Diss. ETH No. 18017

Statistical Models for Human Body Pose Estimation from Videos

A dissertation submitted to the
ETH ZURICH

for the degree of
Doctor of Sciences

presented by

Tobias Jaeggli
MSc, Université de Fribourg
born July 26, 1976
citizen of Winterthur ZH

accepted on the recommendation of

Prof. Dr. Luc Van Gool, examiner
Prof. Dr. Cristian Sminchisescu, co-examiner

2008

Abstract

To investigate the task of multidimensional continuous inference from video sequences on a concrete example application, we focus on the problem of articulated 3D human tracking from monocular video. This is an interesting topic because of its relevance for biological vision systems, as well as its many applications in various domains. Estimating body pose and motion of humans is a challenging task, with difficulties such as self-occlusions and ambiguities. To account for unresolvable uncertainties of the visual analysis of such footage, we formulate the task as a probabilistic inference problem. The pose estimation and tracking algorithms are based on statistical models that can be automatically learned from a set of example data. Thanks to this architecture, the proposed approaches remain general and can be tailored to a specific task by the choice of training data sets. Prior knowledge can be provided in a flexible and theoretically well-motivated way. First, we propose an approach that is based on a model of the joint probability distribution of body pose and the corresponding human shape, as it can be observed in video images. Both body pose and shape are treated as multivariate random variables, by choosing suitable representations. The statistical model uses a mixture of Gaussian distributions to approximate the density, which enables efficient discriminative inference of body poses from shape descriptors. When additionally taking the unknown image locations of the persons into account, the posterior distributions become non-parametric. Therefore, a hybrid inference scheme based on a Rao-Blackwellised particle filter combines parametric inference with sample based inference. A second approach is based on a generative predictive model of human shape, using nonlinear regression. To enable efficient learning and sample based inference, a low-dimensional embedding of human locomotion is determined, with a nonlinear dynamical model. This method is implemented using Locally Linear Embedding, and Relevance Vector Machines for sparse nonlinear regression. We also propose an integrated formulation of the model, fully based on Gaussian Process regression. The resulting tracking algorithms are tested on realistic video sequences with low resolution and image noise. We present extensions of the framework, for simultaneously tracking multiple persons that occlude each other, and for recognising the performed activity along with the pose estimation.

Zusammenfassung

In dieser Arbeit wird das Problem untersucht, wie aus Videosequenzen mittels probabilistischer Inferenz mehrdimensionale kontinuierliche Vektoren geschätzt werden können. Als konkretes Beispiel dient dabei das visuelle Verfolgen von Fussgängern, und insbesondere das Bestimmen ihrer Körperposen und -bewegungen. Das wissenschaftliche Interesse an dieser Fragestellung ergibt sich aus ihrer Relevanz für biologische visuelle Systeme, sowie aus der Vielzahl möglicher Anwendungen in verschiedenen Bereichen, die eine automatische Analyse von solchem Bildmaterial benötigen. Zu den Herausforderungen dieser Aufgabe gehören u. a. (Selbst-) Verdeckungen und Ambiguitäten. Die probabilistische Formulierung des Problems erlaubt es, verbleibende Unsicherheiten miteinzubeziehen. Die vorgeschlagenen Algorithmen basieren auf statistischen Modellen, welche automatisch von Beispieldaten gelernt werden können. Dank dieser Architektur bleiben die Ansätze allgemein, und können durch die Wahl von Trainingsdaten an verschiedene Szenarien angepasst werden. Auch kann *a priori*-Wissen auf flexible und fundierte Weise in den Analyseprozess eingespiesen werden. Der erste vorgeschlagene Ansatz basiert auf einem Model der gemeinsamen Wahrscheinlichkeitsverteilung von menschlichen Körperposen und dazugehörigen Körperformen, wie sie in Bildern beobachtet werden können. Körperposen und -formen werden dabei als mehrdimensionale Zufallsvariablem behandelt, indem passende Deskriptoren definiert werden. Das statistische Modell besteht aus einer Kombination von mehreren Normalverteilungen (*Gaussian Mixture Model*), wodurch auf diskriminative Weise effizient von beobachteten Form-Deskriptoren auf Posen geschlossen werden kann. Wenn zusätzlich die Bildposition der Figuren geschätzt werden soll, können die *a posteriori*-Wahrscheinlichkeitsverteilungen nicht mehr parametrisch dargestellt werden. Deshalb wird ein hybrides Inferenz-Schema vorgeschlagen, welches parametrische Inferenz mit Stichproben-basierter Inferenz (*Partikelfilter*) kombiniert. Ein zweiter Ansatz lernt ein generatives Modell zur Voraussage von Körperformen anhand von Posen, basierend auf nicht-linearer Regression. Um den Lernprozess und Stichproben-basierte Inferenz zu ermöglichen, werden Körperposen mittels Mannigfaltigkeiten modelliert, die in Vektorräumen von relativ geringer Dimensionalität eingebettet sind, verglichen mit der Dimensionalität der ursprünglichen Deskriptoren. Zusätzliches *a priori*-Wissen

wird in Form eines nicht-linearen dynamischen Modells gelernt. Die Algorithmen wurden auf anspruchsvollen realistischen Videosequenzen mit geringer Auflösung und schlechter Bildqualität getestet. Zusätzliche Erweiterungen erlauben die gleichzeitige Analyse der Bewegungen mehrerer Personen und die automatische Erkennung und Unterscheidung verschiedener Klassen von Bewegungsmustern.