Diss. ETH No. 20290

Video-Based Rendering Techniques

A dissertation submitted to
ETH Zurich

for the Degree of
Doctor of Sciences

presented by
Marcel Germann
Dipl. Informatik-Ing., ETH Zurich, Switzerland
born 28. March 1980
citizen of Switzerland

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

2012

Abstract

The goal of video based rendering is to render images or videos of a scene from novel viewpoints, based on video footage from one or more real cameras. This has a high potential especially for outdoor sports events, where usually several cameras are available, but adding other technology into the scene is expensive or not allowed. However, in such uncontrolled setups, the input suffers from several drawbacks. The cameras are usually sparsely placed, causing wide baselines between them. Positioned far from the scene, the cameras are also difficult to calibrate, i.e., to compute their positions, viewing directions and internal parameters based on the images. Therefore, the set of usable cameras is reduced to those with wide-angle shots, causing the coverage of subjects in the scene to be at low resolution. In this thesis we present two different approaches to render novel views in such difficult outdoor setups.

The first approach is based on a body pose estimation used to construct *articulated billboards*, a novel representation of the human body. First, a coarse pose guess is established according to comparisons of silhouettes with a database. From the k best 2D pose estimations of the individual cameras, the optimal combination is chosen according to errors in the 3D triangulation, which results in a 3D pose estimation. After a consistency test to remove left/right flips of arms or legs between frames, the body poses are optimized in a spatio-temporal energy minimization. This includes terms for smoothness, silhouette fitting as well as data-driven terms to favor plausible poses. The articulated billboards are placed according to the results of this pose estimation and consist of a billboard fan per body part. In a novel view-dependent blending and rendering technique they can be shown from arbitrary viewpoints.

The second approach introduces an adaptive reconstruction method and a view-dependent geometry morph. A separate 2.5D representation is obtained in every camera image by a coarse-to-fine reconstruction. It uses sparse feature match correspondences as well as back-projection errors to find optimal depth values for the vertices of a 2.5D triangulation and to adaptively subdivide triangles only where it is required. A refinement step according to back-projections and neighbor look-ups improves the found vertex depths. It results in a 2.5D reconstruction per camera, which are merged into a final 3D representation. Novel viewpoints

are rendered not only with a view-dependent blending but also with a view-dependent geometry. For this, a morph of the geometry is achieved by a force field computed from non-epipolar feature matches. The reconstruction is robust to several errors occurring particularly in outdoor setups and the view-dependent rendering corrects for calibration errors, for which no 3D reconstruction would fit to all camera images.

For both approaches we present results based on conventional TV camera footage of several soccer scenes. The quality of the images and videos is comparable to those of the input footage and the results show the potential of both approaches. We conclude this thesis with a comparison of the two approaches as well as a collection of ideas for future work.

Zusammenfassung

Das Ziel von Video-basiertem Rendering ist es von einer Szenerie, die mit einer oder mehreren Kameras aufgenommen wurde, Bilder oder Videos von einem neuen Blickwinkel zu zeigen, wo keine Kamera plaziert war. Dies hat ein hohes Potential, speziell für Aussenaufnahmen von Sportereignissen, die normalerweise von mehreren Kameras aufgenommen werden, wo es aber teuer oder nicht erlaubt ist, weitere Technologie der Szene hinzuzufügen. Die Eingangsdaten solcher Aussenaufnahmen weisen jedoch verschiedenen Mängel auf. Normalerweise sind nur wenige Kameras vorhanden, was bedeutet, dass die Distanzen zwischen ihnen gross sind. Da diese auch weit vom Geschehen entfernt sind, sind sie ausserdem schwierig zu kalibrieren, bzw. es ist schwierig anhand der Bilder zu bestimmen, wo sie positioniert sind, wohin sie schauen und welche internen Parameter sie haben. Dies reduziert die Menge der brauchbaren Kameras auf solche mit Weitwinkel-Ansichten, was wiederum bedeutet, dass die Personen in der Szene mit niedriger Auflösung abgebildet sind. In dieser Arbeit präsentieren wir zwei verschiedene Ansätze, um in solch schwierigen Aussenaufnahmen neue Ansichten zu rendern.

Der erste Ansatz basiert auf einer Körperposenschätzung die für die Konstruktion von *Articulated Billboards*, einer neuartigen Repräsentation des menschlichen Körpers, verwendet wird. Zuerst wird anhand von Silhouettenvergleichen mit einer Datenbank eine grobe Schätzung der Pose berechnet. Von den k besten 2D Posenschätzungen der einzelnen Kameras wird anhand von 3D Triangulierungsfehlern die optimale Kombination ausgewählt, woraus eine 3D Posenschätzung entsteht. Nach einem Konsistenztest zur Entfernung von Rechts/Links-Verwechslungen der Arme und Beine wird die Pose in einer spatio-temporalen Energieminimierung optimiert. Diese beinhaltet Terme für Glätte, Terme für das Passen auf Silhouetten, sowie datengetriebene Terme um plausible Posen zu bevorzugen. Die Articulated Billboards werden dann anhand der Resultate der Posenschätzung plaziert und bestehen aus einem Billboard-Fächer pro Körperteil. In einer neuen blickwinkelabhängigen Rendertechnik können diese von beliebigen Ansichten gerendert werden.

Der zweite Ansatz führt eine adaptive Rekonstruktionsmethode und ein blickwinkelabhängigen Geometriemorph ein. In jeder Kamera wird eine separate 2.5D Repräsentation mittels eines grob-zu-fein Verfahrens errechnet. Es verwendet dünngesiedelte Merkmalspaarungen sowie Rückprojektionsfehler um optimale

Tiefenwerte für die Ecken einer 2.5D Triangulierung zu finden und um adaptiv nur solche Dreiecke zu unterteilen wo es notwendig ist. Ein Verfeinerungsschritt anhand von Rückprojektionsfehlern und Nachbarschaftsabfragen verbessert die gefundenen Ecktiefen. Dies resultiert in einer 2.5D-Rekonstruktion pro Kamera, welche dann in eine finale 3D Repräsentation vereint werden. Neue Ansichten werden nicht nur mit einem blickwinkelabhängigen Blenden sondern auch mit einer blickwinkelabhängigen Geometrie gerendert. Hierzu wird ein Morph der Geometrie mittels eines Kraftfeldes erreicht, das durch nicht-epipolare Merkmalspaarungen berechnet wird. Die Rekonstruktion ist robust gegen verschiedene Fehler die speziell bei Aussenaufnahmen auftreten und das blickwinkelabhängige Rendering korrigiert Kalibrationsfehler, bei denen keine 3D Rekonstruktion in alle Kamerabilder passen würde.

Für beide Ansätze präsentieren wir Resultate die auf Aufnahmen von normalen TV-Kameras von verschiedenen Fussballspielen basieren. Die Qualität der Bilder und Videos ist vergleichbar mit denjenigen der Eingangsdaten und die Resultate zeigen das Potential beider Ansätze. Zum Schluss dieser Arbeit vergleichen wir die zwei verschiedenen Ansätze und erläutern Ideen füer zukünftige Arbeiten.