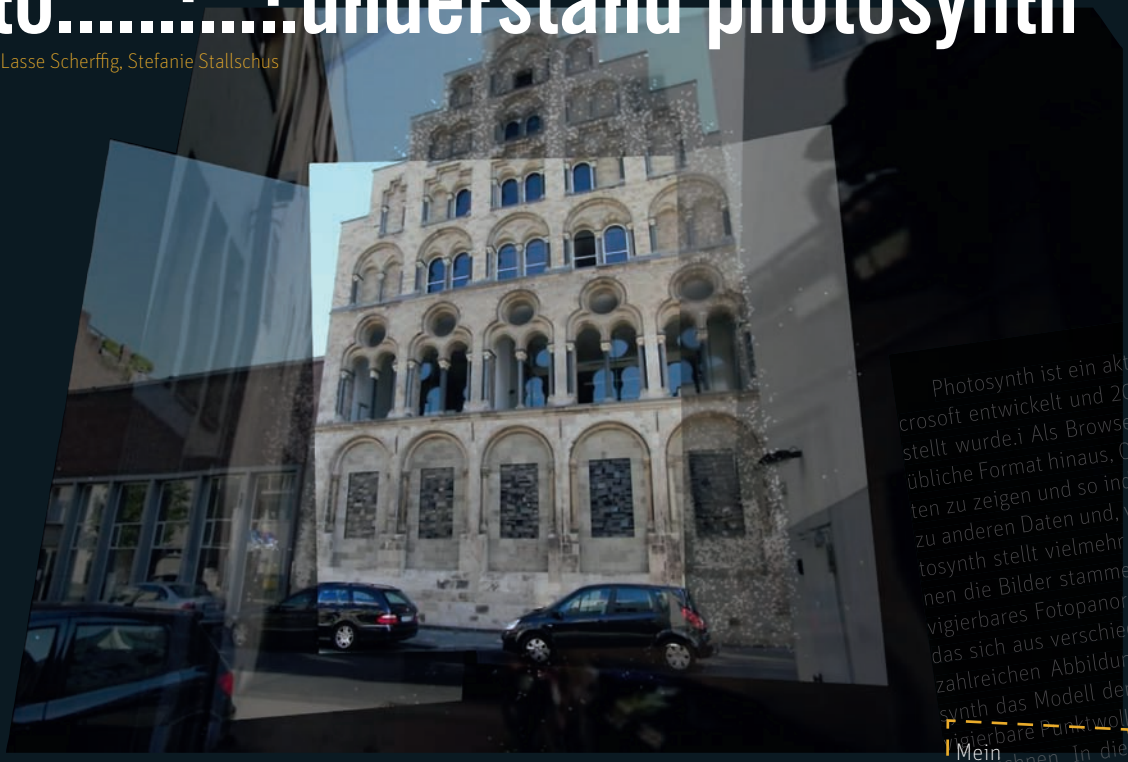
how to understand photosynth

Urs Fries, Lasse Scherffig, Stefanie Stallschus



Mein
Spießschliffdickes
Horn schnitt ich
von der Stirn weg.

Photosynth ist ein aktueller Browser für Bilder, der von Microsoft entwickelt und 2007 von Blaise Aguera y Arcas vorgestellt wurde.[1] Als Browser geht die Anwendung aber über das übliche Format hinaus, Ordnungsrelationen als Listen oder Karten zu zeigen und so individuelle Daten in den Zusammenhang zu anderen Daten und, vor allem, Suchanfragen zu stellen. *Photosynth* stellt vielmehr Bildmengen als die Räume dar, aus denen die Bilder stammen. Was man als User_in sieht, ist ein navigierbares Fotopanorama, in das man hineinzoomen kann und das sich aus verschiedenen Blickwinkeln betrachten lässt. Aus zahlreichen Abbildungen einer Szenerie rekonstruiert *Photosynth* das Modell der Szenerie selber, um es daraufhin als navigierbare Punktwolke, Karte oder dreidimensionale Umgebung zu zeichnen. In diesen Darstellungsformen bleiben dabei die einzelnen Bilder, die Ausgangspunkt der Rekonstruktion sind, weiterhin als Bilder enthalten. Einzelbilder spannen so einen Raum auf, der immer wieder in diese Einzelbilder zerfällt.

Den Hintergrund der Anwendung bildet die Verbindung zweier Strömungen der Informatikforschung: Algorithmische Bildverarbeitung und *Human Computing* (oder *Crowd Computing*). Erstere baut auf dem Ansatz auf, Bilder ihrem Inhalt nach zu analysieren, wobei Inhalt als der Datensatz definiert wird, der einem digitalen Bild zu Grunde liegt.[2] Auf diesen Daten arbeiten inhaltsbasierte Algorithmen um Bilder in Beziehung zueinander setzen zu können. *Human Computing* setzt dagegen nicht direkt auf die Datensätze, die etwa ein Bild ausmachen, sondern auf ihren Gebrauch. Es kommt zum Einsatz, wenn beispielsweise ein Bild nicht nach dem, was es zeigt, in eine Ordnungsrelation überführt wird, sondern nach dem, wie es behandelt wurde. Die Daten, auf denen die Algorithmen des *Human Computing* operieren, sind damit Tags, Verknüfungen und Bewertungen, die von menschlichen Benutzer_innen stammen.[3] Anders als inhaltsbasierte Methoden ist *Human Computing*, da es auf eine große Menge von Verhaltensdaten angewiesen ist, eine direkte Folge der massenhaften Vernetzung von Menschen im Internet und vor allem im *Web 2.0*. Die Ordnungsbeziehung, die *Photosynth* auf der Grundlage dieser beiden Ansätze zwischen Bildern herstellt, ist eine räumliche.

Wahrnehmung von Raum ist nicht unbedingt eine Folge stereoskopischen Sehens mit zwei Augen. Bereits um 1900 hat der Mathematiker Henri Poincaré argumentiert, dass nicht stereoskopisches Sehen, sondern Bewegung eine notwendige und hinreichende Voraussetzung für die Wahrnehmung von Raum ist. Räumliche Wahrnehmung entsteht vor allem dadurch, dass Beobachter_innen ihre eigene Bewegung und die Veränderung ihrer Wahrnehmung, die aus dieser Bewegung folgt, in Beziehung zueinander setzen können: Damit hängt »die Konstruktion von Wahrnehmung vom Prozeß der Veränderung der eigenen Sinneswahrnehmungen durch die Bewegung des Körpers und

von der Korrelierung dieser Veränderung der Sinneswahrnehmung mit den willkürlichen Bewegungen« ab.[4] Sowohl die menschliche Fähigkeit, räumliche Wahrnehmung aus den Bildern auf der Retina zu erzeugen, als auch die rechnerische Fähigkeit Rauminformationen aus den Bilddaten eines Bilderstroms zu extrahieren, bezeichnet die Kognitionswissenschaft als »Structure from Motion« oder »Kinetic Depth.«[5] Und während für die menschliche Wahrnehmung das Verhältnis eigener Bewegung zur Veränderung der Sinneswahrnehmung Basis dieser Fähigkeit ist, müssen für die Informatik die Bilder des Stroms keineswegs aus einer zusammenhängenden Bewegung stammen. Sie müssen lediglich unterschiedliche Perspektiven einer einzigen Szenerie festhalten.[6] Ist diese Voraussetzung erfüllt, können in jedem einzelnen Bild herausragende »Keypoints« gesucht werden. Wie bei der algorithmischen Bestimmung auffälliger »salienter« Bildregionen in der digitalen Bildanalyse sind diese Keypoints zusammenhängende Teile eines Bildes, die sich strukturell von ihrer Umgebung unterscheiden.[7] Auf Grundlage dieser Keypoints können Bilder nun miteinander verglichen werden: Lässt sich ein und die selbe herausragende Bildregion auf zwei Bildern finden, bedeutet diese strukturelle Übereinstimmung auf Ebene des Bildträgers vermutlich, dass beide Bilder an dieser Stelle auch auf der Ebene des Bildobjektes das gleiche zeigen.[8] Der *Photosynth*-Algorithmus versucht auf diese Weise Paare von Bildern zu finden, die das gleiche Objekt abbilden und die sich daher überlagern. Auf Grund dieser Überlagerungen von Bildern lässt sich nun ein »image connectivity graph« konstruieren, der alle zur Verfügung stehenden Bilder an Hand der Objekte, die sie gemeinsam abzubilden scheinen, miteinander verknüpft.

Sobald so zwei unterschiedliche Fotos gefunden wurden, die das Gleiche abbilden, können die möglichen Orte errechnet werden, von denen aus die Bilder gemacht sein könnten. An Hand des »image connectivity graph« können nun schrittweise zusätzliche Bilder in die Berechnung einbezogen werden und so, ausgehend von zwei Bildern, die möglichen Standorte der Kameras aller Bilder immer weiter eingegrenzt werden. Es entsteht eine Punktwolke, bestehend aus den Positionen aller Kameras. Jede Kamera bildet dabei den Fluchtpunkt des dazugehörigen Bildes, und Kamerapositionen und Bilder ergeben eine dreidimensionale Rekonstruktion der Szenerie. Dabei werden nicht nur die Teile des Bildträgers genutzt, die tatsächlich für das Zeichnen der digitalen Bilder verantwortlich sind. Die Bilddaten enthalten im Allgemeinen auch Daten, die ausschließlich für die Kommunikation von Maschinen untereinander bestimmt sind: Die *EXIF-Tags* digitaler Bildformate bergen Informationen, die die Kamera hinterlassen hat und die vor allem für ihre algorithmische Weiterverarbeitung bestimmt sind. Hat eine Kamera die Brennweite zum Zeitpunkt der Aufnahme im Bild gespeichert, kann *Photosynth* mit dieser Informationen die Bestimmung ihres Standorts verbessern.

Die wichtigste Information aber, die der Rekonstruktion eines solchen Bildraumes zu Grunde liegt, ist ebenfalls nicht Teil der Bilder. Sind die Bilddaten im Sprachgebrauch der Informatik der »content«, so bilden diese Informationen deren »context«. Erst wenn über ein Bild bekannt ist, von wo es stammt, kann versucht werden, es zur Rekonstruktion dieses ›Wo‹ zu benutzen. Und anders als die Brennweite ist der ungefähre Ort eines Bildes heute noch oft ausschließlich über manuell hinzugefügte Informationen zu erhalten: Für *Photosynth* müssen Autor_innen und Konsument_innen der Bilder diese mit Informationen darüber anreichern, was gezeigt wird, oder wo es gezeigt wird. Die inhaltsbasierten Methoden zur Konstruktion des »image connectivity graph« müssen um solche erweitert werden, die im Sinne des *Human Computing* auf Verhalten setzen. »Combining Context with Content« wird so zum Schlagwort eines Ansatzes, der sich auf Algorithmen stützt, die gleichermaßen mit Bilddaten arbeiten wie auf standardisierten Daten der zwischenmaschinellen Kommunikation (wie EXIF-Tags) und den digitalen Spuren von Nutzerverhalten.[9] *Human Computing* oder *Crowd Computing* trifft so auf Ansätze, die auf »content«, also Datensätze, zurückgreifen. *Photosynth* ist ein besonders exponiertes Beispiel für dieses Zusammentreffen – aber im Kern arbeitet auch die Suchmaschine Google mit der Kombination von »content« und »context«, wenn sie den Inhalt eines Dokuments in Beziehung zu den Verweisen setzt, die Menschen im Internet auf dieses Dokument gerichtet haben. Die Forscher hinter *Photosynth* machen klar, dass ihnen diese Verbindung wichtig ist, indem sie sich explizit auf die nutzergenerierten Inhalte im Internet beziehen: »There are billions of photographs on the Internet, comprising the largest and most diverse photo collection ever assembled. How can computer vision researchers exploit this imagery?«[10] Der Raum, den *Photosynth* aus Einzelbildern rekonstruiert, wird vor dem Hintergrund dieser »largest photo collection ever assembled« zum Abbild der Welt: Die Forscher überschreiben ihre Arbeit mit »Modeling the World from Internet Photo Collections«.[11] Diese so modellierte Welt ist dabei alles: Bilddatenbank, dreidimensionales Modell, navigierbarer Raum und Grundlage der Erzeugung neuer Bilder und realer Objekte.[12]

Der Raumrekonstruktion von *Photosynth* liegt also zunächst ein »connectivity graph« und dann eine Punktwolke zu Grunde. Punktwolken spielen bei der Konstruktion von 3D-Bildern eine wichtige Rolle, weil sie als räumliche Positionen zwischen der Vermessung von Objekten im Raum und ihrer räumlichen Darstellung im Bild vermitteln. Punktwolken machen dabei

Grober Regen fitzt in nassen Schauern und spritzt von unten.

deutlich, welche Verschiebungen mit dem *Spatial* oder *Topographical Turn* in Hinblick auf Funktionen, Praxen und vielleicht langfristig auch die Definitionen von Bildern einhergehen. Das ist ein Grund dafür, dass sie auch jüngst in medienkünstlerischen Arbeiten auf der Anschauungsebene thematisiert werden.

Bei Punktwolken handelt es sich einerseits um Diagramme, da sie aus verzeichneten Raumkoordinaten bestehen. *Photosynth* erzeugt zunächst eine Rekonstruktion der Koordinaten von Bildern und Kameras im Raum. Andererseits sind diese Diagramme schon ausgelegt auf eine Repräsentation mit hoher syntaktischer Dichte,[13] wenn der zu Grunde liegende Raum etwa als dreidimensionale Umgebung gezeichnet wird, die aus den zweidimensionalen Fotos besteht. Die Punktwolke markiert die Schnittstelle zwischen der Vermessung, der Sammlung und der Verbindung von Daten und ihrer Visualisierung. Insofern hat sie das Potenzial zur Metapher für die medientechnische enge Verschränkung von ausgewerteten optischen und anderen Informationen, Metakodierungen und dem Raum als Plattform für die Integration verschiedener Medien,[14] wie sie mit der algorithmischen Verknüpfung von »content« und »context« im Geoweb und ähnlichen Phänomenen virulent werden.

Mit den Medientechnologien verändert sich grundlegend die Erfahrung von Räumen und Orten. Statt den Ort als eine Konstellation von gegebenen, festen Punkten aufzufassen und den Raum als einen dimensionierten Behälter für materielle Objekte zu betrachten, werden Orte mehr und mehr als Netzwerke von Relationen, Verbindungen, Handlungen und gesammeltem Wissen erfahren. Der »image connectivity graph« im Herzen des Photosynth-Algorithmus lässt sich als paradigmatisches Beispiel eines solchen ortserzeugenden Netzwerks lesen, ist er doch Ergebnis eines maschinisierten Wissens über Fotografie und zentralperspektivische Abbildung, reale Orte und Handlungen im Umgang mit Bildern, sowie etwa deren Verschlagwortung. Mit Michel de Certeau würde man sagen, dass der Ort dadurch zum Raum wird, dass man etwas mit ihm macht; es sind Aktivitäten, die Räume schaffen und konstrurieren.[15] Nicht nur in der »Konstruktion von Wahrnehmung […] durch die Bewegung des Körpers«, sondern auch in der hybriden Aktivität zwischen »content« und »context«, zwischen Algorithmen und ihrer Benutzung, die *Photosynth* zu Grunde liegt.

Eichhorn,
mein Vögelchen blutet
Bärchen, den
Apfel der Wälder.

Wenn man an Panofskys Abhandlung zur Perspektive als symbolische Form zurückdenkt, in der er die Unterscheidung zwischen Aggregatraum der Antike und Systemraum der Neuzeit stark gemacht hat, dann stellt sich die Frage, wie man das heutige Raumkonzept anhand neuer Visualisierungstechniken interpretieren kann. Einen eingängigen Begriff scheint es in der Theorie bisher noch nicht zu geben. Manuel DeLanda ist von der Philosophie Gilles Deleuzes ausgegangen, um die Vermaschung oder Vernetzung von Räumen zu problematisieren.[16] In einem größeren Zusammenhang der historischen Anthropologie und unter Berücksichtigung der epistemologischen Verschränkung des Raumkonzeptes mit der neuzeitlichen Wissenschaft hat Löffler den Vorschlag gemacht den Raum als ein »Extensionsvermögen« aufzufassen.[17] Es würde zu weit führen, diese Vorschläge hier zu diskutieren, deshalb lässt sich abschließend hier nur das Folgende festhalten: Das, was wir mit dem Wahrnehmungsapparat unseres Körpers als Raum erfahren, stellt lediglich die Aktualisierung von bestimmten Tendenzen einer Multiplizität dar, die jederzeit durch die Aktualisierung von anderen Tendenzen mit Hilfe anderer Apparate komplexer gestaltet werden kann. Raum wird somit zur Variablen verschiedener Aktualisierungsprozesse, abhängig von Verfahren und technischen Medien. Jeglicher Raum – Körpererfahrung oder medial induzierte Erfahrung – bildet die Aktualisierung einer virtuellen Räumlichkeit, jeder aktualisierte Raum enthält virtuelle Tendenzen einer Entgrenzung. Wenn der Bildraum von *Photosynth* für dessen Entwickler zum Weltmodell wird, erinnert das vor allem daran, dass die Welt als Raum immer schon als solcher »Aktualraum« zu verstehen war.

Das Projekt *How to understand...* untersucht in unregelmäßigen Abständen eine aktuelle (Bild-)Technologie als Verfahren. Seinen Ausgangspunkt bildet ein Blog, der an der KHM von Urs Fries (technischer Leiter des Holografie-Labors), Stefanie Stallschus (künstlerisch-wissenschaftliche Mitarbeiterin im Bereich Kunstgeschichte im medialen Kontext) und Lasse Scherffig (künstlerisch-wissenschaftlicher Mitarbeiter im Bereich experimentelle Informatik) initiiert wurde.
http://how-to-understand.blog.de

Ich bin ein durch Eimer-Henkel kriechender Kübel.

1 Vergleiche http://www.ted.com/index.php/talks/blaise_aguera_y_arcas_demos_photosynth.html, zuletzt gesehen am 22.08.2009.

2 Zum Inhalsbegriff in der Informatik vergleiche: Lasse Scherffig, Stefanie Stallschus, Urs Fries: *How to understand... seam carving*. In: *off topic, Nr #0: übersetzen*, 2008, Köln, S. 94–97.

3 Vergleiche hierzu auch den Beitrag von Ludwig Zeller in diesem Heft.

4 Heinz von Foerster: *Wissen und Gewissen: Versuch einer Brücke*. Frankfurt am Main (Suhrkamp), 1993, S. 275–276.

5 Richard Andersen, David Bradley: Perception of three dimensional structure from motion. In: *Trends in Cognitive Sciences, 2*(6), 1998.

6 Zum Folgenden vergleiche Snavely et al.: Modeling the World from Internet Photo Collections. In: *International Journal of Computer Vision*, 80(2), 2007, S. 189–210.

7 Zu Salienz vergleiche Scherffig, Stallschus, Fries, S. 94–97.

8 Zu Bildträger und Bildobjekt vergleiche Lambert Wiesing: *Artifizielle Präsenz. Studien zur Philosophie des Bildes,* Frankfurt am Main (Suhrkamp), 2005, S. 48ff.

9 Mor Naaman: Eyes on the World. In: *IEEE Computer Magazine*, 39(10), 2006, S. 108–111.

10 Snavely et al., 2007, S. 189.

11 Ebd.

12 Mit Hilfe eines nicht ganz offiziellen Software *Mash-up*, lassen sich aus *Photosynth* die Koordinaten der modellierten Szene zurückgewinnen. Diese können dann zum Rendern neuer Bilder oder als Grundlage der Konstruktion realer Objekte verwendet werden. Vergleiche hierzu: Urs Fries: *Photosynth to 3D Conversion*, http://www.khm.de/mg/labd/wikka.php?wakka=PhotoSynth, zuletzt gesehen am 10.10.2009.

13 Um die Unterscheidung von Goodmann aufzugreifen, vergleiche Nelson Goodman: *Sprachen der Kunst. Entwurf einer Symboltheorie*, Frankfurt a. M. (Suhrkamp), 1995, (stw 1304), S. 212f.

14 Vergleiche Lev Manovich, Tristan Thielmann: Geomedien. Raum als neue Medien-Plattform? Ein Interview mit Lev Manovich. In: Jörg Döring, Tristan Thielmann (Hg.): *Mediengeographie. Theorie - Analyse – Diskussion*, Bielefeld (Transcript), 2009, S. 383–396.

15 Michel de Certeau: *Kunst des Handelns*, Berlin (Merve), 1988, S. 218.

16 Vergleiche Manuel DeLanda: *Intensive Science and Virtual Philosophy*, London/New York (Continuum), 2002, S. 45–81, bes. S. 69.

17 Vergleiche Davor Löffler: *Endlichkeitskaskaden. Fünf Aufsätze über den Rand*, Berlin (sine causa), 2009, bes. S. 34ff.

Abb. Photosynth-Rekonstruktion des Overstolzenhauses durch Urs Fries