Diss. ETH No. 16682

# Markerless Motion Capture of Complex Human Movements from Multiple Views

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH

for the degree of
Doctor of Sciences ETH Zürich

presented by
Roland Kehl
Dipl. El. Ing. ETH
born $23^{th}$ April 1976
Swiss citizen

accepted on the recommendation of
Prof. Dr. Luc Van Gool, examiner
Prof. Dr. D.M. Gavrila, co-examiner

July 2006

# Zusammenfassung

Forschung die sich mit dem Aufzeichnen von Bewegungen artikulierter Objekte befasst wird immer populärer, speziell im Fall des menschlichen Körpers. Die Möglichkeit solche Bewegungen in Videosequenzen zu erkennen und zu verfolgen ist die Basis für viele interessante Anwendungsbereiche, wie z. Bsp. Überwachung, Mensch-Maschine-Schnittstellen oder computergenerierte Animationen in der Film- und Spiel-Industrie. Diesen Anwendungen zu Grunde liegt die Aufgabe, menschliche Bewegungen auf Videobildern zu erkennen und digital aufzuzeichnen, ein Verfahren das als *Motion Capture* bekannt ist.

Kommerzielle, video-basierte Motion Capture Systeme setzen voraus, dass die Person einen schwarzen, eng anliegenden Anzug mit weissen, klar sichtbaren Markierungen trägt. Allerdings ist solch ein Anzug nicht in jeder Situation erwünscht oder gar anwendbar. Daher stellt diese Arbeit eine Methode vor die es erlaubt, komplexe menschliche Bewegungen ohne jegliche Markierungen am Körper aufzuzeichnen.

Ohne diese Markierungen wird die Aufgabe ungleich schwieriger und erfordert robuste Bilddaten und komplizierte Optimierungs-Methoden. Der vorgeschlagene Ansatz stützt sich auf eine Kombination aus verschiedenen Arten von Daten, wie z. Bsp. Bildkanten, Farbe und eine 3D Rekonstruktion der Person. Wir zeigen dass solch eine Kombination in der Lage ist unsere Methode robust gegen Doppeldeutigkeiten zu machen, wie sie z. Bsp. entstehen wenn Körperteile sich berühren oder sich gegenseitig verdecken.

Ein artikuliertes Modell des menschlichen Körpers, gebaut aus Superellipsoiden, bildet die Basis unseres Ansatzes. Wir verwenden Stochastic Meta Descent (SMD) Optimierung um diejenige Pose zu finden welche die Bilddaten anhand des Modells am besten erklärt. Dabei helfen wenige, zufällig ausgewählte Stichproben SMD von lokalen Minima der Kostenfunktion zu befreien und erlauben es gleichzeitig schneller ans Ziel zu kommen als mit anderen Methoden.

Wir veranschaulichen die Leistungsfähigkeit unseres Ansatzes anhand von mehreren anspruchsvollen Sequenzen und zeigen, dass bereits fünf Kameras ausreichen um komplexe Bewegungen aufzuzeichnen. Dabei wird die menschliche Pose mit 24 Freiheitsgraden mit bis zu 2 Bildern pro Sekunde berechnet.

# Abstract

Tracking of articulated structures such as human bodies has gained in popularity over the past years. Applications include surveillance, human-computer interaction and computer based animations in games and the movie industry. These applications require the solution of a common task: the one of recovering the body pose from observed images. This task of digitally recording the motions of a person is called *Motion Capture*.

Commercial vision-based Motion Capture systems require the person to wear a tight-fitting costume with white, clearly outstanding dots as markers on it. However, markers and the corresponding special clothing are not always desirable or even applicable. We present a method for markerless tracking of complex human motions from multiple camera views.

In the absence of markers, the task of recovering the pose of a person during complex motions is challenging and requires strong image features and robust tracking. We propose a solution that integrates multiple image cues such as edges, color information and a 3D reconstruction of the person. We show that a combination of multiple image cues helps the tracker to overcome ambiguous situations such as limbs touching or strong occlusions of body parts.

Following a model-based approach, we match an articulated body model built from superellipsoids against these image cues. Stochastic Meta Descent (SMD) optimization is used to find the pose which best matches the images. Stochastic sampling makes SMD robust against local minima and lowers the computational costs, as a small set of predicted image features is sufficient for optimization.

We illustrate the performance of our method by several challenging sequences where a person wears different, casual clothes. Five cameras are sufficient for tracking complex motions and full body articulation. The power of SMD is demonstrated by comparing it to the commonly used Levenberg-Marquardt method. Results are shown for several challenging sequences showing complex motions and full articulation, with tracking of 24 degrees of freedom in approx. 1-2 frames per second.