Acquisition, processing and display for 3D live-action cinema and television

Author(s):
van Baar, Jeroen

Publication Date:
2013

Permanent Link:

https://doi.org/10.3929/ethz-a-010140235 →

Diss.-No. ETH 21510

Acquisition, Processing and Display for 3D Live-Action Cinema and Television

A dissertation submitted to
ETH Zurich

for the Degree of
Doctor of Sciences

presented by
Jeroen van Baar
M.Sc. Delft University of Technology, the Netherlands
born 04 September 1973
citizen of the Netherlands

accepted on the recommendation of
Prof. Dr. Markus Gross, examiner
Prof. Dr. Marc Pollefeys, co-examiner
Dr. Paul Beardsley, co-examiner

2013

Abstract

Three-dimensional cinema and television involves the presentation of a separate image to a viewer's left and right eyes, in order to invoke a depth perception. Three-dimensional cinema and television provides filmmakers with an additional cue to aid in their storytelling. Current acquisition and manipulation approaches make it difficult to effectively exploit the additional depth dimension. In this thesis we examine the pipeline of acquisition, processing and display, and propose methods and approaches which make it easier to exploit the depth dimension, while also aiming to improve the quality of the three-dimensional viewing experience.

Computing a depth value for each pixel in the video images of a captured scene is a difficult task. We propose an acquisition system where a central, high quality film camera is supported with additional satellite sensors. Rather than using sensors of a single modality, e.g. visible light cameras, we propose to use additional modalities. Besides lower quality visible light cameras, we also incorporate a Time-of-Flight depth camera and a thermal camera. By combining sensors of different modalities we aim to provide more information for computing per-pixel depth. The satellite cameras allow for better occlusion reasoning of the scene. A depth camera provides a direct measure of scene depth, albeit at a low resolution. Finally, a thermal imaging camera provides information to correctly discern between different scene elements, when those scene elements are imaged as regions with similar colors. We propose a method to combine the information from multiple modalities and demonstrate that we can compute high quality depth maps.

Since we are dealing with motion pictures, it's not sufficient to compute depth only for a single instant in time. The computed depth should be temporally consistent for the video. We argue that the temporally consistent depth is of most importance for foreground objects in a scene. We propose an interactive approach which propagates segmented foreground objects from a begin and end frame of a shot, to the frames in between. By grouping pixels with similar photometric and thermal properties into so-called superpixels, we reduce the complexity from per-pixel to per-superpixel. We then pose the problem as a labeling problem for superpixels over time, where the label that is assigned to each superpixel indicates to which segment that superpixel belongs. We show that this information can be directly exploited in the depth computation, where the segments are used as prior knowledge in that computation.

For three-dimensional acquisition using a stereo pair of cameras, the arrangement of the cameras at acquisition time determines the amount of depth that can be perceived by the viewers. The depth of the underlying scene has to be recovered in order to change the amount of depth perceived. In general, the processing of three-dimensional content cannot be performed independently for the left and right eye images, but should also take the underlying scene depth into account. We propose a processing method which can copy elements from a scene captured with one particular arrangement of cameras, and then paste those elements into a scene with a different arrangement. We demonstrate a method that even in the case where the recovered depth cannot be accurately estimated, we can robustly copy and paste elements. We further demonstrate how the underlying scene depth can be exploited when the element is pasted, and avoid the difficult task of scene in-painting, while aiming to conform to the stereo properties of the target scene.

Display of stereoscopic three dimensional content provides the user the ability to perceive a depth impression. The key factor is to ensure that each eye is only stimulated with the corresponding image of the stereoscopic image pair. The case where information intended for one eye is perceived by the other eye, is denoted as crosstalk or ghosting. The presence of ghosting may result in objects being perceived at the incorrect depth, and even result in the depth impression being entirely lost. Ghosting also puts a relatively heavy burden on the visual system of the viewer, with visual fatigue as the consequence. We identify that display systems are not perfect and propose a computational approach to mitigate the occurrence of ghosting in stereoscopic three dimensional display systems. Our approach is based on incorporating perceptual metrics to compensate the input images in such a way as to provide a perceptually more optimal viewing experience.

Zusammenfassung

In dieser Doktorarbeit befassen wir die Pipeline zur Erfassung, Verarbeitung und Darstellung von dreidimensionalen Inhalten für Kino und Fernsehen.

Die Beiträge die in dieser Doktorarbeit gemacht werden, können als wie folgt zusammengefasst werden:

- **Verfassungssystem auf einem einzigen Referenzkamera, durch multimodale Satellitsensoren unterstützt, für die Berechnung der Tiefekarten und Segmentierung**
- **Fusion multimodaler Sensorinformationen zu berechnen der Tiefekarten mit einem lokalen Verfahren.**
- **Interaktive Video Segmentierung Methode mit multimodalen Sensorinformationen. Das Ergebnis der Segmentierung wird zur Berechnung verbesserte Tiefekarten verwendet.**
- **Ein System für Kopieren und Einfügung Bearbeitung von stereoskopischen 3D-Inhalten mit Tiefe und Segmentierung Informationen.**
- **Ein Wahrnehmungsbasiertes System für die Kompensation der Lichtverschmutzung durch sogenannte Geisterbilder in stereoskopische 3D Abbildungssysteme. Das Kompensationssystem ist allgemein und kann für alle Formen der additiven Lichtverschmutzung in Abbildungssysteme angewendet werden.**

Cinematographers und Kamerabetreiber haben sich gewöhnt um mit einer einzigen Kamera zu erfassen. Wir schlagen daher eine multimodalen Verfassungssystem vor, mit einer zentralen hochqualitativen Kamera der mit verschiedenen Arten von Sensoren erweitert wird, um die Berechnung der Tiefekarten und Segmentierung zu unterstützen. Unser Prototyp zeigt, dass es relativ einfach ist, ein solches System zu bauen. Auch der Durchführung von geometrischen Kalibrierung und Farbkalibrierung ist relativ einfach.

Wir beschreiben eine lokale Methode basiert auf der Fusion der verschiedenen Modalitäten für die Berechnung von Tiefekarten. Tiefekarten werden für die hohe Qualität Referenz Kamera in das Verfassungssystem berechnet. Experimentelle Ergebnisse werden für Szenen mit dynamischen Objekten und Hintergrundkram gezeigt. Okkludierungen, Texturlose Regionen, wiederholende Texturen, oder ähnlich farbigen Vorder- und Hintergrundobjekte können vor Probleme sorgen in Methoden die sich nur auf

Farbkonsistenz verlassen. Mehrere SatellitKameras ermöglichen es uns Okklusionsregionen besser abzuschatzen, durch der Farbkonsistenz zwischen der Referenzkamera und die SatellitKameras auf der linken Seite, und die Farbkonsistenz zwischen der Referenzkamera und die SatellitKameras auf der rechten Seite, zu vergleichen. Die Fusion von Stereo mit Time-of-Flight Tiefedaten ergibt der richtigen Rekonstruktion von Texturlose Regionen wie Hintergrundwänden. Daneben können gleichfarbige, aber in unterschiedlichen Tiefen überdeckende Flächen, korrekt rekonstruiert werden. Von besonderem Interesse ist der Fall, wenn menschliche Subjekte oder Körperteile sich überdecken. Wir haben gezeigt, dass verschiedene Subjekte unterschiedliche thermische Signaturen haben können. Daher, durch die Fusion von thermischen Daten kann ein okkludierende Kontur gefunden werden, obwohl die Hautfarbe ähnlich ist. Wir vergleichen die Fälle von Fusion von Stereo mit dem Time-of-Flight Tiefedaten, Fusion mit den thermischen Daten und Fusion mit den beiden Time-of-Flight Tiefe und thermischen Daten. Obwohl jede dieser Modalitäten separat zu ein verbesserte Tiefekarte beitragen können, die Kombination aus beiden gibt Tiefekarte mit den beste Ergebnisse.

Eine zentrale Aufforderung bei der Berechnung der Tiefekarten ist die Schätzung der Okklusionsbereichen in einer Szene. Die Verwendung mehrerer Satellitkameras auf beiden Seiten einer Referenzkamera, verhilft zu einer besseren Schätzung. Um Kosten und Stellfläche der Verfassungssystem zu reduzieren, schlagen wir vor, geringere Qualität Satellitkameras zu verwenden. Dagegen hatten geringere Qualität Kameras mehr Bildrauschen als die hochqualitativen Kamera. Dies gilt vor allem in dunkleren Bildbereichen. Darüber hinaus haben die Satelliten und Referenzkameras auch sehr unterschiedliche Farbräume. Diese Eigenschaften beeinflussen die Genauigkeit der Farbkonsistenz zwischen den Satellitkameras und dem Referenzkamera, und damit die gesamte Präzision. Das berechnen von Tiefekarten basiert auf Farbkonstanz allein, werd immer von Vieldeutigkeiten leiden. Fusion mit zusätzlichen Modalitäten ist daher eine vielversprechende Richtung zur Lösung einiger dieser Vieldeutigkeiten. Die Auflösung des Time-of-Flight Kameras ist sehr niedrig im Vergleich zu den Referenzkamera. Feine Details, wie die Blätter einer Pflanze, können daher nicht genau durch der Time-of-Flight Kamera erfasst werden. Fusion mit niedriger Auflösung Time-of-Flight Tiefe funktioniert deshalb am besten für Bereiche ohne feine Details. Wärmebilder sind besonders nützlich, wenn das thermische Kontrast ausreichend hoch ist. Dies ist typischerweise der Fall bei Szenen mit menschlichen Akteuren.

Wir beschreiben ein interaktives Video Segmentierung Methode für die Segmentierung mehrere Vordergrund Objekte vom Hintergrund. Unsere Meth-

ode propagiert bekannte Segmentierungen für das erste und letzte Bild zu der Zwischenbilder in einer Videosequenz. Die propagierung stützt sich auf der Übereinstimmung von Superpixeln in der Videosequenz, ohne Annahme auf die Art der Bewegungen in einer Szene. Unsere Methode kann deshalb bewegten Kameras und nicht-starr bewegten Objekten verarbeiten. Das Ausnutzen mehreren Modalitäten trägt zu ein mehr robuster Übereinstimmung der Superpixel zwischen Frames einer Sequenz bei. Die Propagierung von bekannte Segmentierungen kann okkludierende Objekte verarbeiten. In sofern dass Objekte im Vordergrund in einer Sequenz sowohl in der ersten und letzten Rahmens sind, können sie dann für der Zwischenbilder verschwinden und wieder erscheinen. Falls optical flow Informationen verfügbar sind, kann es einfach für die Übereinstimmung von Superpixel eingearbeitet werden.

Ein vollautomatisches System kann zu ein falschen Segmentierung führen. Wir schlagen daher vor, um einen Benutzer interaktiv die Propagierung einer Segmentierung zu begleiten. Wir benötigen Korrekturen nur auf einen groben Niveau statt auf Pixelniveau. Das verringert die Belastung für den Benutzer. Genaue Segmentgrenzen werden dann in einem nachfolgenden Schritt produziert. Mehrere Modalitäten können dann genutzt werden zur Lösung der Farbenvieldeutigkeiten, und produzieren dann besseren Segmentgrenzen. Die Segmentgrenzen sind zeitlich stabil, weil sie genau die Grenzen der Objekte im Video passen. Wir können die Grenzen als Randbedingungen in die Berechnung der Tiefekarten verwenden, so dass die Tiefesilhouetten zeitlich mehr stabil werden.

Wir beschreiben ein System für 3D Kopieren & Einfügen. Das System baut das 2D Kopieren & Einfügen für Standbilder zu stereoskopischen 3D aus. Die Rekonstruktion der Tiefekarte für die Szene ist die grundlegende Operation in diesem System. Die rekonstruierten Tiefekarten können bei der Durchführung des interaktiven Segmentierung benutzt werden, für die Propagierung des Segmentierungsergebnis für das Bild entsprechend mit einem Auge, zu das Bild entsprechend mit dem anderen Auge. Die rekonstruierten Tiefekarten können auch bei die Zusammensetzung der segmentierten Objekte in der Zielszenen benutzt werden. Segmentierung, die Propagierung, und Zusammensetzung werden alle von höherer Qualität Tiefkarten profitieren. Direktes Zusammensetzung mit benutzung eine Tiefekarte, würde eine fehlerfreie Tiefekarte benötigen. Fehlerfreie Tiefekarten können jedoch selten für allgemeine Szenen erhalten werden. Zusammensetzung basiert auf Tiefekarten mit Fehlern, können stattdessen Proxygeometrie und parametrische Verzögungen benutzen.

Bei Zusammensetzung unter verschiedenen Orientierungen, oder in ein Ziel-

szene mit verschiedenartige stereo Parameter, könnte Okklusionsbereichen wieder sichtbar werden. Für realistische Ergebnisse müssten diese wieder sichtbare Okklusionsbereichen eingemalt werden. Wir zeigen stattdessen dass, durch die Anwendung der entsprechenden Einschränkungen für das berechnen der parametrische Verzögungen, wieder sichtbare Okklusionsbereichen ganz vermieden werden können. Die Ergebnisse bleiben jedoch immer noch überzeugend.

Wir beschreiben ein System für die Kompensation von Geisterbildern und Lichtstreuung. Durch die Formulierung der Kompensation als Optimierungsproblem, können wir den System in alle Fälle von zusatzliches Lichtverschmutzung anwenden. Als solches ist unsere Formulierung eine Verallgemeinerung der vorhandenen subtraktiven Kompensationsverfahren. Da wir für den menschlichen Beobachter kompensieren werden, sollten wir die Eigenschaften des menschlichen visuellen System nutzen. Wir zeigen wie Beobachtungsmetriken in die Optimierungsformulierung eingearbeitet werden können. Insbesondere durch die Integrierung der Kontrastempfindlichkeitsfunktion und das Lösen des resultierenden Optimierungsproblem, wird der restliche Fehler in Regionen, in denen das menschliche visuelle System weniger empfindlich ist, verbreitet. Am wichtigsten ist, dass die Wahrnehmbarkeit von möglicherweise widersprüchliche Randcues für Stereosehen, für Wahrnehmungsbasierte Deghosting reduziert ist. Dies macht gerade das beobachten von stereoskopische 3D Abbildungssysteme noch komfortabler. Ein Benutzerstudie wurde durchgeführt, um sicherzustellen, dass unsere Methode in der Tat von Benutzer bevorzugt wird, statt einfache subtraktive Kompensation.

Summary

In this thesis we address the pipeline for acquisition, processing and display of three-dimensional contents for cinema and television.

The contributions made in this thesis can be summarized as:

- **Acquisition system based on a single reference camera, supported by multi-modal satellite sensors, for computing depth maps and segmentation**
- **Fusion of multi-modal sensor information to compute depth maps using a local method.**
- **Interactive video segmentation approach using multi-modal sensor information. The result of the segmentation is used for computing improved depth maps.**
- **A framework for copy and paste editing of stereoscopic 3D content using depth and segmentation information.**
- **A perceptually-based framework for the compensation of light pollution due to ghosting in stereoscopic 3D displays. The framework is general and can be applied to all forms of additive light pollution in display systems.**

Cinematographers and camera operators are used to capture with a single camera. We therefore propose a multi-modal capture system, using a central high quality reference camera augmented with different types of sensors to support the computation of depth maps and segmentation. Our prototype system demonstrates that it is relatively straightforward to build such a system, including performing geometric calibration and color calibration.

We describe a local method based on fusion of the different modalities for computing depth maps. Depth maps are computed for the high quality reference camera in the capture system. Experimental results are shown for scenes with dynamic objects and background clutter. Occlusions, textureless regions, repeated textures, or similarly colored fore- and background objects may pose problems in methods that rely only on color consistency. Multiple satellite cameras allow us to better estimate occlusion regions by comparing the color consistency between the reference camera and the satellite cameras on the left side, to the color consistency between the reference camera and the satellite cameras on the right side. The fusion of stereo with Time-of-Flight

depth data results in the correct reconstruction of textureless regions such as background walls. In addition, surfaces of the same color, but overlapping at different depths can be correctly reconstructed. Of particular interest is the case where human subjects or body parts are overlapping. We showed that different subjects may have different thermal signatures. Therefore, by also fusing the thermal data, an occluding contour can be found even though the skin color is similar. We compared the cases of fusion of stereo with only the Time-of-Flight depth data, fusion with only the thermal data, and fusion with both Time-of-Flight depth and thermal data. Although each of these modalities separately can help improve the depth map, the combination of both gives the best result.

A key challenge in computing depth maps is the estimation of occlusion areas in a scene. Using multiple satellite cameras on either side of a reference camera, helps to better estimate occlusions. To reduce cost and physical footprint of the acquisition system, we propose to use lower quality satellite cameras. However, lower quality cameras exhibit more noise than the high quality reference camera. This is particularly true in low light areas. In addition, the satellite and reference cameras also have very different color spaces. These properties affect the accuracy of the color consistency between the satellite cameras and the reference camera, and therefore the overall accuracy as well. Computing depth based on color consistency alone will always suffer from ambiguities. Fusion with additional modalities is therefore a promising direction to help solve some of these ambiguities. The Time-of-Flight depth camera resolution is very low compared to the reference camera. Fine details, such as the leaves of a plant, are therefore not accurately captured with the Time-of-Flight depth camera. Fusion with low resolution Time-of-Flight depth thus works best for areas without fine details. Thermal images are most useful when thermal contrast is sufficiently high. This is typically the case for scenes with human actors.

We describe an interactive video segmentation approach to segment multiple foreground objects from the background. Our approach propagates known segmentations for the first and last frame to the intermediate frames in a video sequence. The propagation relies on the matching of superpixels across the video sequence, without any assumption on the motions in a scene. Our method can thus handle moving cameras and non-rigidly moving objects. Exploiting multiple modalities helps to make the matching of superpixels between frames of a sequence more robust. The propagation of known segmentations can handle occluding objects. Provided that foreground objects within a sequence are present in both the first and last frame, they may then disappear and re-appear for the intermediate frames. If optical flow information is available, it can be easily incorporated for the matching of superpixels.

A fully automated method may produce the wrong segmentation. We thus propose to employ a user to interactively guide the propagation of a segmentation labeling. We require corrections only at a coarse level, rather than at the pixel level, which reduces the burden on the user. Accurate segment boundaries are produced in a subsequent refinement step. Multiple modalities can then be exploited to help resolve color ambiguities and result in better refinement boundaries. The segment boundaries are temporally stable as they accurately match the object boundaries in the video. We can use the boundaries as constraints in the computation of depth maps, so that the depth silhouettes become temporally more stable as well.

We describe an end-to-end system for 3D copy & paste, which extends 2D copy & paste for still images to stereoscopic 3D. The reconstruction of the depth map for the scene is the fundamental operation in this system. The reconstructed depth maps can be used when performing the interactive segmentation, for the propagation of the segmentation result for one eye image to the other eye image, and for composition of the segmented objects into the target scenes. Segmentation, propagation, and composition will all benefit from higher quality depth maps. Direct composition based on the depth map on the other hand, would require an error-free depth map. Error-free depth maps are rarely obtained for general scenes however. Compositing based on depth maps with errors can instead be done using proxy geometry and parametric warps.

When compositing under different orientation, or into a target scene with different stereo parameters, disocclusions could occur. For realistic results these disocclusions would have to be inpainted. Instead we show that by applying the appropriate constraints to compute the parametric warps, disocclusions can be avoided altogether, while still achieving compelling results.

We describe a framework for the compensation of ghosting and scattering. By formulating the compensation as an optimization problem, we can apply the framework to additive light pollution in general. As such, our formulation is a generalization of existing subtractive compensation methods. Since we are compensating for human observers, we should exploit the properties of the human visual system. We show how we can incorporate perceptually-based metrics into the optimization formulation. Specifically, by incorporating the Contrast Sensitivity Function and solving the resulting optimization problem, the residual error is distributed to regions where the human visual system is less sensitive to them. Most importantly, the perceptibility of possibly conflicting edge cues for stereopsis is reduced for perceptual-based deghosting. This makes watching stereoscopic 3D displays more comfortable. A user study was conducted to verify that our perceptually-based compensa-

tion method is indeed generally preferred over straightforward subtractive compensation.