Visual media editing using scene understanding

Author(s):
Mansfield, Alexander P.

Publication Date:
2014

Permanent Link:

https://doi.org/10.3929/ethz-a-010144656 →

DISS. ETH NO. 21756

Visual Media Editing Using Scene Understanding

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Sciences (Dr. sc. ETH Zürich)

presented by

Alexander Paul Mansfield
M.A. M.Eng., University of Cambridge
born on 29th July 1987
citizen of the United Kingdom of Great Britain and Northern Ireland

accepted on the recommendation of

Prof. Dr. Luc Van Gool, examiner
Prof. Dr. Carsten Rother, co-examiner
Prof. Dr. Niloy Mitra, co-examiner

2014

Abstract

Visual media are a powerful and important means of communication. However, the creation of high impact visual media requires editing of the raw data captured with cameras and 3D scanners. This editing process requires significant time and expert skill to carry out effectively, while cameras and 3D scanners are increasingly widely available and easy to use. This gap in effort motivates the development of automated editing tools.

Previous work in automated editing has mostly made use of low-level properties of the data, such as gradients and correlations. While these cues allow the preservation of local structures, they alone cannot capture complex structures such as the appearance of a person or the geometry of a building.

In contrast, the goal of this thesis is to exploit state-of-the-art computer vision algorithms to understand scenes represented in visual media and hence develop more intelligent automated visual media editing tools. In particular, we examine the use of scene understanding for the problems of retargeting and completion.

Retargeting is the problem of resizing visual media while minimising visible distortion, for example for viewing the visual media on devices with different screen sizes. Scene understanding allows distortion to be placed where it is not disruptive. We propose a method for image retargeting that decomposes the image into layers which are processed separately, allowing us to ensure that objects in the scene are not distorted, and that their relative depth ordering is preserved.

Completion is the problem of synthesising new visual information in certain parts of the visual media. This problem arises when holes are left by insufficient data for the reconstruction of a 3D model, or when artefacts need to be removed from a photograph. We propose a method for image completion which uses geometrically and photometrically transformed image patches for

completion. This allows the set of source patches to be increased, allowing for example the exploitation of structures in the scene with rotational symmetry. For 3D model completion, we propose two methods which exploit structural similarities in 3D shape within a class of objects. The first combines state-of-the-art methods from image completion with local shape models which can be learned for any class. The second globally aligns incomplete models of articulated objects, so that each can be used to fill holes in the others. Both of these methods allow the completion of large holes in a semantically correct manner.

Our proposed methods demonstrate the advantages of using scene understanding techniques in visual media editing: they allow complex editing to be performed with significant automation and minimal user effort. With these methods, we make a step towards closing the gap in effort between the capturing and editing of visual media.

Zusammenfassung

Visuelle Medien sind eines der wichtigsten Kommunikationsmittel der heutigen Zeit. Allerdings erfordert die Produktion ausdrucksstarker visueller Medien eine aufwendige und intensive Bearbeitung von Rohdaten, die beispielsweise durch Fotografie oder dreidimensionale Scanverfahren aufgenommen wurden. Des Weiteren setzt der Bearbeitungsvorgang Wissen und Fertigkeiten eines Experten voraus. Andererseits ermöglicht der technische Fortschritt und die weite Verbreitung von Aufnahmegeräten eine immer umfangreichere und schnellere Rohdatenerzeugung. Die Diskrepanz zwischen Bearbeitungsaufwand einerseits und Einfachheit und Geschwindigkeit der Datenaufnahme andererseits verlangt nach der Entwicklung von computerunterstützten Bearbeitungsmethoden.

Bisherige Methoden zur computerunterstützten Bearbeitung nutzen einfache lokale Dateneigenschaften wie Gradienten oder Korrelation, die geeignet sind, um lokale Strukturen wie Kanten zu beschreiben. Obwohl diese Bottom-Up-Repräsentationsform lokal eine konsistente Bearbeitung gewährleistet, können komplexe Strukturen, beispielsweise das Aussehen einer Person oder die Geometrie eines Gebäudes, mit diesen Methoden nicht beschrieben werden.

Im Gegensatz dazu ist das Ziel dieser Dissertation, eine dargestellte Szene im 2D- oder 3D-Raum mit modernen Methoden des maschinellen Sehens (*Computer Vision*) zu verstehen und darauf aufbauend computerunterstützten Bearbeitungsmethoden zu entwickeln. Die vorliegende Dissertation untersucht im Besonderen den Nutzen von Szenenverstehen für inhaltssensitive 2D- und 3D-Skalierung (*Retargeting*) sowie -Vervollständigung (*Completion*).

Retargeting ist eine Methode des Skalierens ohne sichtbare Verzerrung, zum Beispiel zur Darstellung auf Geräten mit unterschiedlichen Bildschirmgrössen. Szenenverstehen erlaubt es, die Verzerrung auf Bereiche zu beschränken, in denen sie nicht störend auffällt. Unser vorgeschlagenes Bildskalierensverfahren zerlegt das Bild in Ebenen, wobei jede Ebene einem Objekt entspricht. Die

Ebenen werden getrennt verarbeitet und somit sichergestellt, dass die dargestellten Objekte unverzerrt bleiben und die Anordnung der Objekte hinsichtlich ihrer Tiefe innerhalb der Szene beibehalten wird.

Completion ist eine Methode, bestimmte Teile von Bildern oder 3D-Modellen automatisch inhaltssensitiv auszufüllen. Dies wird beispielsweise benötigt, um Löcher zu füllen, die durch unzureichende Daten oder störende Objekte im Aufnahmeprozess entstehen und zur unvollständigen Rekonstruktion eines 3D-Objektes oder zu Artefakten in Fotoaufnahmen führen. In unserer Bildvervollständigungsmethode nutzen wir photometrisch und geometrisch transformierte Bildbestandteile zum Auffüllen. Dies ermöglicht, den Suchraum für passende Bildbereiche zu vergrössern, zum Beispiel durch Wiederverwenden von Bildbereichen, die sich unter Rotation ähnlich sind.

Bezüglich 3D-Modellvervollständigung werden im Rahmen dieser Dissertation zwei Methoden vorgestellt, die strukturelle Ähnlichkeiten innerhalb von Objektklassen ausnutzen. Die erste Methode kombiniert Bildvervollständigung mit lokalen Formmodellen, die für beliebige Objektklassen gelernt werden können. Die zweite Methode richtet unvollständige 3D-Modelle unter artikulierter Bewegung global aneinander aus, damit sie gegenseitig als Vorlage verwendet werden können, um fehlende Teile zu ersetzen. Dies erlaubt, selbst grosse strukturelle Löcher semantisch korrekt zu füllen.

Die im Rahmen dieser Dissertation entwickelten computerunterstützten Bearbeitungsmethoden demonstrieren den Nutzen des maschinellen Szenenverstehens in der Bearbeitung von visuellen Medien. Sie ermöglichen eine signifikante Automatisierung komplexer Bearbeitungsvorgänge, aufgrund derer aufwendige manuelle Benutzerintervention minimiert wird. Die vorgestellten Methoden bilden somit einen ersten Schritt, um die Diskrepanz zwischen Aufnahme- und Bearbeitungsaufwand visueller Medien zu reduzieren.