Training Argus. Ansätze zum automatischen Sehen in der Kunstgeschichte

Argus war zwar ein ‚Allesseher', doch eben kein ‚Allesversteher'. Weder erkannte er Merkur, der sich als Hirte ausgab, noch durchschaute er die List, dass er mit Geschichten und Flötenspiel eingeschläfert werden sollte. Es wurde Argus zum Verhängnis, dass er schlecht instruiert war. Dadurch konnte seine überlegene Sehkraft und übrigens auch seine Stärke und Sturheit seinen Tod letztlich nicht verhindern. Das Bild des Argus passt in vielerlei Hinsicht auf die Sehen lernende Maschine. Das zugehörige Fachgebiet, die *Computer Vision*, nutzt die Argusaugen der Maschine, um Inhalte aus digitalen Bildern zu erschließen. Wie der Riese des Mythos kann auch der Computer rein quantitativ mehr sehen als der Mensch. Hunderttausende Bilder können schnell nach wiederkehrenden Mustern durchsucht werden. Gleich dem mythologischen Wächter kann die Maschine eine Objektkategorie, beispielsweise eine Kuh oder, noch konkreter, die Kuh Io, im Blick behalten. Während das Urteilsvermögen des Kunsthistorikers den Einflüsterungen von Kustoden oder Sammlern erliegen oder auch zu ungerechtfertigtem Widerspruch gereizt werden könnte, hält sich die Maschine stur an die einmal aufgestellten Regeln, ebenso wie Argus sich nicht einmal von Ios Vater erweichen ließ.

Die letzte Größe, mit der dieser Vergleich strapaziert werden soll, ist das Monströse des Argus und der riesige Datenmengen ‚sichtenden' Maschine. Ressentiments gegen die Technologie sind nicht selten. Allerdings scheint es ebenso unangemessen, dass sich Kunsthistoriker von einem automatischen Such- und Analyseinstrument in ihrer Kompetenz bedroht fühlen, wie es unpassend ist, dass Artikel mit der Abschaffung von Kunsthistorikern durch *Computer Vision* polemisieren (vgl. Matthew Sparkes, Could Computers put art historians out of work?, in: *The Telegraph* vom 18.8.2014). Neben solchen Befürchtungen reagieren Fachkollegen bisweilen ungläubig und bezeichnen den Einsatz von Algorithmen zur Bildanalyse als utopisch. Das ist erstaunlich in einem Fach, das in Bezug auf seine Textquellen deutlich von einer anderen *Computer Vision*-Technologie, nämlich der Volltexterkennung mittels OCR-Programmen, profitiert hat.

ZWEI BILDWISSENSCHAFTEN

Der folgende Überblick über das Forschungsgebiet möchte die transdisziplinären Methodenansätze von *Computer Vision* und Kunstgeschichte präsentieren sowie erste Prototypen vorstellen. Dabei sollen die Möglichkeiten des Erkennens von Szenen und Objekten in Bildern auf einer visuellen und semantischen Ebene diskutiert werden. Die sich überschneidenden Problemstellungen von Informatik und Kunstgeschichte sind die visuelle Suche nach Bildinhalten, der Bildvergleich und das Verstehen von Szenen und Objekten im Bild. Der Computer wird dabei als keine den Experten ersetzende, sondern als den Experten ergänzende Maschine betrachtet, in der die Chance einer schnellen Bildverarbeitung in Bezug auf große Datenmengen genutzt wird. Das Problem der Ähnlichkeit kann dadurch strukturierter angegangen und besser validiert werden.

Computer Vision arbeitet mehrheitlich mit Fotografien und Videos aus unserer gegenwärtigen Alltagswelt. Die Repositorien der Kunstgeschichte bieten der *Computer Vision* dagegen genuine Herausforderungen durch wechselnde Stile, zeichnerische Variation, die Darstellung unterschiedlicher Realitätsgrade, Abstraktionen und andere

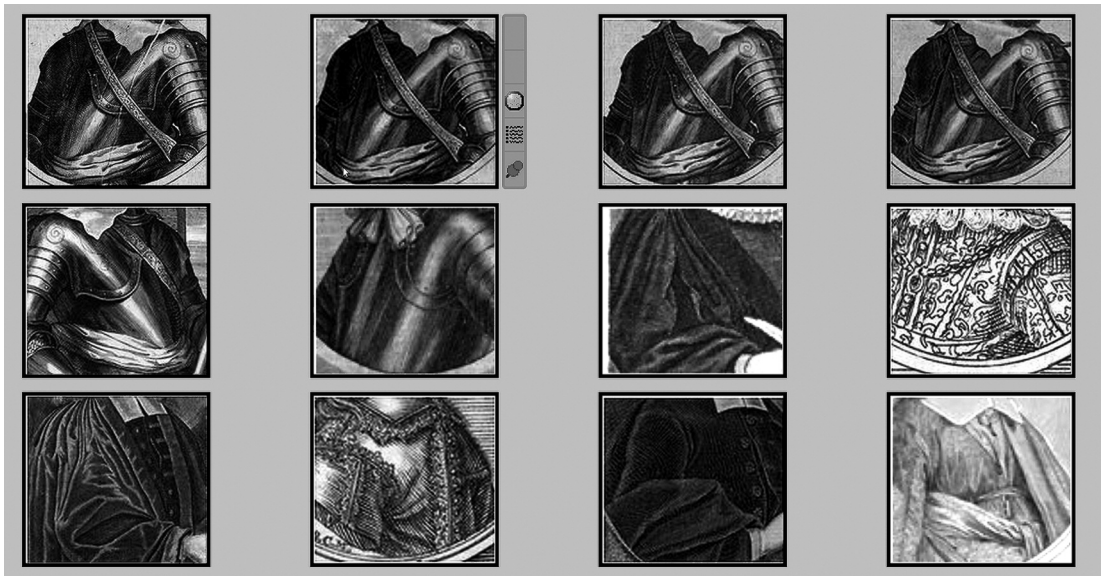

Abb. 1 Suchergebnisse aus dem Marburger Porträtindex im Prototyp der Computer Vision Group Heidelberg. Suchauswahl oben links (Sämtliche Abbildungen sind Screenshots oder Grafiken der Computer Vision Group Heidelberg)

Bildkonventionen als in der Fotografie. Insgesamt scheint der Sonderstatus von nicht-abstrakter Kunst gegenüber der Fotografie jedoch weniger groß zu sein als erwartet. Es lassen sich mit anhand von Fotos trainierten Algorithmen durchaus vergleichbare Suchergebnisse bei kunsthistorischen Datensätzen generieren. Einige Bildcorpora sind sogar standardisierter und einfacher automatisch bearbeitbar als Fotografien. In der Kunst wurden über Jahrhunderte Sehaufgaben in skalierender Komplexität gelöst. Diese Darstellungen lassen sich dadurch nicht nur als hochdifferenzierte Trainingsdaten heranziehen, sondern enthalten zudem Informationen über menschliche Sehstrategien, mit denen das automatische Sehen unterstützt werden kann. Künstler aller Epochen abstrahieren und pointieren die charakteristischen Eigenschaften von Objekten durch prägnante Formen. Wenn also beispielsweise mittelalterliche Miniaturen auf einfache Lesbarkeit hin angelegt wurden, sind an den Objekten Partien betont, die für deren Identifikation besonders signifikant sind. Der Computer „lernt" somit an den in der Kunst repräsentierten Objekten menschliches Differenzierungsvermögen. Dazu extrahiert der Algorithmus zuerst eine Vielzahl von Merkmalen aus Bildern, um diese anschließend zusammenzufassen. Mittels statistischer Mustererkennung wird dann eine kompakte Bildbeschreibung erlernt. Dabei werden die Charakteristika automatisch gefunden, die ein oder mehrere positive Objektbeispiele von einer großen Menge irrelevanter Hintergrundregionen unterscheiden. Nach diesem Training ist der Algorithmus in der Lage, ähnliche Objekte in umfangreichen Bildsätzen selbsttätig zu finden. Die Empirie künstlerischer Erfahrung kommt der Informatik in dieser Zusammenarbeit ebenso zugute wie kunsthistorische Methodenansätze zum Bildverstehen. So können beispielsweise Panofskys dreistufiges Interpretationsschema und eine bildwissenschaftliche Rezeptionsästhetik die informatischen Trainingsbilder und kunsthistorischen Datensätze kritisch beschreiben und die jeweilige Betrachterrolle, die dem Computer antrainiert wird, identifizieren.

Die theoretische Auseinandersetzung mit der *Computer Vision* und die Entwicklung von entsprechenden Anwendungen sind für die Kunstgeschichte zur Erschließung der derzeit laufenden, umfangreichen Digitalisierungsmaßnahmen dringend notwendig. In den Kunst- und Bildwissenschaften verlief die technische Entwicklung – nicht zuletzt wegen der benötigten Rechnerkapazitäten bei Bildern – langsamer, heterogener und lokaler als in den textbezogenen Disziplinen. Das Internet und wissenschaftliche Datenbanken waren in den ersten Dekaden ihrer Entwicklung vornehmlich textbasiert. Durch die im ersten Schritt einfacher erscheinende automatische Erschließung von Texten, durch OCR und Volltextsuche,

blieb eine vergleichbare Aufbereitung von Bildern bislang ein Desiderat. Dabei wird verkannt, dass auch die Computerlinguistik bei einer tieferen Erschließung der Texte auf für künstliche Intelligenz nur schwer lösbare Probleme stößt. Die Zurückhaltung gegenüber rein visuellen Ansätzen scheint auch methodische Gründe zu haben: In der Wahl ihrer Analyseverfahren ist die Kunstgeschichte in vielen Fällen keine Bildwissenschaft, als die sie oft postuliert wird, sondern sieht einen Zugang erst in und nach der Übersetzung der visuellen Untersuchungsgegenstände in Sprache. Der automatisierte visuelle Zugriff könnte jedoch komplementär zur Abstraktionsleistung des Textes kunsthistorische Inhalte erschließen und aufbereiten.

BILDWISSENSCHAFT REDUZIERT AUF METADATEN

Viele der großen Bilddatenbanken können ihre Digitalisate kaum kursorisch in Schlagwörtern charakterisieren oder Bezüge herstellen. Dies liegt nicht nur an der großen Menge aller im Netz vorhandenen Reproduktionen, sondern auch an der Schwierigkeit, Bildinhalte überhaupt in wissenschaftliche Textinformationen zu übersetzen. Die Annotation kunsthistorischer Datenbanken erschöpft sich oft in Informationen zu Künstler, Titel, Datierung, Standort, kurzen Angaben zur Ikonografie und gelegentlichen realienkundlichen Stichworten. Einzelformen werden hingegen kaum beschrieben; Komposition, Interaktion und Struktur der Szenen können nicht verglichen werden. Es muss an dieser Stelle nicht ausgeführt werden, wie sehr diese Ordnungssysteme Zugänge und Methoden der Kunstgeschichte mitbestimmen und wie wenig sich darin Rezeptionsverhalten und Bildbefunde widerspiegeln. Taxonomien wie ICONCLASS stellen engmaschige Beschreibungswerkzeuge bereit, durch die eine textliche Ordnung der Bilder möglich wird. Gleichzeitig ergeben sich Überschneidungen mit semiotisch und strukturanalytisch geprägten Methodenansätzen der Kunstwissenschaft, die das Bild als zu dechiffrierenden Text verstehen.

Allerdings beschränken sich die genannten Taxonomien auf einzelne, wenn auch zentrale Erschließungsmodi wie etwa die Ikonografie und sind oft eurozentrisch ausgerichtet. Wer nach Individuen oder nach einem bestimmten Gegenstand sucht oder Aufschlüsse über die Rezeption gewinnen möchte, wird damit nur selten ans Ziel kommen. Dies betrifft nicht nur den Bildinhalt, sondern z. B. auch die Zuschreibung an einen Künstler. Denn wenn ein Bild zu Anfang des 20. Jahrhunderts unter einem aus heutiger Sicht falschen Künstlernamen auktioniert wurde, ist es trotz digitalisierter Kataloge schwer auffindbar. Ansätze der *Computer Vision*, die es ermöglichen, Bilder automatisch inhaltsbasiert zu verarbeiten, statt sie nur auf der Ebene von Metainformationen zu erfassen, werden vor dem Hintergrund der enormen Datenmengen zu einer essentiellen Methode der bildhistorischen Forschung und letztlich auch für öffentliche Sammlungen relevant. Die Zusammenarbeit von Kunstgeschichte und *Computer Vision* kann Such- und Analysestrategien erarbeiten, die unmittelbar bei der visuellen Information des Bildes ansetzen.

In der Kunstgeschichte sind die verschiedenen Teilgebiete der Bildverarbeitung wenig bekannt und werden oft nur mit Firmen wie Adobe oder Google in Verbindung gebracht. Zu unterscheiden ist grundsätzlich zwischen *Image Processing*-Ansätzen (*Low-Level Vision*), bei denen beispielsweise über die Textur von Pinselstrichen die Autorschaft von Kunstwerken bestimmt werden soll (Johnson/Wang 2008), und Problemstellungen im Zusammenhang mit dem semantischen Bildverstehen, bei dem Objekte und Szenen erkannt werden (*High-Level Vision*). Um letzteres soll es im Folgenden gehen.

GRUNDLAGEN UND FRÜHE BILDSUCHEN

Die frühen aussichtsreichen und breit rezipierten Versuche einer automatischen Bildsuche (Vaughan 1997) demonstrierten bereits die große Herausforderung, die eine kunsthistorische Bildanalyse für die *Computer Vision* darstellt; sie hatten daher wenige Nachfolger. Ein hoher Grad an Standardisierung innerhalb des Bilddatensatzes er-

Abb. 2 Suchergebnisse aus Sachsenspiegel-Handschriften vom Prototyp der Computer Vision Group Heidelberg. Suchauswahl oben links, mit Kreuz markierte Ausreißer

leichtert allerdings das Erkennen von Objekten maßgeblich. So war ein aus 2800 Zeichnungen bestehendes florentinisches Wappencorpus mit seinen signifikanten heraldischen Objekten ein guter Testfall für die Kooperation zwischen dem Kunsthistorischen Institut in Florenz (MPI) und der Bildverarbeitung des Istituto di Scienza e Tecnologia dell'Informazione (ISTI) in Pisa (http://wappen.khi.fi.it/). Die Bayerische Staatsbibliothek implementierte eine Bildersuche, durch die ähnliche Bilder oder vorgegebene Bildsegmente innerhalb des eigenen Corpus gefunden und die Suche über vom Nutzer hochgeladene Vorlagen ermöglicht wird (http://bildsuche.digitale-sammlungen.de/). Die *Visual Geometry Group* der University of Oxford präsentiert neben weiteren Fallstudien im Kunstbereich eine Demoversion zum Durchsuchen von illustrierten Lieddrucken der Bodleian Library (http://zeus.robots.ox.ac.uk/ballads/), ein Ansatz, der aufgrund der zahlreichen Nachdrucke und Varianten der Illustrationen in diesem Corpus sinnvoll erscheint.

Beide Bildsuchen – wie auch die kommerziellen Anbieter (z. B. Google, TinEye) – werden der Komplexität der kunsthistorischen Fragestellungen nicht gerecht (vgl. Felix Thürlemann, Christus eingegeben und Hitler gefunden beim Ikonogoogeln, in: *Frankfurter Allgemeine Zeitung* vom 14.9.2011), unter anderem, weil sie nur identische Bilder oder zufällig korrespondierende Formen auffinden und außerdem nicht den gegenwärtigen Forschungsstand der *Computer Vision* widerspiegeln (Everingham/Zisserman 2014). Dennoch geben diese früheren Entwicklungen eine Vorstellung davon, was mit Hilfe von automatischem Sehen gewonnen werden kann: eine Ordnung der Bilder nach visueller Ähnlichkeit. Das automatische Sehen ist dadurch nicht nur als Bildsuche al-

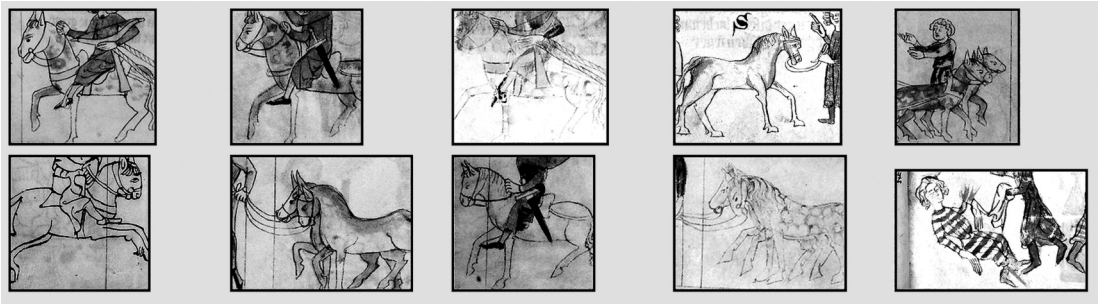

Abb. 3 Suchergebnisse aus Sachsenspiegel-Handschriften im Prototyp der Computer Vision Group Heidelberg. Kombinierte Suche nach Pferdekopf und -hufen

ternativ zum Text zu verstehen. Der Datensatz kann von vorneherein nach visuellen Gesichtspunkten aufbereitet und dargestellt werden, indem beispielsweise eine Anordnung visuell ähnlicher Bilder statt einer alphabetischen Sortierung erscheint. Die Ähnlichkeit kann sich dabei entweder auf das ganze Bild beziehen oder auf mit Hilfe von Auswahlboxen bestimmte Bereiche beschränkt werden.

Für jede *Computer Vision*-Anwendung muss vorweg entschieden werden, welchen Anteil das maschinelle Lernen haben soll. So ließe sich für einen Datensatz, in welchem eine Kategorie von Objekten eine besondere Bedeutung besitzt, diese durch Beispiele erlernen. In der stark standardisierten Zeichensprache mittelalterlicher Handschriften lassen sich so etwa die piktogrammartigen Formen von Kronen oder Gesten an einer Auswahl trainieren und dann über den ganzen Datensatz abfragen (Bell/Ommer/Schlecht 2013). Die Bildsuchen von Oxford und München verwenden hingegen sogenannte *Bag-of-words*-Ansätze, in denen Bildelemente wie Wörter in einem Glossar geordnet werden. Diese Registrierung der Daten führt zu einer schnellen Suche, bedarf aber einer aufwendigen Vorbereitung des Datensatzes und führt dazu, dass in der Suche nur Dinge gefunden werden können, die zuvor registriert wurden. Darüber hinaus gibt es keine räumliche Zuweisung, so dass beispielsweise bei der Suche nach einer Person es für das Programm irrelevant ist, ob die Person ihren Kopf auf dem Hals oder unter dem Arm trägt. Methodische Ansätze, in denen der Datensatz zuvor nicht bzw. nur minimal durch Trainingsbeispiele oder andere Registrierungsverfahren erschlossen wurde, ergeben entsprechend längere Suchzeiten, aber auch die Möglichkeit einer offenen Bildsuche.

SUCHE NACH ÄHNLICHKEITEN

Der hier vorzustellende Suchalgorithmus soll die Schwächen des *Bag-of-words*-Modells beseitigen und damit für beliebige Suchanfragen nutzbar sein (Takami/Bell/Ommer 2014). Der Prototyp kann seit dem Frühjahr 2015 mit Daten des Prometheus-Bildarchivs ausprobiert werden (http://hci.iwr.uni-heidelberg.de/COMPVIS/projects/suchpassion). Der Nutzer kann über ein einfach zu bedienendes Webinterface Bildpartien aus dem Datensatz oder eigenen Uploads markieren, um die ausgewählten Bereiche im Datensatz zu suchen. Während der Suche können positive Ergebnisse als weitere Lernbeispiele hinzugenommen werden, um die Resultate zu verbessern. Der Algorithmus ermöglicht nicht nur das Auffinden von identischen oder sehr ähnlichen Partien, sondern eben auch von größeren Abweichungen, wobei die Ergebnisse nach dem Grad der Ähnlichkeit sortiert werden *(Abb. 1)*. Oben links erscheint der gesuchte Ausschnitt, es folgen weitere Versionen des gleichen Stichs, eine spiegelverkehrte Variante sowie ein anderer Harnisch und weitere Obergewänder mit ähnlicher Linienführung. Die Nutzer können somit selbst den Suchvorgang nachvollziehen und entscheiden, wann das Ergebnis zu weit von der Sucheingabe entfernt ist.

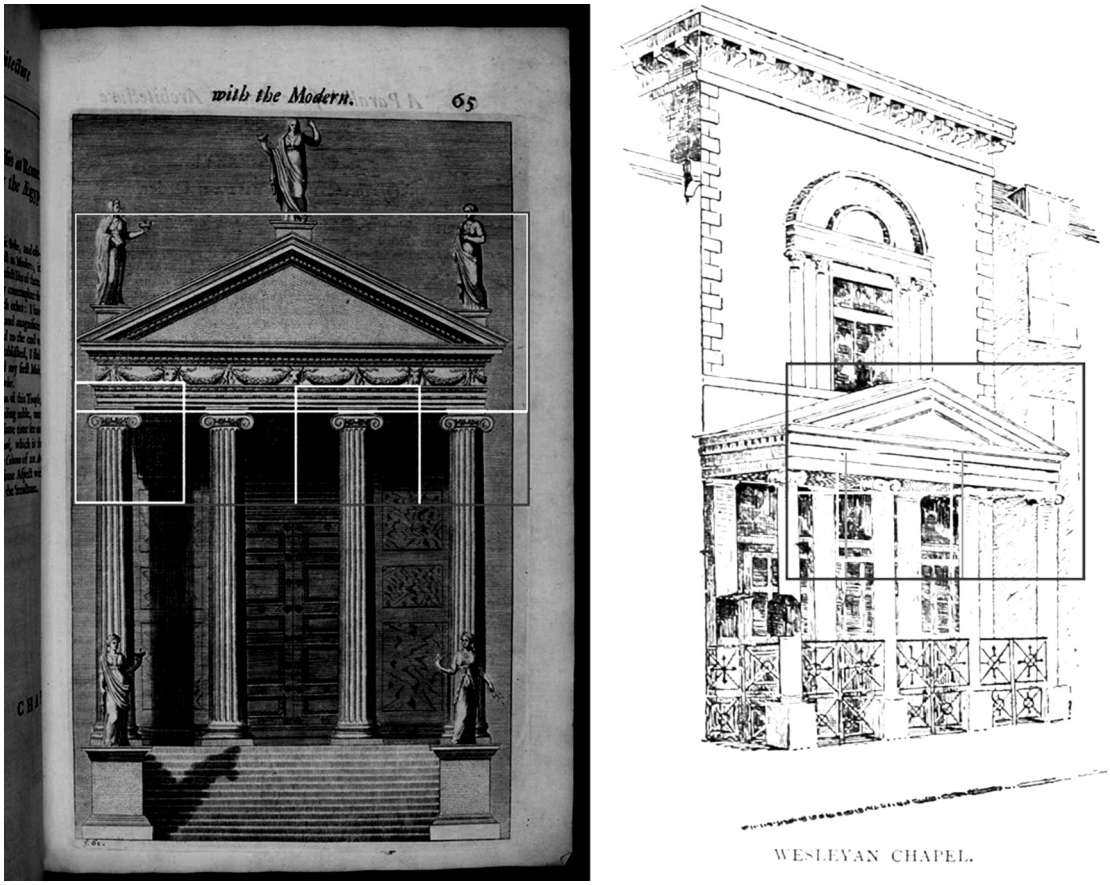

Abb. 4 links: kombinierte Suche nach Dreiecksgiebel und ionischen Kapitellen, rechts: räumlich variierendes Ergebnis

Ohne dass im Programm eine Semantik der Objekte angelegt wäre, ergeben sich in einigen Suchen dennoch semantische Zusammenhänge durch die vom Nutzer festgelegte Auswahl. So entsteht bei der Suche nach einer liegenden Figur in den vier Sachsenspiegelausgaben eine Bildfolge *(Abb. 2)*, in der zunächst weitere alte Männer mit Bart erscheinen, dann Alter, danach Geschlecht wechseln, bis schließlich ähnlich gelagerte Objekte als offensichtliche Fehler den Zusammenhang aufheben. Die Haltung der Figur ist so signifikant, dass sie lange stabil von anderen Objekten unterschieden werden kann, während kleinere Veränderungen in Gesicht und Kleidung der Gestalt nur eine untergeordnete Rolle spielen. Die Suchenden können die semantische Signifikanz ihrer Anfrage erhöhen, indem sie die Auswahlboxen geschickt wählen. Es ist möglich, nur einen besonders charakteristischen Teil eines Objekts für die Suche zu verwenden oder auch mehrere Auswahlfenster einzufügen. So lässt sich ein Bischof etwa über Krummstab und Mitra suchen, Pferde lassen sich hingegen gut durch ihre Köpfe und Beinstellungen von anderen Motiven abgrenzen *(Abb. 3)*.

Im Webinterface kann der Nutzer entscheiden, wie stark die räumliche Distanz der Auswahlboxen berücksichtigt werden soll. Bei einem geringen Wert geht der räumliche Zusammenhang verloren, der durch die Anatomie des Pferdekörpers oder den Ornat des Bischofs grob vorgegeben ist. Gleichzeitig werden dadurch überraschende Funde möglich, zum Beispiel der eines Bischofs, der seine Amtsinsignien abgelegt hat. Besonders erfolgreich ist diese Suchmethode jedoch bei Datensätzen, in denen gleiche Objekte in variierenden Abständen vorkommen. Dies trifft in besonderem Maße auf Architekturdarstellungen zu. Kapitelle, Dreiecksgiebel und Balustraden lassen sich als Suchfenster markieren und können dann in engem oder weiterem räumlichen Zusammenhang zueinander gesucht werden. Durch die Zugabe von weiteren Suchfenstern wie Kapitellen oder

Fensterlaibungen lassen sich auch zunächst gleiche Formen wie Dreiecksgiebel über Fenstern und Dreiecksgiebel als Bekrönungen von Fassaden semantisch differenzieren *(Abb. 4)*.

Durch die rein visuelle Suche ergibt sich jedoch auch die Möglichkeit, sich aus semantischen und ikonografischen Konventionen zu lösen und Ähnlichkeiten jenseits der Ikonografie aufzudecken und zu interpretieren. Ein Großteil der menschlichen Wahrnehmungsleistung an Kunstwerken besteht aus Abstraktionsfähigkeit, Assoziation und Verständnis der künstlerischen Umsetzung. Daran kann sich die Maschine nur langsam herantasten. Hier sind die Qualität der Suchergebnisse und die Laufzeit der Suche abhängig vom Aufwand händisch erstellter Trainingsbeispiele und vom gewählten informatischen Ansatz. Auch die Interaktion mit den Nutzern, das Vorschlagen und Evaluieren von Ergebnissen, kann die Suche beschleunigen und effizienter gestalten. *Computer Vision* kann Vorschlagssysteme entwickeln, durch welche die Erschließung von Bildrepositorien grundsätzlich und individuell befördert wird. Das automatische Sehen schafft nicht nur einen neuen ‚Betrachter', sondern ermöglicht auch, neu über kunsthistorische Methoden zu reflektieren. Dieser maschinelle Argus kann Bildmengen überblicken, die von Menschen nicht gesichtet und verglichen werden können. Ob er dabei so gut instruiert ist, dass er keinen Augentäuschungen erliegt, hängt letztlich an der Mitarbeit und methodischen Reflexion der Kunstwissenschaft.

LITERATUR

Peter Bell/Björn Ommer/Joseph Schlecht: Nonverbal Communication in Medieval Illustrations Revisited by Computer Vision and Art History, in: *Visual Resources: An International Journal of Documentation* (Special Issue: *Digital Art History*) 29/1–2, 2013, 26–37

Mark Everingham/Andrew Zisserman et al.: The Pascal Visual Object Classes Challenge. A Retrospective, in: *International Journal of Computer Vision*, June 2014, 1–38

C. Richard Johnson/James Z. Wang et al.: Image Processing for Artist Identification – Computerized Analysis of Vincent van Gogh's Painting Brushstrokes, in: *IEEE Signal Processing Magazine* 25/4, 2008, 37–48

Hubertus Kohle: *Digitale Bildwissenschaft*, Glückstadt 2013, insbes. 37–57, 70–76

Masato Takami/Peter Bell/Björn Ommer: Offline Learning of Prototypical Negatives for Efficient Online Exemplar SVM, in: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, IEEE, 2014, 377–384

William Vaughan: Computergestützte Bildrecherche und Bildanalyse, in: Hubertus Kohle (Hg.), *Kunstgeschichte digital. Eine Einführung für Praktiker und Studierende*, Berlin 1997, 97–105

DR. DES. PETER BELL, PROF. DR. BJÖRN OMMER
Computer Vision Group, Heidelberg Collaboratory for Image Processing (HCI),
Ruprecht-Karls-Universität Heidelberg,
Speyerer Str. 6, 69115 Heidelberg,
bell@uni-heidelberg.de,
ommer@uni-heidelberg.de