Dubletten- und Ähnlichkeitssuche im LostArt-System

Duplicate Search and Visual Image Retrieval in the LostArt System

Ingolf Geist, Eike Schallehn, Nadine Schulz
Otto-von-Guericke Universität
Universitätsplatz 2
39104 Magdeburg
Tel.: +49 391 67 18800 Fax: +49 391 67 12020
E-Mail: {geist|eike|nschulz}@cs.uni-magdeburg.de

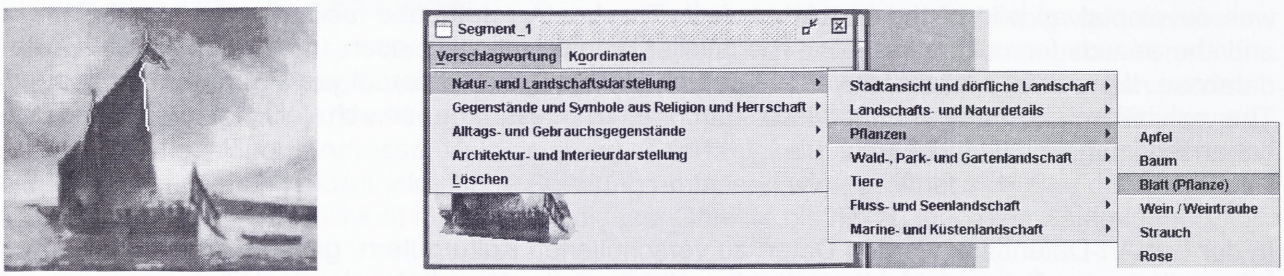
Zusammenfassung:
Die Recherche nach Kulturgütern, die in Folge des zweiten Weltkrieges und des Nationalsozialismus geraubt wurden oder verloren gingen, ist auch heute noch eine aktuelle Aufgabe – nicht nur für Kunsthistoriker, sondern auch für betroffene Privatpersonen, Institutionen und die Politik. Aus diesem Grund wurde im Rahmen des LostArt-Projektes eine Datenbank aufgebaut, die eine Vielzahl von Informationen zu den registrierten Kulturgütern verwaltet. Diese Informationen wurden als Anwendung in Form einer Web-Datenbank der Öffentlichkeit zugänglich gemacht. Dabei werden verschiedene Such- und Navigationsmodi in verschiedenen Sprachen angeboten. An dieser Stelle werden erweiterte Recherchemöglichkeiten vorgestellt. Dazu gehört die Dublettensuche sowie die erweiterte Ähnlichkeitssuche.

Abstract:
The search for cultural assets, that are lost as a result of persecution by the Nazi regime and World War II, is an ongoing task. This is not only of interest for art historians but also for private persons, various institutions as well as for politicians. In order to support these inquiries a database was developed as part of the LostArt project. The LostArt database facilitates the registration of and the search for cultural assets. The available information is open to the public via a web-database. It provides support for various search and navigation alternatives in different languages. The exhibition shows two enhanced search alternatives: the search for duplicates and the enhanced image similarity search.

In der LostArt-Datenbank werden Daten zu verschollenen Kulturgütern gesammelt. Dabei werden Informationen zu Sammelobjekten verwaltet, die im Fall von vorliegenden Detailinformationen mit Informationen zu Einzelobjekten ergänzt werden. Zu diesen Detailinformationen zählen die teilweise vorhandenen bildlichen Darstellungen der Objekte. Bereits im LostArt-Projekt umgesetzt, ist die Suche nach Bildern und Kulturobjekten auf Basis textueller Informationen, wie Titel, Künstler, Herkunft, etc. Nachteilig an diesem textbasierten Retrieval ist jedoch, dass die manuelle Beschreibung der Bilder sehr aufwändig und durch mögliche Tippfehler fehleranfällig ist. Hinzu kommt die Subjektivität der Beschreibungen von Abbildungen, denn der Bildinhalt kann von verschiedenen Personen auf unterschiedliche Weise interpretiert werden. Aus diesem Grund bietet LostArt zusätzlich zur textuellen Suche eine inhaltsbasierte Suche an. Sie ermöglicht eine textuell unabhängige Suche. Eine Erweiterung davon stellt die Suche nach einzelnen in einer Abbildung enthaltenen Objekten sowie deren Lagebeziehungen dar. Für diesen Zweck werden aus den Bilddaten spezielle Bildeigenschaften (Features), wie z.B. Farbe, Textur oder Struktur, extrahiert. Die Ähnlichkeitssuche erlaubt somit bereits eine einfache Dublettensuche auf Basis der Abbildungen.
Neben der Ähnlichkeitssuche auf Bilddaten ist es notwendig, Dubletten auch mit Hilfe von Textbeschreibungen zu finden, da nicht stets zu jedem Kulturobjekt eine Abbildung vorliegt. Diese Aufgabe wird im Lostart-System an drei Stellen benötigt:

1. Bei der Abgleichung von Fund- und Suchdaten können die Texteingaben leichte Unterschiede in der Beschreibung haben, die durch Tippfehler oder unterschiedliche Quellen entstanden sind.
2. Beim Eingang von neuen Meldungen muss überprüft werden, ob diese Objekte schon in der Datenbasis vorhanden sind. Hierbei treten die gleichen Probleme auf wie im vorhergehenden Punkt.
3. Das dritte Anwendungsgebiet liegt im LostArt-Mediator-Projekt. Hier werden verschiedene Kulturgut-Webdaten integriert, so dass ebenfalls Duplikate, die durch Mehrfacheintragungen in verschiedenen Datenbanken entstehen, entfernt werden können.

Die Dublettensuche basiert auf Zeichenkettenbasis, d.h. Objekte, die in gleichen Attributen (z.B. Titel, Autor) ähnliche Zeichenketten aufweisen, werden als identische Objekte erkannt. Die Ähnlichkeit basiert dabei entweder auf der Editierdistanz oder einer einfachen Textähnlichkeit. Die Editierdistanz zwischen zwei Zeichenketten entspricht der Anzahl der Editierschritte (Löschen, Einfügen, Ändern), die nötig sind, um eine Zeichenkette in eine zweite zu überführen. Im LostArt-System ist diese Funktionalität in der internen Verwaltung mit Hilfe eines relationalen Datenbanksystems realisiert. Diese Art der Dublettenerkennung ist sehr rechenintensiv und wird daher im Batchbetrieb ausgeführt. Nachdem der Nutzer die Suche parametrisiert und angestoßen hat, läuft diese im Hintergrund ab. Anschließend werden die gefundenen, möglichen Übereinstimmungen vom Nutzer überprüft.
Um eine erweiterte Ähnlichkeitssuche unterstützen zu können, wurde für die interne Anwendung des LostArt-Systems ein zusätzliches Programmmodul entwickelt. Dieses Modul unterstützt eine semi-automatische Segmentierung und Schlagwortungbildung der Abbildungen, so dass einzelne Bildobjekte in das Retrieval integriert werden können. Abbildung 1a zeigt, wie durch Setzten weniger Randpunkte, Objekte in den Abbildungen segmentiert werden können. Daran anschließend werden die Objekte entsprechend einer vorgegeben Begriffshierarchie mit einem Schlagwort versehen (Abbildung 1b). Die Lagebeziehungen aller segmentierten Objekte im Bild werden bestimmt und in der Datenbank gespeichert. Ferner wird für jedes Objekt das Schlagwort und die automatisch extrahierten Features in der Datenbank abgelegt.



a) b)

Abbildung 1: a) Segmentierung von Objekten in einer Abbildung
b) Zuordnung eines Schlagwortes zu einem Segment entsprechend einer vorgegebenen Begriffshierarchie

Für die Recherche wurde die Web-Anwendung entsprechend erweitert. Dabei kann der Nutzer entweder ein Bild von seiner Festplatte auswählen und anschließend das gewünschte Objekt segmentieren oder er wählt ein Bild bzw. Objekt aus der Datenbank. Für ein gewähltes Nutzerbild ist es notwendig, die Features aus den Bilddaten zu extrahieren. Die Ähnlichkeitssuche erfolgt auf Basis dieser Features.
Des Weiteren wird eine räumliche Suche auf Basis der Lagebeziehungen und der Schlagworte angeboten. Hierbei hat der Nutzer die Möglichkeit, verschiedene Objekte beliebig zu arrangieren und mit Schlagworten zu versehen. Jedem Objekt wird eine ikonisierte Abbildung zugeordnet. Nachdem der Nutzer mindestens zwei Objekte platziert hat, werden die Lagebeziehungen ermittelt und diese zusammen mit den Schlagworten und den Informationen in der Datenbank verglichen. Die Ergebnisse der räumlichen Anfrage werden ebenso wie bei der Ähnlichkeitssuche in einem Browserfenster dargestellt.