MARSILIUS KOLLEG

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Lasst die Bilder sprechen!

Bild-Text-Beziehungen innerhalb einer informatischen, linguistischen und kunsthistorischen Praxis

Björn Ommer

Auszug aus dem Jahresbericht
„Marsilius-Kolleg 2015/2016“

Björn Ommer

Lasst die Bilder sprechen!

Bild-Text-Beziehungen innerhalb einer informatischen, linguistischen und kunsthistorischen Praxis

Ein Bild sagt mehr als tausend Worte, jedoch sprechen die wenigsten Bilder tatsächlich. Aber was würden sie sagen, wenn sie es könnten? Was würden sie uns mit nur hundert Wörtern mitteilen, was mit nur 10? Wie würden sie einzelne Bildregionen mit nur einem Substantiv oder einem Verb beschreiben und welche Regionen wären dies? Denn auch wenn ein Bild stark mit Bedeutung aufgeladen ist – und diese Annahme ist grundlegend für die kunsthistorische Arbeit –, so verbalisieren doch die wenigsten Bilder diese ihnen immanente Information. Selbst wenn die Aussage eines Bildes für den menschlichen Betrachter offensichtlich ist, so stellt ihre Verbalisierung doch eine Herausforderung dar. Dies wird insbesondere deutlich, wenn wir uns fragen, wie genau sich die Bedeutung aus dem Bild ableiten lässt und welche Regionen wie zur Aussage eines Bildes beitragen. Wie gelangen wir von einzelnen Pinseltupfern eines Gemäldes, von einzelnen Strichen einer Zeichnung oder einzelnen Bildpunkten eines digitalen Bildes zu Objekten oder auch nur ihren Bestandteilen? Wie erschließt sich die gesamte Szene oder gar die tieferliegende Bedeutung?

Diese Herausforderung wird umso deutlicher, wenn wir einer Maschine beim Sehen zuschauen. So ermöglicht die Disziplin der Computer Vision, die Algorithmen entwickelt, mit denen Computer aus Bildern das Sehen lernen sollen, einen anderen Blickwinkel. Für einen Computer beginnt der Prozess des Bildverstehens mit einzelnen Bildpunkten. Aber was sagt ein sogenannter Pixel aus? Ein Bildpunkt ist schließlich nur eine lokale Messung von Helligkeit oder Farbe in einem kleinen Bereich einer Szene. Die Farbe mag uns vielleicht Anhaltspunkte zur Identität von Objekten geben, aber eine eindeutige Identifikation findet dadurch nicht statt. Nehmen wir

zum Beispiel die Farbe Blau: Sie kann sowohl auf Wasser als auch auf einen Himmel an dieser Stelle im Bild hinweisen, aber auch schlichtweg Teil eines Gegenstandes sein. Außerdem kann eine Änderung der Beleuchtung wiederum die wahrgenommene Farbe einer Oberfläche verändern. Die Betrachtung einzelner Bildpunkte führt folglich nur zu Anhaltspunkten. Sehen ist offensichtlich mehr als ein bloßes Betrachten einzelner Bildpunkte. Die Form eines Objekts, die Mimik eines Gesichts, die Stimmung eines Bildes – was wir sehen sind typischerweise komplexe, weil emergente Eigenschaften, die sich erst durch das Zusammenspiel vieler Bildbereiche ergeben. Der Zusammenschluss einzelner Bereiche bildet dabei bedeutungstragende Kompositionen, die wiederum durch Gruppierung mit weiteren Regionen eine ganze Hierarchie ausbilden, die den Bildinhalt zusammenfasst. Und wenn auch dies nicht mehr ausreicht, so wird der Einbezug von Kontext in Form von verwandten Bildern und erklärenden Texten nötig. Dass dieser gesamte Prozess hochgradig nichtlinear verläuft, dass also das Ganze nicht ein bloßes Aufsummieren seiner Teile ist, stellte schon Aristoteles fest und es wurde zu einem der Pfeiler der Gestaltpsychologie vor etwa hundert Jahren.

Die Auseinandersetzung mit Bildern stellt demnach unweigerlich den Terminus der Bedeutung in den Fokus und fragt nicht zuletzt nach der Wesensart und Konstitution des Begriffes. Indem sich Kunsthistoriker oder Linguisten mit dem Begriff der Bedeutung auseinandersetzen, verlassen sie den Bereich ihres Faches, der sich ausschließlich mit formalen Qualitäten eines Kunstwerkes oder Textes befasst und wenden sich der Ikonographie, beziehungsweise der Semantik oder gar Pragmatik zu. Letztere beschäftigt sich außerdem mit kontextabhängiger und nicht wörtlich zu nehmender Bedeutung, wie sie zwischen sprachlichen Zeichen und dem Benutzer entsteht. Wo entsteht Bedeutung schließlich? In der Kunstgeschichte finden sich dafür zwei wesentliche Orte. Zunächst innerhalb des Kunstwerkes selbst, in welchem durch die Abbildung bestimmter Motive und Symbole oder durch ganze Szenen Bedeutung generiert wird. Die Erfassung des Inhalts erfolgt demnach, indem einzelne Bildteile, die gesamte Szene oder sogar das Kunstwerk selbst – inklusive Pinselführung, Farbauftrag oder verwendetes Material – analysiert werden. Zudem entsteht eine weitere Bedeutungsebene durch die Verbindung und Öffnung des Werkes nach außen hin durch den Betrachter. Nun entsteht Bedeutung im Zusammenspiel und mit dem Wissen um den jeweiligen Kontext. So kann das Kunstwerk auch eine politische und gesellschaftliche Relevanz aufweisen. In der Linguistik lässt sich eine ähnliche Bedeutungsentstehung beobachten, die ebenfalls das Aggregieren einzelner

Zeichen zu komplexeren bedeutungstragenden Kompositionen voraussetzt: Entweder ist sie in einzelnen Wortkörpern oder in ganzen Abschnitten zu finden. Schließlich entsteht Bedeutung auch durch die Verschränkung von Linguistik und Kunstgeschichte oder Bild und Text. Anhand der Ausführungen lassen sich also zwei Arten der Bedeutungsgenese festhalten: einmal die objekt-immanente und dann die, die in der Verbindung zweier Elemente entsteht – seien es Bild und Bild, Text und Text oder Bild und Text.

Es stellt sich also nicht nur die Frage, wie das Bildverstehen algorithmisch durch die Computer Vision abgebildet werden kann. Es ist auch offen, inwieweit ein gemeinsames Bild- und Textverstehen einander unterstützen können. Denn gerade die Computerlinguistik steht vor ähnlichen Herausforderungen, wenn beispielsweise Texte automatisch von einer Sprache in eine andere übersetzt werden sollen. So kommt es häufig bei der Übersetzung des Inhalts zu Missverständnissen, die vor allem aufgrund von Ambiguitäten oder fehlendem Hintergrundwissen entstehen. In Text-Bild Systemen – Ordnungen, die eine enge Beziehung zwischen Text und Bild implizieren – können Text und Bild die beim jeweils anderen auftretenden Mehrdeutigkeiten viel eher auflösen. So unterstützen Objektbegriffe im Text das Gruppieren von Bildregion zu bedeutungstragenden Kompositionen. Umgekehrt profitiert das maschinelle Übersetzen des Texts von der Detektion von relevanten Objekten im Bild. Für die Untersuchung einer Verbindung von Computerlinguistik und Computer Vision stellt die Kunstgeschichte ein wichtiges Bindeglied dar, da sie gerade die inhaltliche Analyse der Bilder zum Ziel hat. Auch sind gerade in dieser Disziplin durch große Digitalisierungskampagnen große Mengen an Bilddaten und textueller Beschreibung vorhanden, die durch ihre enge Text-Bild Beziehung eine ideale Basis für die Verschränkung der Fächer bieten. Ein wesentliches Ziel ist dabei die Auseinandersetzung mit Mehrdeutigkeiten, wie sie beim Text- und Bildverstehen auftreten.

Für die Erkennung und Beschreibung eines Objekts bedeutet dies beispielsweise, dass wir die semantische Lücke zwischen sichtbarem Bild und unsichtbarem Inhalt schließen müssen. Es gilt, Bildpunkte so zu Kompositionen zusammenzufassen und

zu abstrahieren, dass die resultierende Repräsentation unwichtige Details vernachlässigt und wesentliche Charakteristika herausstellt. Alle Instanzen einer Kategorie von Objekten sollen so einander ähnlicher werden, während sie sich von anderen Kategorien absetzen sollen. Anhand einer Menge Beispielbilder lernt die Maschine dabei implizit ein bildbasiertes Maß für die Ähnlichkeit von Objekten. Das Bestreben ist insbesondere, dass sich dieses Maß gut auf neue, zuvor nicht gesehene Instanzen einer Kategorie verallgemeinern lässt. Ein robustes Maß für die Ähnlichkeit von Objekten oder ihrer Bestandteile ist essentiell für das Bildverstehen. Es ermöglicht, Korrespondenzen zwischen ähnlichen Objektteilen herzustellen und sie zueinander und zu anderen Objekten in Beziehung zu setzen. Weiterhin ist dies die Grundlage für eine automatische Annotation. Korrespondierende Bildbereiche in verschiedenen Bildern können einer Objektkategorie zugeordnet werden, und die räumliche Anordnung und Deformation eines Objekts kann anhand der relativen Positionierung seiner Bestandteile nachvollzogen werden. Schließlich können so Begriffe, die im die Bilder ergänzenden Text vorkommen, im Bild lokalisiert werden. Umgekehrt können verwandte Objektinstanzen aus verschiedenen Bildern auf Text abgebildet werden, so dass Mehrdeutigkeiten und Synonyme durch die Bilder aufgelöst werden können.

Die Zusammenarbeit von Text und Bildern in der kunsthistorischen Praxis

Eine Zusammenarbeit von Text und Bildern zur Übersetzungshilfe von zweisprachigen Texten, zur Klärung von Bedeutung oder zur Auflösung von Ambiguitäten hat in der Kunstpraxis bereits mehrfach Anwendung gefunden. Als Materialpool für eine Erprobung des maschinellen Übersetzens eignen sich besonders Werke, die bereits aufgrund ihrer Definition eine enge Beziehung von Text und Bild implizieren oder einen schablonenhaften Aufbau aufweisen. Paradigmatisch sind hierbei illustrierte Bücher aus den ersten Jahrzehnten des Buchdrucks zu nennen, da die neuen technischen Produktionsvoraussetzungen dieser medialen Umbruchsphase u.a. eine Tendenz zur Standardisierung und Formelhaftigkeit zur Folge hatten. Ein Beispiel hierfür sind Inkunabeln, die zu großen Teilen Mitte/Ende des 15. Jahrhunderts entstanden sind und mit beweglichen Lettern gedruckt wurden. Dabei entstanden Wiederholungen und große Ähnlichkeiten zwischen den einzelnen Drucken.

Zwei Inkunabeln scheinen für den oben beschriebenen Ansatz besonders geeignet. Sie stammen aus dem Jahr 1488, beschreiben Heiligenleben und haben – im Stil der Inkunabeln – einen fast identischen Aufbau. Entscheidend ist, dass sie die

Geschichten der Heiligen nicht nur im Text erzählen, sondern auch durch Bilder visualisieren und dabei wiederholt auf gleiche Bilder zurückgreifen. Auch Blockbücher fallen in diese Kategorie, obwohl sie im Unterschied zu Inkunabeln nicht mit beweglichen Lettern gedruckt, sondern im Holzschnittverfahren, aus einem Stück, gefertigt wurden. Obwohl damit zwar eine wiederholte Nutzung der Druckstöcke nicht ausgeschlossen wird, verhindert die Geschlossenheit der Platten einen variablen und modulartigen Einsatz. Trotzdem handelt es sich bei beiden Varianten um illustrierte Bücher, sodass sie für den verfolgten Ansatz von Interesse sind. Auch moderne Formen wie Comics oder sogenannte *Scrapbooks* – Alben, in denen Text und Bilder gleichermaßen ausschnitthaft und anekdotenhaft von persönlichen Ereignissen berichten – greifen auf ein ähnlich enges Verhältnis von Text und Bild zurück. Die Schablonenhaftigkeit dieser „Gattungen" und Techniken erleichtert das Herausarbeiten von grundlegenden Mustern und visuellen Ähnlichkeiten, die folglich eine Gegenüberstellung von komplexen bildlichen Darstellungen und sprachlichen Phrasen erlaubt.

Weitergehende Herausforderungen

Neben eindeutigen Vorteilen bringen die angenommenen engen Text-Bild Systeme auch Beschränkungen mit sich. Objekte und ihre Beziehungen lassen sich so gut erschließen, mit eindeutigen Vorteilen für die beteiligten Disziplinen. Der Zugang zu tiefer liegenden Bedeutungsebenen ist jedoch weiterhin eine große Herausforderung. Außerdem finden sich zahllose Beispiele, in denen mit der Interpretierbarkeit gespielt wird. Gerade die Werke der Surrealisten lehnen häufig eine direkte Bedeutungszuschreibung ab. Hier sei nur an das populäre Wort-Bild *La trahison des images* (1929) des Belgiers René Magritte erinnert, das eine hölzerne Pfeife zeigt aber laut Künstler keine ist. Es macht die Grenzen der Interpretierbarkeit von Text und Bild und ihrer Zusammenhänge deutlich. Wer hier versucht, die Ambiguität des Bildes aufzulösen, vernachlässigt dessen essentielle Kritik an der Gleichsetzung vom Objekt, seiner Bezeichnung und der tatsächlichen Repräsentation. Eine andere Herausforderung besteht in der Erweiterung des Ansatzes auf Text-Bild Kombinationen, die nicht in direkter räumlicher Nähe auftreten. Die richtige Zuordnung gelingt hier nur mit ausreichend groß bemessenen Trainingsdatenmengen, um auch schwache Beziehungen aufdecken zu können. Es existieren also direkte Anknüpfungspunkte zu großen Bilddatenbankprojekten und den textuellen Metadaten, die sie zu Künstlern, Kunstwerken und deren Zusammenhängen sammeln.