Vision-Based Human Motion Analysis

Author(s):
Yao, Angela

Publication Date:
2012

Permanent Link:

https://doi.org/10.3929/ethz-a-007563521 →

DISS. ETH NO. 20366

Vision-Based Human Motion Analysis

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences (Dr. sc. ETH Zürich)

presented by

Yingjie Angela Yao
MSc ETH in Biomedical Engineering
born February 18, 1983, citizen of Canada

accepted on the recommendation of

Prof. Dr. Luc Van Gool, ETH Zürich and KU Leuven, examiner
Dr. Cordelia Schmid, Inria Grenoble, co-examiner

2012

Abstract

Interpreting human activity from video is at the core of a wide spectrum of applications such as content-based indexing, intelligent surveillance, human-computer interfacing and sports video analysis. Cheap hardware and growing storage capacity has led to an explosion of video data and there is a critical need for machine vision algorithms that automatically analyze video content.

This thesis provides a collection of methods for video-based human action recognition, *i.e.* the application of semantic labels to a person's movements over time in a video sequence. We present two approaches for this task, one appearance-based and one pose-based. The appearance-based method uses no structural modeling of the human body and relies only on the statistical distribution of appearance features such as edges, shapes and flow to classify actions. The pose-based method, on the other hand, explicitly estimates a 3D articulated pose of the body and classifies the action based on geometric relations between specific joints in a single pose or a short sequence of poses.

In addition to determining action labels, we examine how action recognition could be leveraged to help with the closely related task of human pose estimation. We integrate action recognition and pose estimation into a single system, taking output from appearance-based action recognition as a prior for 3D pose estimation. The estimated poses are then used to for pose-based action recognition to refine the action label. Finally, we examine the temporal aspect of labeling actions and propose a method to both segment and classify actions from a continuous stream of body poses.

Zusammenfassung

Die Auswertung menschlicher Tätigkeiten in Videos ist zentral für ein breites Spektrum von Anwendungen, wie zum Beispiel Indizierung von Videos, intelligente Videoüberwachung, Mensch-Computer-Schnittstellen oder Sportvideoanalyse. Die stetig fallendenden Kamerapreise und die zunehmenden Speicherkapazitäten führten jedoch zu einem enormen Zuwachs an Videodaten und damit zu einem dringenden Bedarf an Lösungen zur automatischen Videoanalyse.

Diese Dissertation handelt davon, die Bewegungen von Personen in Videosequenzen semantisch zu annotieren. Zu diesem Zweck wird eine bildbasierte und eine posenbasierte Methode präsentiert. Der bildbasierte Ansatz interpretiert Bewegungen ohne Vorwissen über Skelett oder Körperform eines Menschen und basiert auf der statistischen Verteilung von Bildmerkmalen wie Kanten, Gradienten und optischer Fluss. Der posenbasierte Ansatz hingegen verwendet die geschätzte Pose in Form eines Skelettes im dreidimensionalem Raum und benutzt die raum-zeitlichen Relationen der Gelenke zueinander.

Da die Schätzung der menschlichen Pose und die Erkennung der menschlichen Tätigkeit eng miteinander verbunden sind, untersuchen wir, inwieweit das Lösen der einen Aufgabe hilfreich für die andere ist. Hierzu vereinen wir Lösungsansätze für beide Probleme in ein einzelnes System und verwenden die Resultate der bildbasierten Methode zur Tätigkeitserkennung als apriorisches Wissen für die Poseschätzung. Die hiermit ermittelten Posen verwenden wir wiederum, um die annotierten Tätigkeiten mit Hilfe des posenbasierten Ansatzes zu verbessern. Zum Schluss stellen wir eine Methode vor, die nicht nur die Tätigkeiten klassifiziert, sondern auch die Länge der Tätigkeit genau segmentiert.