Vom Foto zum virtuellen Raum

# Aufgeklappt

**Jürgen Schuck**

Automatic Photo Pop-up nennt sich eine Methode zur Umwandlung von Fotos in virtuelle Räume. Pate steht das Prinzip von Kinder- und Sachbüchern, deren Illustrationen sich beim Umblättern der Seiten zu dreidimensionalen Szenen entfalten.

Forscher der Carnegie Mellon University haben eine Software entwickelt, die zweidimensionale Bilder ohne menschliche Hilfe in dreidimensionale Ansichten verwandelt. In der Tat waren es zwei Motive, von denen sich die Urheber der Automatic Photo Pop-up genannten Technik leiten ließen: Die wissenschaftliche Herausforderung, die Abbildung einer realen Szene in einen entsprechenden Raum zu überführen, sowie der praktische Nutzen für den Betrachter von Alltagsfotos, dem die Methode einen individuellen und kurzweiligen Zugang auf den Schauplatz der Ablichtung ermöglicht.

Ein Bild wie das in Abbildung 1 ist für einen Betrachter leicht zu erfassen. Auch wenn er es zum ersten Mal sieht, kann er sich ein realistisches Bild von der Struktur der Szene machen. Er unterscheidet Boden und Himmel und bestimmt die Topologie der dazwischenliegenden Objekte. Anhand dieser Erkenntnisse kann er sich vorstellen, wie die Bildszene von beliebigen anderen Standorten betrachtet aussieht.

Ein Automat oder Programm benötigt für den Prozess der Raumbestimmung realer Gegenstände auf einem Foto eine Vorstellung von der allgemeinen Beschaffenheit der Welt: Materie schwebt in der Regel nicht lose im Raum, sondern befindet sich auf dem Boden. Die Schwerkraft bestimmt das Übereinander aller Objekte. Daraus lassen sich Annahmen über ihre Orientierung ableiten. Viele Dinge ähneln einander, beispielsweise sehen im Prinzip die Oberflächen (Texturen) von Häuserfassaden stets gleich aus. Gegenstände mit ähnlichen Charakteristiken lassen sich zu Klassen zusammenfassen, die Zusammenhänge ausdrücken wie „Gras ist meist grün und befindet sich auf dem Boden“.

Es handelt sich um erlerntes Wissen, das die praktisch unendliche Vielfalt geometrischer Interpretationen eines Fotos deutlich einschränkt, da die meisten denkbaren Topologien nach menschlicher Erkenntnis in der realen Welt nicht vorkommen. Außerdem liegt der Schluss nahe, dass menschliches Erkennen durch Sehen ein statistischer Vorgang ist, der nur wenig mit Geometrie zu tun hat. Automatic Photo Pop-up ist daher auch ein Beitrag zum klassischen Problem der geometrischen Rekonstruktion durch statistisches Lernen. Es bestimmt die geometrischen Parameter einer Abbildung durch einen Erkennungsprozess, der sie mit einer Gruppe von Referenzfotos vergleicht. Heraus kommt ein statistisches Modell geometrischer Klassen, die die Orientierung jedes Gegenstands in der Szene beschreiben. Das sogenannte Training-Set enthält verschiedene Außenaufnahmen mit rechtwinklig vom ebenen Boden aufragenden Gegenständen und Himmel. Auf dieses Bildmotiv ist Automatic Photo Pop-up zurzeit beschränkt.

Jedes Pixel der Bildvorlage fällt in eine der drei geometrischen Klassen: Boden, Vertikal und Himmel. Hinzu kommen die Horizontlinie und die relativen Positionen der als „Vertikal“ markierten Pixel. Zusammen ergeben die Parameter ein grobes Modell der 3D-Szene auf der Bildvorlage. Zur Markierung der Pixel verwendet der mit Entscheidungsbäumen arbeitende Machine-Learning-Algorithmus Eigenschaften wie Farbe, Textur, Bildposition und geometrische Merkmale.

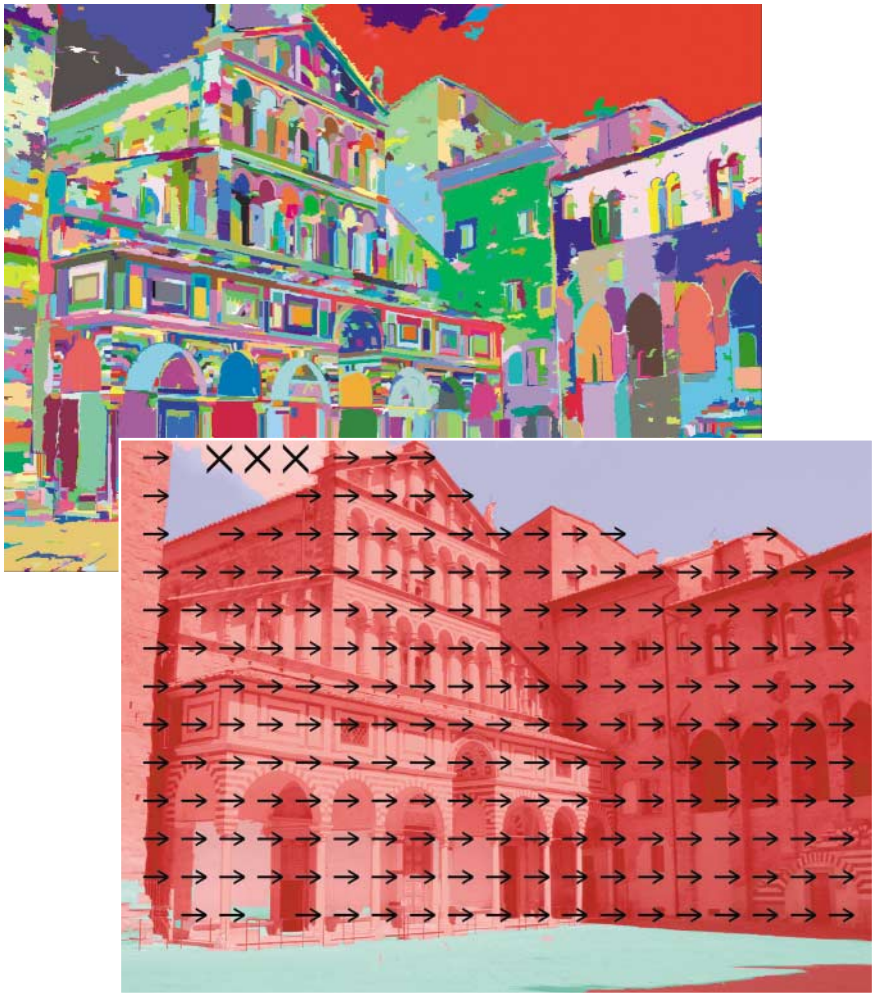
## Erkennungsdienstliche Behandlung

Ohne Wissen über die Struktur der Bildszene lassen sich aus den Farbwerten der einzelnen Pixel kaum nennenswerte Eigenschaften ermitteln. Der erste Schritt besteht daher in der Bestimmung von Regionen, sogenannter Superpixel, die im Idealfall ganze Gegenstände der Szene repräsentieren, zumindest jedoch große Teile davon. Das Verfahren nennt sich Segmentierung. Es stellt die im Originalbild gefundenen Regionen durch gleichfarbige Pixel dar (siehe Abbildung 2). Auf der Webseite des von den Entwicklern der Software empfohlenen Segmentierers findet sich eine allgemeine Einführung [1].

**X-TRACT**

- Automatic Photo Pop-up erstellt Raummodelle von Außenaufnahmen.
- Das Verfahren funktioniert automatisch und kommt mit einem einzigen Foto aus.
- Aus circa dreißig Prozent typischer Amateurfotos entstehen realitätsnahe 3D-Szenen.

**Ansichten: Das Ausgangsfoto der Piazza (links) als aufgeklappte Szene (rechts) mit verschiedenen Perspektiven der beiden Fassaden (untere Reihe). Liegt die gewählte Perspektive entgegen der ursprünglichen (unten links), ähnelt das Bild den paradoxen Raumdarstellungen M. C. Eschers (Abb. 1).**



**Aufgeteilt: Superpixelbild (hinten) des Fotos und der daraus abgeleitete geometrische Kontext mit Himmel (blau), Boden (grün) und Vertikal (rot). Die Pfeile geben die Orientierung der Ebenen an, Kreuze kennzeichnen unebene massive Objekte und Kreise (hier nicht vorhanden) durchsichtige (Abb. 2).**

Aus den Superpixeln, von denen ein Foto immer noch mehrere Hundert enthalten kann, lassen sich Farb- und Texturverteilung ermitteln. Die gewonnenen Werte ermöglichen eine weitere Gruppierung zu sogenannten Konstellationen, die Regionen mit wahrscheinlich gleichen Labels enthalten. Die Schätzung bedient sich der Werte aus dem Training-Set. Durch die Aufteilung in Konstellationen entstehen ausreichend große Bildbereiche zur Gewinnung des übrigen potenziell brauchbaren Zahlenmaterials. Im Idealfall entspricht eine Konstellation einem Gegenstand der Szene, beispielsweise einer Gebäudefront, oder einem großen Stück des Bodens oder des Himmels. Da das Gruppierungsergebnis unscharf ist, berechnet das Programm mehrere Sets möglicher Konstellationen. So ist die Wahrscheinlichkeit höher, dass es die Superpixel korrekt markiert. Die Bestimmung der Label ist die Kernfunktion von Automatic Photo Pop-up. Der dazu verwendete Machine-Learning-Algorithmus modelliert die Erscheinungsformen der geometrischen Klassen aus dem Trainings-Set. Darüber hinaus vertraut er auf die Annahme, dass alle Superpixel einer Konstellation dasselbe Label tragen und somit zum selben Gegenstand gehören.

Aus den markierten Pixeln lässt sich ein dreidimensionales Modell erstellen, das zunächst die Position der „vertikalen“ Objekte relativ zum Boden bestimmt. Dazu betrachtet der Algorithmus die Grenzlinien zwischen den Objekten der beiden geometrischen Klassen. Die Horizontlinie ergibt sich aus der Boden-Klasse und einigen weiteren geometrischen Eigenschaften. Der gefundene Horizont und die Annahme einer einzigen Boden-Ebene ermöglichen eine Abbildung der Boden-Pixel auf diese Fläche. Pixel der Klasse Vertikal repräsentieren aufragende Dinge wie Fahrzeuge, Gebäude oder Bäume, aber auch Menschen. Der Himmel ist in der einfachen Welt von Automatic Photo Pop-up durchsichtig und kann daher entfallen. Schließlich projiziert ein Texture-Mapping das Originalbild auf das berechnete Modell und stellt damit die 3D-Szene fertig.

## Vom Feature zur Klasse

Bei der Bestimmung der geometrischen Klasse berücksichtigt das Programm

Farbe, Textur, Bildposition und Form sowie die 3D-Geometrie einer Konstellation. Anhand der Farbe etwa lassen sich Aussagen über das Material einer Fläche machen. Der Himmel etwa ist meist blau oder grau, Boden oft grün oder bräunlich. Eine weitere Differenzierung erlaubt die Textur, etwa zwischen Wasser und Himmel oder zwischen Blättern und Gras. Dazu setzt die Software Gaußsche Filter in Verbindung mit dem Berkeley Segmentation Dataset [4] ein.

Durch die Bildpositionen der Pixel respektive Superpixel und Konstellationen lassen sich ebenfalls Erkenntnisse über die potenzielle Klassenzugehörigkeit gewinnen: Boden befindet sich in der Regel im unteren Teil des Bilds, Himmel im oberen. Vertikal-Konstellationen haben meist konvexe Form, während Konstellationen der Klasse Himmel selten konvex sind und oft eine größere Fläche besitzen.

Die 3D-Geometrie schließlich ermöglicht die Berechnung eines realistischen Raummodells der Szene. Über die Fluchtlinie einer Ebene lässt sich zum Beispiel leicht deren Orientierung relativ zum Betrachter festlegen [5]. Leider ist die Information aus einer relativ unstrukturierten Außenaufnahme schwer zu gewinnen, weshalb Automatic Photo Pop-up den indirekten Weg über eine Statistik der Schnittpunkte von durchgezogenen Linien geht, die natürlich erst zu finden sind [6]. Das Verfahren fasst die Schnittpunkte mehrerer beinahe paralleler Linien zusammen, wobei es deren Richtung und Entfernung vom Bildzentrum berücksichtigt. Ausführliche Beschreibungen des Verfahrens sowie weiterführende Literatur finden sich im Web ([2], [3]).

**Onlineressourcen**

| | |
|---|---|
| Automatic Photo Pop-up Software | www.cs.cmu.edu/~dhoiem/projects/popup/app.zip |
| Automatic Photo Pop-up Training-Set | www.cs.cmu.edu/~dhoiem/projects/popup/popupTrain.zip |
| Matlab Runtime-Umgebung | www.cs.cmu.edu/~dhoiem/projects/popup/MCRInstaller.zip |
| Segmentierer von Felzenszwalb und Huttenlocher | people.cs.uchicago.edu/~pff/segment/ |
| VMware Player | www.vmware.com/de/download/player/ |
| VMware Linux-Appliance | www.vmware.com/vmtn/appliances/ |
| Liste von VRML-Betrachtern | de.wikipedia.org/wiki/VRML |
| ImageMagick | www.imagemagick.org |

## Schneiden und Falten

Eine 3D-Szene entsteht durch Abschneiden des Himmels und Auffalten der als Vertikal markierten Pixel entlang der Vertikal-Boden-Linie. Die in der 3D-Geometrie gefundenen Linien organisieren die Vertikal-Pixel zu Flächen und beschreiben sie als Polyline-Objekte. An der gemeinsamen Kante

der Objekte mit der Boden-Linie erfolgt die Auffaltung um 90 Grad. Eine einfache Strategie steigert den Übereinstimmungsgrad der gefunden Geometrie mit der tatsächlichen Situation der Bildszene: Im Zweifel verzichtet das Programm darauf, einen Falz einzusetzen. Mit der Horizontlinie und festen Kameraparametern projiziert der Algorithmus die Szene auf 3D-Koordinaten. Das abschließende Texture-Mapping des Bodens und der darauf stehenden Gegenstände verwendet zwei teilweise transparente Varianten des Originalfotos, auf denen nicht zur Klasse Boden beziehungsweise Vertikal gehörende Pixel den Alpha-Wert Null haben.

Leider funktioniert das Verfahren nicht bei jedem Bild. Dafür arbeitet es vollständig automatisch und benötigt als Arbeitsgrundlage nur ein einziges Foto. Manchmal sind die errechneten Szenen schlichtweg falsch, nicht selten jedoch liefert es überraschende Ergebnisse. Andererseits schaffen es in der modernen Digitalfotografie von vielen Aufnahmen meist auch nur wenige aufs DVD-Album. Ein denkbares Szenario wäre, die Bilder in der Kamera umzurechnen und vorzusichten, damit sie als fertige 3D-Szenen auf den Computer gelangen, in die der Betrachter sofort eintreten kann – im Prinzip wie beim sogenannten Stitching, das ebenfalls schon in der Kamera die Einzelaufnahmen zu Panoramaansichten zusammensetzt. (mr)

## Installation

Wer die Software ausprobieren will, muss leider einige Pakete zusammentragen und installieren. Allerdings handelt es sich um wenig mehr als einen Proof of Concept, für den solche Laborbedingungen durchaus zulässig sind. Das eigentliche Programm ist im Archiv *app.zip* enthalten (siehe Kasten „Onlineressourcen“). Da es eine Matlab-Anwendung ist, benötigt man zusätzlich die Runtime-Umgebung, die die Entwickler leider nur für Linux zur Verfügung stellen können. Nutzer anderer Betriebssysteme können sich mit dem VMware-Player und einem vorgefertigten Image behelfen. Außerdem benötigt man den empfohlenen Segmentierer von Felzenszwalb und Huttenlocher sowie einen VRML-Viewer. Eine Liste von Kandidaten ist im Wikipedia-Artikel über VRML zu finden. Der Autor hat unter Windows gute Erfahrungen mit dem kostenlos erhältlichen Cortona VRML Client von Parallel Graphics gemacht.

Soll die Software hochauflösende Bilder bearbeiten, etwa von einer Digitalkamera, benötigt sie mindestens 1 GByte Hauptspeicher. Auch kann es unter Umständen notwendig sein, mit dem Shell-Kommando *ulimit –s unlimited* die Begrenzung des Stack aufzuheben. An dieser Stelle ein Wort zur Performance: Ein 1,7 GHz schnelles Notebook mit 1,5 GByte RAM und VMware (1,2 GByte Hauptspeicher) berechnete die Szene aus Abbildung 1 in etwa 45 Minuten. In der *iX*-Redaktion benötigte die Anwendung auf einem Core Duo T2400 (1,833 MHz, 1 GByte RAM) unter Linux rund 8 Minuten für ein Foto mit 2048 × 1536 Pixeln. Leider nutzte das Programm dabei nur einen Kern der Dual-Core-CPU.

Hat man alle Einzelteile beisammen, sollte man ein frisches Verzeichnis für die Installation anlegen, etwa *~/aphup*, und dort den Segmentierer (*segment.zip*) sowie die Pakete *app.zip* und *MCRInstaller.zip* auspacken. Dazu muss man den Linux-Entpacker *unzip* verwenden, da das Matlab-Paket symbolische Links enthält. Außerdem muss der Nutzer im Unterverzeichnis *segment* mit dem Kommando *make* den Segmentierer kompilieren. Er lässt sich anschließend unter dem Namen *~/aphup/segment/segment* aufrufen. Die eigentliche Anwendung ist in *~/aphup/app/photoPopup* zu finden. Damit sie ihre Laufzeitumgebung findet, muss man zusätzlich zwei Environment-Variablen setzen:

```
export XAPPLRESDIR=~/aphup/v73/X11/\
app-defaults
export LD_LIBRARY_PATH=~/aphup/v73/\
runtime/glnx86:~/aphup/v73/sys/os/glnx86:\
~/aphup/v73/sys/java/jre/glnx86/jre1.5.0/\
lib/i386/native_threads:~/aphup/v73/sys/\
java/jre/glnx86/jre1.5.0/lib/i386/client:\
~/aphup/v73/sys/java/jre/glnx86/jre1.5.0/\
lib/i386:~/aphup/v73/bin/glnx86
```

Wer sich die Tipparbeit sparen will, kann die Kommandos in *~/.profile* eintragen oder ein Wrapper-Skript schreiben, das die Variablen setzt und anschließend photoPopup aufruft.

Zum Ausprobieren eignen sich am besten Gebäudeaufnahmen, die Boden und Himmel durch eine jeweils gleichmäßige Färbung eindeutig erkennen lassen. Außerdem sollten mehrere Gebäudeseiten sichtbar sein. photoPopup erwartet die Bilder im JPEG-Format. Andere Formate lassen sich zum Beispiel mit dem Programm *convert* aus der ImageMagick-Suite umwandeln. Der Segmentierer hingegen kann nur Dateien im PPM-Format lesen. Eventuell ist daher ein zweiter Konvertierungsschritt nötig. Anschließend lässt sich mit *~/aphup/segment/segment 0.8 100 100 <bild>.ppm <bild>.pnm* das Superpixel-Bild erzeugen. Die Parameter sind Empfehlungen der Entwickler.

Sind alle Vorarbeiten erledigt, kann man ins Unterverzeichnis app wechseln und dort mit *./photoPopup ./classifiers_08_22_2005 <bild>.jpg pnm /tmp* die Metamorphose des Fotos in eine dreidimensionale Szene veranlassen. Die Datei *classifiers_08_22_2005.mat* enthält die geometrischen Klassen des Training-Set. Es folgen der Dateiname des Fotos, die Endung des Superpixel-Image (der Anfang muss dem des Fotos entsprechen) und schließlich das Verzeichnis, das die Ergebnisse aufnehmen soll: die VRML-Datei *<bild>.wrl* und die Texturen für die Boden- und Vertikal-Klassen in *<bild>.g.png* beziehungsweise *<bild>.v.png*. Darüber hinaus hinterlässt die Software einige Bilddateien mit Zwischenergebnissen im PGM-Format sowie die Datei *<bild>.l.jpg*, die den gefundenen geometrischen Kontext enthält.

JÜRGEN SCHUCK

ist Mitarbeiter der Materna GmbH und als Projektleiter im Bereich IT-Service-Management tätig.

## Literatur

[1] Pedro F. Felzenszwalb, Daniel P. Huttenlocher; Efficient Graph-Based Image Segmentation; people.cs.uchicago.edu/~pff/papers/seg-ijcv.pdf

[2] Derek Hoiem, Alexei A. Efros, Martial Hebert; Geometric Context from a Single Image; ICCV 2005; www.cs.cmu.edu/~dhoiem/publications/Hoiem_Geometric.pdf

[3] Derek Hoiem, Alexei A. Efros, Martial Hebert; Automatic Photo Pop-up; ACM SIGGRAPH 2005; www.cs.cmu.edu/~dhoiem/publications/popup.pdf

[4] David Martin, Charless Fowlkes, Doron Tal, Jitendra Malik; A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics; www.cs.berkeley.edu/~fowlkes/papers/mftm-iccv01.pdf

[5] Richard Hartley, Andrew Zissermann; Multiple View Geometry in Computer Vision, Second Edition; Cambridge University Press

[6] Jana Kosecka, Wei Zhang; Video Compass; European Conference on Computer Vision; Springer-Verlag

[7] André T. Martins, Pedro M.Q. Aguiar, Mário A.T. Figueiredo; Orientation in Manhattan: Equiprojective Classes and Sequential Estimation; www.lx.it.pt/~mtf/MartinsAguiarFigueiredo.pdf