. 19

[99] André Klahold. *Empfehlungssysteme: Grundlagen, Konzepte und Systeme*. Vieweg+Teubner, 2009. 105

[100] Boriana Koleva, Stefan Rennick Egglestone, Holger Schnädelbach, Kevin Glover, Chris Greenhalgh, Tom Rodden, and Martyn Dade-Robertson. Supporting the creation of hybrid museum experiences. In *proceedings of the 27th international conference on Human factors in computing systems*, pages 1973–1982, New York, NY, USA, 2009. ACM. 24

[101] Krzysztof Kolodziej and Hjelm Johan. *Local Positioning Systems: LBS applications and services*. CRC Press Inc, 2006. 15

[102] Kooaba. http://www.kooaba.com/. 27

[103] Antti Kotanen, Marko Hännikäinen, Helena Leppäkoski, and Timo Hämäläinen. Experiments on local positioning with bluetooth. In *ITCC*, pages 297–303. IEEE Computer Society, 2003. 21

[104] Jens Krösche, Susanne Boll, and Jörg Baldzer. Mobidenk – mobile multimedia in monument conservation. volume 11(2), pages 72–77, 2004. 22

[105] Uwe Kubach and Kurt Rothermel. Exploiting location information for infostation-based hoarding. In *proceedings of the 7th annual international conference on Mobile computing and networking*, pages 15–27, New York, NY, USA, 2001. ACM. 35

[106] H. Laitinen, J. Lahteenmaki, and T. Nordstrom. Database correlation method for gsm location. volume 4, pages 2504–2508 vol.4, 2001. 22

[107] Layar. http://www.layar.com/. 2, 1

[108] Jimmy Addison Lee and Kin Choong Yow. Image recognition for mobile applications. In *proceedings of ICIP*, pages 177–180. IEEE, 2007. 26, 29

[109] J.P. Lewis. Fast normalized cross-correlation. In *proceedings of Vision Interface*, pages 120–123, 1995. 61

[110] Binghao Li, Yufei Wang, Hyung Keun Lee, Andrew Dempster, and Chris Rizos. Method for yielding a database of location fingerprints in wlan. *Communications*, 152(5):580–586, 2005. 20

[111] Lin Liao, Dieter Fox, Jeffrey Hightower, Henry Kautz, and Dirk Schulz. Voronoi tracking: Location estimation using sparse and noisy sensor data. In *proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS*, 2003. 98

[112] Joo-Hwee Lim, Jean-Pierre Chevallet, and Sihem Nouarah Merah. An augmented reality museum guide. In *proceedings of Mobile & Ubiquitous Information Access*, 2004. 26, 29

[113] Dimitri A. Lisin, Marwan A. Mattar, Matthew B. Blaschko, Erik G. Learned-Miller, and Mark C. Benfield. Combining local and global image features for object class recognition. In *proceedings of CVPR - Workshops*, page 47, Washington, DC, USA, 2005. IEEE Computer Society. 30

[114] LookTel. http://www.looktel.com/. 27

[115] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Journal of Computer Vision*, 60(2):91–110, 2004. 25, 26, 28, 30, 49, 61, 114

[116] Patrick Morris Luley, Lucas Paletta, and Alexander Almer. Visual object detection from mobile phone imagery for context awareness. *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, pages 385–386, 2005. 5, 3, 24, 29

[117] Ping Luo, Hui Xiong, Kevin Lü, and Zhongzhi Shi. Distributed classification in peer-to-peer networks. In *proceedings of the 13th international conference on Knowledge discovery and data mining*, pages 968–976, 2007. 34

[118] Wanji Mai, Gordon Dodds, and Chris Tweed. A pda-based system for recognizing buildings from user-supplied images. In Fabio Crestani, Mark D. Dunlop, and Stefano Mizzaro, editors, *proceedings of Mobile HCI Workshop on Mobile and Ubiquitous Information Access*, volume 2954 of *Lecture Notes in Computer Science*, pages 143–157. Springer, 2003. 26, 29

[119] Wanji Mai, Chris Tweed, and Peter Hung. Building identification by low-resolution mobile images. In Gabriele Kotsis, David Taniar, Eric Pardede, and Ismail Khalil Ibrahim, editors, *proceedings of Advances in Mobile Computing and Multimedia*, volume 230 of *books@ocg.at*, pages 15–30. Austrian Computer Society, 2007. 27

[120] The MathWorks. http://www.mathworks.com/. 109

[121] Sven Meyer and Andry Rakotonirainy. A survey of research on context-aware homes. In *proceedings of the Australasian information security workshop conference on ACSW frontiers 2003*, pages 159–168, Darlinghurst, Australia, Australia, 2003. Australian Computer Society, Inc. 2, 1

[122] Damianos Gavalas Michael Kenteris and Daphne Economou. Electronic mobile guides: a survey. *Personal and Ubiquitous Computing*, 2010. 15

[123] C. Millet, I. Bloch, P. Hede, and P.-A. Moellic. Using relative spatial relationships to improve individual region recognition. In *proceedings of Integration of Knowledge, Semantic and Digital Media Technologies*, pages 119–126, 2005. 33

[124] T Miyashita, P Meier, T Tachikawa, S Orlic, T Eble, V Scholz, A Gapel, O Gerl, S Arnaudov, and S Lieberknecht. An augmented reality museum guide. In *proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 103–106, Washington, DC, USA, 2008. IEEE Computer Society. 28

[125] MobileTag. http://www.mobiletag.com. 23

[126] Hitomi Murakami, Atsushi Ito, Yu Watanabe, and Takao Yabe. Mobile phone based ad hoc network using built in bluetooth for ubiquitous life. pages 137–146, 2007. 34

[127] Kavitha Muthukrishnan, Maria Eva Lijding, and Paul J. M. Havinga. Towards smart surroundings: Enabling techniques and technologies for localization. In Thomas Strang and Claudia Linnhoff-Popien, editors, *proceedings of Location-and Context-Awareness*, volume 3479 of *Lecture Notes in Computer Science*, pages 350–362. Springer, 2005. 15

[128] Tamer Nadeem, Sasan Dashtinezhad, Chunyuan Liao, and Liviu Iftode. Trafficview: traffic data dissemination using car-to-car communication. *proceedings of SIGMOBILE Mobile Computing and Communications Review*, 8(3):6–19, 2004. 34

[129] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *proceedings of CVPR*, pages 2161–2168, Washington, DC, USA, 2006. IEEE Computer Society. 30, 51

[130] Bill Schilit Norman, Norman Adams, and Roy Want. Context-aware computing applications. In *proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications*, pages 85–90, Washington, DC, USA, 1994. IEEE Computer Society. 2, 1, 13

[131] Specification of the Bluetooth System. http://www.bluetooth.com/, 2003. 84

[132] Reinhard Oppermann and Marcus Specht. Adaptive mobile museum guide for information and learning on demand. In Hans-Jörg Bullinger and Jürgen Ziegler, editors, *proceedings of HCI*, pages 642–646. Lawrence Erlbaum, 1999. 18

[133] Reinhard Oppermann and Marcus Specht. A nomadic information system for adaptive exhibition guidance. In *ICHIM*, pages 103–109, 1999. 77

[134] Dimitris Papadias and Yannis Theodoridis. Spatial relations, minimum bounding rectangles, and spatial data structures. *Journal of Geographical Information Science*, 11(2):111–138, 1997. 32

[135] Stelios Papakonstantinou and Vesna Brujic-Okretic. Framework for context-aware smartphone applications. *The Visual Computer*, 25(12):1121–1132, 2009. 22

[136] NFC Forum (White paper). Near field communication and the nfc forum: The keys to truly interoperable communications, 2007. 18

[137] Thang V. Pham and Arnold W. M. Smeulders. Learning spatial relations in object recognition. *Pattern Recognition Letters*, 27(14):1673–1684, 2006. 32

[138] Nicolas Pinto, David D Cox, and James J DiCarlo. Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4(1):151–156, 2008. 4, 3

[139] Hubert Piontek, Matthias Seyffer, and Jörg Kaiser. Improving the accuracy of ultrasound-based localisation systems. *Personal Ubiquitous Computing*, 11(6):439–449, 2007. 21

[140] Nokia Point and Find. http://pointandfind.nokia.com/. 27

[141] Nissanka B. Priyantha, Anit Chakraborty, and Hari Balakrishnan. The cricket location-support system. In *proceedings of the 6th annual international conference on Mobile computing and networking*, pages 32–43, New York, NY, USA, 2000. ACM. 21

[142] Nissanka B. Priyantha, Allen KL Miu, Hari Balakrishnan, and Seth Teller. The cricket compass for context-aware mobile applications. In *proceedings of the 7th annual international conference on Mobile computing and networking*, pages 1–14, New York, NY, USA, 2001. ACM. 21

[143] Omer Rashid, Will Bamford, Paul Coulton, and Reuben Edwards. Pac-lan: the human arcade. In *proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*, page 33, New York, NY, USA, 2006. ACM. 16

[144] Omer Rashid, Will Bamford, Paul Coulton, Reuben Edwards, and Jurgen Scheible. Pac-lan: mixed-reality gaming with rfid-enabled mobile phones. *Computers in Entertainment*, 4(4):4, 2006. 16

[145] Nishkam Ravi and Liviu Iftode. Fiatlux: Fingerprinting rooms using light intensity. In *proceedings of of the Fifth International Conference on Pervasive Computing (Pervasive)*, 2007. 21

[146] Nishkam Ravi, Pravin Shankar, Andrew Frankel, Ahmed Elgammal, and Liviu Iftode. Indoor localization using camera phones. In *proceedings of the Seventh IEEE Workshop on Mobile Computing Systems & Applications*, page 19, Washington, DC, USA, 2006. IEEE Computer Society. 21, 91

[147] Nishkam Ravi, Peter Stern, Niket Desai, and Liviu Iftode. Accessing ubiquitous services using smart phones. In *proceedings of the 3rd IEEE International Conference on Pervasive Computing and Communications*, pages 383–393, 2005. 34, 83

[148] Jun Rekimoto and Katashi Nagao. The world through the computer: computer augmented interaction with real world environments. In *proceedings of the 8th annual ACM symposium on User interface and software technology*, pages 29–36, New York, NY, USA, 1995. ACM. 24

[149] Jörg Roth. Context-aware web applications using the pinpoint infrastructure. In *proceedings of ICWI*, pages 3–10. IADIS, 2002. 22

[150] Walter Rudametkin, Lionel Touseau, Maroula Perisanidi, Andrés Gómez, and Didier Donsez. Nfcmuseum: an open-source middleware for augmenting museum exhibits. In *proceedings of International Conference on Pervasive Services*, 2008. 18

[151] Boris Ruf, Effrosyni Kokiopoulou, and Marcin Detyniecki. Mobile museum guide based on fast SIFT recognition. In *proceedings of 6th International Workshop on Adaptive Multimedia Retrieval*, Lecture Notes in Computer Science. Springer, 2008. Compares the performance (computational time and matching accuracy) of SIFT and SURF features. 26, 29

[152] Enrico Rukzio. *Physical Mobile Interactions: Mobile Devices as Pervasive Mediators for Interactions with the Real World.* PhD thesis, Ludwig-Maximilians-Universität München, 2006. 14, 22

[153] Enrico Rukzio, Karin Leichtenstern, Victor Callaghan, Paul Holleis, Albrecht Schmidt, and Jeannette Shiaw-Yuan Chin. An experimental comparison of physical mobile interaction techniques: Touching, pointing and scanning. In Paul Dourish and Adrian Friday, editors, *proceedings of Ubicomp*, volume 4206 of *Lecture Notes in Computer Science*, pages 87–104. Springer, 2006. 3, 2, 22

[154] Corina Sas and Ronan Reilly. G.: A connectionist model of spatial knowledge acquisition in a virtual environment. In *proceedings of 2nd Workshop on Machine Learning, Information Retrieval and User Modeling, 2003*, pages 40–48, 2003. 36

[155] Dieter Schmalstieg and Wagner Daniel. A handheld augmented reality museum guide. In *proceedings of IADIS International Conference on Mobile Learning*, 2005. 24

[156] Albrecht Schmidt, Michael Beigl, and Hans W. Gellersen. There is more to context than location. *Computers and Graphics*, 23(6):893–901, 1999. 3, 2

[157] W. Schwinger, Ch. Grün, B. Pröll, W. Retschitzegger, and A. Schauerhuber. Context-awareness in mobile tourism guides. Technical report, Johannes Kepler University Linz, 2005. 15

[158] Andrea Selinger and Randal C. Nelson. Appearance-based object recognition using multiple views. In *proceedings of CVPR*, pages 905–911. IEEE Computer Society, 2001. 46

[159] Hao Shao, Tomás Svoboda, Vittorio Ferrari, Tinne Tuytelaars, and Luc J. Van Gool. Fast indexing for image retrieval based on local appearance with re-ranking. In *proceedings of ICIP*, pages 737–740, 2003. 51

[160] Stefan Siersdorfer and Sergej Sizov. Automatic document organization in a p2p environment. In *proceedings of European Conference on Information Retrieval*, pages 265–276, 2006. 35

[161] Michael Skalak, Jinyu Han, and Bryan Pardo. Speeding melody search with vantage point trees. In *proceedings of the 9th International Conference on Music Information Retrieval*, pages 95–100, 2008. 51

[162] Snaptell. http://snaptell.com/. 27

[163] Prashant Solanki and Huosheng Hu. Techniques used for location-based services: A survey. Technical report, University of Essex, 2005. 15

[164] A. Souissi, H. Tabout, and A. Sbihi. Mir system for mobile information retrieval by image querying. *Pervasive Computing*, 8(4):65–73, 2008. 26, 29

[165] Oliviero Stock, Massimo Zancanaro, Paolo Busetta, Charles Callaway, Antonio Krüger, Michael Kruppa, Tsvi Kuflik, Elena Not, and Cesare Rocchi. Adaptive, intelligent presentation of information for the museum visitor in peach. *User Modeling and User-Adapted Interaction*, 17(3):257–304, 2007. 25

[166] Michael J Swain and Dana H Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991. 26

[167] Duy-Nguyen Ta, Wei-Chao Chen, Natasha Gelfand, and Kari Pulli. Surftrac: Efficient tracking and continuous object recognition using local feature descriptors. In *proceedings of CVPR*, pages 2937–2944. IEEE, 2009. 28

[168] S. M. M. Tahaghoghi, James A. Thom, and Hugh E. Williams. Are two pictures better than one? In *proceedings of the 12th Australasian database conference*, pages 138–144, Washington, DC, USA, 2001. IEEE Computer Society. 31, 46

[169] Gabriel Takacs, Vijay Chandrasekhar, Natasha Gelfand, Yingen Xiong, Wei-Chao Chen, Thanos Bismpigiannis, Radek Grzeszczuk, Kari Pulli, and Bernd Girod. Outdoor augmented reality on mobile phone using loxel-based visual feature organization. In *proceedings of Multimedia Information Retrieval*, 2008. 4, 3, 28, 29, 31, 114

[170] Kai Ming Ting and Ian H. Witten. Issues in stacked generalization. *Artificial Intelligence Research*, 10(1):271–289, 1999. 105, 117

[171] Konrad Tollmar, Tom Yeh, and Trevor Darrell. Ideixis: image-based deixis for finding location-based information. In *proceedings of Mobile HCI*, 2004. 27

[172] James Trulove. *Build Your Own Wireless LAN*. McGraw-Hill Professional, 2002. 19

[173] Esa Tuulari and Arto Ylisaukko-oja. Soapbox: A platform for ubiquitous computing research and applications. In Friedemann Mattern and Mahmoud Naghshineh, editors, *proceedings of Pervasive*, volume 2414 of *Lecture Notes in Computer Science*, pages 125–138. Springer, 2002. 17

[174] Tinne Tuytelaars and Krystian Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008. 3

[175] Stollmann Entwicklungs und Vertriebs-GmbH. http://www.stollmann.de. iv, 43

[176] Alex Varshavsky, Eyal de Lara, Jeffrey Hightower, Anthony LaMarca, and Veljo Otsason. Gsm indoor localization. *Pervasive Mobile Computing*, 3(6):698–720, 2007. 22

[177] Alex Varshavsky, Denis Pankratov, John Krumm, and Eyal Lara. Calibree: Calibration-free localization using relative distance estimations. In *proceedings of the 6th International Conference on Pervasive Computing*, pages 146–161, Berlin, Heidelberg, 2008. Springer-Verlag. 121

[178] Andrea Vedaldi. http://www.vlfeat.org/ vedaldi/code/sift.html/. 113

[179] E. Veron and M. Levasseur. Ethnographie de l'exposition. *Bibliothèque publique d'Information*, 1989. 92

[180] Lucian Vintan, Arpad Gellert, Jan Petzold, and Theo Ungerer. Person movement prediction using neural networks. In *proceedings of Workshop on Modeling and Retrieval of Context*, 2004. 36, 95, 97

[181] Paul Viola and Michael Jones. Robust real-time object detection. Technical report, Compaq CRL, 2002. 65

[182] Pasi Välkkynen and Timo Tuomisto. Physical browsing research. In Enrico Rukzio, Jonna Häkkilä, Mirjana Spasojevic, Jani Mäntyjärvi, and Nishkam Ravi, editors, *proceedings of PERMID*, pages 35–38. LMU Munich, 2005. 17

[183] Philipp Vorst, Jürgen Sommer, Christian Hoene, Patrick Schneider, Christian Weiss, Timo Schairer, Wolfgang Rosenstiel, Andreas Zell, and Georg Carle. Indoor positioning via three different rf technologies. In *4th European Workshop on RFID Systems and Technologies (RFID SysTech 2008)*, number 209 in ITG-Fachbericht, Freiburg, Germany, June 10-11 2008. VDE Verlag. 22

[184] Daniel Wagner, Gerhard Reitmayr, Alessandro Mulloni, Tom Drummond, and Dieter Schmalstieg. Pose tracking from natural features on mobile phones. In *proceedings of ISMAR*, pages 125–134. IEEE, 2008. 28, 58

[185] Nayer M Wanas and Mohamed S Kamel. Decision fusion in neural network ensembles. *proceedings of International Joint Conference on Neural Networks*, 4:2952–2957, 2001. 40, 47, 63, 91, 98

[186] Alf Inge Wang, Michael Sars Norum, and Carl-Henrik Wolf Lund. Issues related to development of wireless peer-to-peer games in j2me. In *proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services*, page 115, 2006. 34

[187] Roy Want. An introduction to rfid technology. *Pervasive Computing*, 5(1):25, 2006. 16

[188] Roy Want, Andy Hopper, ao Veronica Falc and Jonathan Gibbons. The active badge location system. *Transactions on Information Systems*, 10(1):91–102, 1992. 17

[189] Mark Ward, Ronald Azuma, Robert Bennett, Stefan Gottschalk, and Henry Fuchs. A demonstrated optical tracker with scalable work area for head-mounted display systems. In *proceedings of the 1992 symposium on Interactive 3D graphics*, pages 43–52, New York, NY, USA, 1992. ACM. 17

[190] Denso Wave. http://www.denso-wave.com/qrcode/. 23

[191] Mark Weiser. The computer for the 21st century. *Scientific American*, 265(3):94–104, 1991. 2, 1, 13

[192] Kai Wendl, Marcus Berbig, and Patrick Robertson. Indoor localization with probability density functions based on bluetooth. In *proceedings of 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 160–164. 21

[193] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christopher von der Malsburg. Face recognition by elastic bunch graph matching. *Journal of Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997. 33

[194] Jason Wither, Stephen DiVerdi, and Tobias Höllerer. Annotation in outdoor augmented reality. *Computers & Graphics*, 33(6):679–689, 2009. 77

[195] Ran Wolff and Assaf Schuster. Association rule mining in peer-to-peer systems. In *proceedings of the 3rd IEEE International Conference on Data Mining*, page 363, 2003. 34

[196] Peter N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms*, pages 311–321, Philadelphia, PA, USA, 1993. Society for Industrial and Applied Mathematics. 50

[197] Xiang S. Zhou and Thomas S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, April 2003. 31

[198] Mustafa Özuysal, Pascal Fua, and Vincent Lepetit. Fast keypoint recognition in ten lines of code. In *proceedings of CVPR*. IEEE Computer Society, 2007. 28