DISS. ETH NO. 22207

Camera Calibration and Human Pose Estimation for Sports Broadcasts and Human Performances

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Jens Puwein

MSc Computer Science
ETH Zurich

born on 03.02.1983
citizen of Buchs (AG), Switzerland

accepted on the recommendation of

examiner
Prof. Marc Pollefeys

co-examiners
Prof. Adrian Hilton
Prof. Markus Gross

2014

Abstract

3D reconstruction and augmented reality of sports events or, in general, human performances are popular applications of computer vision in the entertainment industry. Often, all that is available to create a 3D reconstruction of an event is the video footage provided by the uncalibrated cameras filming it. However, knowing the positions, orientations, and the internal properties of the cameras is crucial for most of those applications. Furthermore, the reconstruction of humans and their movements requires the knowledge about the 3D positions of their joints. In this thesis, methods to obtain the camera parameters from video footage in three different scenarios are proposed. The estimation of human poses is addressed for single images and multi-view videos.

The typical setting encountered in sports broadcasts is characterized by wide-baseline camera setups and planar playing fields. In this thesis, such scenes are categorized into one of two common cases. First, scenes with distinctive textures and second, scenes that are mostly homogeneous. In the first case, distinctive textures are leveraged to match features across wide baselines and to build an abstract feature representation of the playing field. New video sequences are calibrated and integrated into the system, updating the feature representation. To address the second case, the geometric constraints that need to be fulfilled by player trajectories recorded in different cameras are used to calibrate a network of cameras.

If, in contrast to broadcasts of team sports, only a single person is captured, this person usually covers large parts of the images. In a third method for camera calibration, this is exploited to obtain an initial camera calibration by identifying different joints of the human body to establish correspondences between cameras in a static camera network. Given the correspondences, the camera parameters and the 3D joint positions representing the moving person are initialized. Subsequently, the camera parameters and the 3D joint positions are optimized jointly, further improving the accuracy of both the human poses

and the camera parameters.

The reliable estimation of human poses from single images is an important basic component of computer vision. Besides the fact that in some applications a single image is all that is available, this component is a valuable tool for many other applications. For example, it can serve as an initialization for pose estimation in videos or from multiple views. Often, a coarse segmentation of a person is available a priori. To incorporate such information during pose estimation from a single image, a globally optimal branch-and-bound algorithm augmenting basic pose estimation models with a global foreground term is proposed. To deal with the exponential size of the search space, a combination of custom tailored upper bounds is proposed along with methods for more efficient inference.

Zusammenfassung

Dreidimensionale Rekonstruktionen und erweiterte Realität sind verbreitete Anwendungen des maschinellen Sehens im Bereich der Unterhaltungsindustrie. Um eine dreidimensionale Rekonstruktion eines Ereignisses zu erzeugen, ist oftmals bloss das Videomaterial der unkalibrierten Kameras vorhanden, welche das Ereignis filmen. Es ist jedoch für die meisten Anwendungen unabdingbar, dass die Positionen, die Orientierungen und die internen Eigenschaften der Kameras bekannt sind. Hinzu kommt, dass die Rekonstruktion von Menschen und derer Bewegungen die Kenntnis der dreidimensionalen Positionen der Gelenke erfordert. In dieser Doktorarbeit werden Methoden zur Bestimmung der Kameraparameter für drei verschiedene Szenarien vorgeschlagen. Die Schätzung von menschlichen Posen wird behandelt für Einzelbilder und für Videos, welche von mehreren Kameras aufgenommen wurden.

Für den typischen Aufbau bei Sportübertragungen sind grosse Abstände zwischen den Kameras und planare Spielfelder charakteristisch. In dieser Arbeit werden zwei gängige Fälle unterschieden. Erstens, Szenen mit markanten Texturen und zweitens, Szenen, welche grösstenteils homogen sind. Im ersten Fall werden markante Texturen dazu benutzt, um Korrespondenzen zwischen den Kameras zu finden und um eine abstrakte Darstellung des Spielfeldes zu erstellen. Neue Videosequenzen werden kalibriert und in das System integriert. Zugleich wird die abstrakte Darstellung aktualisiert. Um den zweiten Fall anzugehen, werden geometrische Bedingungen, welche Spielertrajektorien aus unterschiedlichen Kameras erfüllen müssen, dazu verwendet, um ein Kameranetzwerk zu kalibrieren.

Wird im Gegensatz zu Übertragungen von Teamsportarten nur eine einzelne Person gefilmt, so ist diese meistens gross im Bild sichtbar. Dies wird in einem dritten Ansatz zur Kalibrierung von Kameras ausgenutzt, indem die verschiedenen Gelenke des menschlichen Körpers identifiziert werden, um Korrespondenzen zwischen den Kameras eines statischen Kameranetzwerks zu finden.

Anhand der Korrespondenzen werden die Kameraparameter und die dreidimensionalen Positionen der Gelenke der sich bewegenden Person initialisiert. Anschliessend werden die Kameraparameter und die dreidimensionalen Positionen der Gelenke gemeinsam optimiert, was zu einer höheren Genauigkeit der menschlichen Posen und Kameraparameter führt.

Das zuverlässige Schätzen von menschlichen Posen anhand einzelner Bilder ist eine wichtige grundlegende Komponente des maschinellen Sehens. Davon abgesehen, dass in bestimmten Anwendungen nur einzelne Bilder zur Verfügung stehen, stellt diese Komponente für viele andere Anwendungen ein wertvolles Werkzeug dar. Ein Beispiel ist die Initialisierung der Schätzung von Posen in Videos oder aus mehreren Perspektiven. In vielen Fällen ist eine grobe Segmentierung der Person im Bild von vornherein vorhanden. Um solche Informationen in die Schätzung von Posen anhand einzelner Bilder miteinzubinden, wird ein global optimaler Algorithmus basierend auf Branch-and-Bound vorgeschlagen, welcher elementare Modelle zur Schätzung von Posen mit einem globalen Term für Vordergrund erweitert. Um die exponentionelle Grösse des Suchraums handzuhaben, werden eine Kombination von massgeschneiderten oberen Grenzen und Methoden zur effizienteren Inferenz vorgeschlagen.