Joint face and head tracking inside multi-camera smart rooms

Zhenqiu Zhang · Gerasimos Potamianos · Andrew W. Senior · Thomas S. Huang

Received: 26 October 2006 / Revised: 26 March 2007 / Accepted: 30 March 2007 / Published online: 30 May 2007
© Springer-Verlag London Limited 2007

Abstract The paper introduces a novel detection and tracking system that provides both frame-view and world-coordinate human location information, based on video from multiple synchronized and calibrated cameras with overlapping fields of view. The system is developed and evaluated for the specific scenario of a seminar lecturer presenting in front of an audience inside a “smart room”, its aim being to track the lecturer’s head centroid in the three-dimensional (3D) space and also yield two-dimensional (2D) face information in the available camera views. The proposed approach is primarily based on a statistical appearance model of human faces by means of well-known AdaBoost-like face detectors, extended to address the head pose variation observed in the smart room scenario of interest. The appearance module is complemented by two novel components and assisted by a simple tracking drift detection mechanism. The first component of interest is the initialization module, which employs a spatio-temporal dynamic programming approach with appropriate penalty functions to obtain optimal 3D location hypotheses. The second is an adaptive subspace learning based 2D tracking scheme with a novel forgetting mechanism, introduced to reduce tracking drift and increase robustness. System performance is benchmarked on an extensive database of realistic human interaction in the lecture smart room scenario, collected as part of the European integrated project “CHIL”. The system consistently achieves excellent tracking precision, with a 3D mean tracking error of less than 16 cm, and is demonstrated to outperform four alternative tracking schemes. Furthermore, the proposed system performs relatively well in detecting frontal and near-frontal faces in the available frame views.

Keywords Person tracking · Face detection · Multi-camera tracking · Dynamic programming · Adaptive subspace tracking · Mean-shift tracking · AdaBoost · Lecture data · Smart rooms

1 Introduction

Visual detection and tracking of humans is an important problem with numerous applications that range from automated surveillance to interfaces for human–computer interaction. In general, robust human tracking in complex scenes is challenging. In some circumstances however, multiple time-synchronized and calibrated camera sensors with overlapping fields of view may be available, from which both frame-view and world-coordinate human location information can be derived. In such scenarios, efficiently combining frame-level appearance-based human detection with temporal and spatial constraints constitutes a viable approach. This can simultaneously provide both desired types of location information with improved accuracy, while avoiding reliance on any form of background modeling or motion estimation.
The paper introduces a novel human tracking vision system employing these principles, developed and evaluated for the specific scenario of tracking a seminar lecturer presenting inside a “smart room” in front of an audience.

This scenario is of central focus in the European integrated project Computers in the Human Interaction Loop (CHIL [1]). In CHIL, smart rooms have been set up, equipped with multiple audio and visual sensors that include a minimum of four calibrated and time-synchronous cameras with highly overlapping fields of view, located at the room corners (see also Figs. 1 and 2). Numerous seminars have been recorded in such rooms providing a large multi-sensory and multimodal database of real human interaction [2]. The resulting CHIL corpus, annotated with a wealth of multimodal information, has been crucial to the development and evaluation of technologies for perception of humans in the lecture scenario of interest [3,4]. Prominent among such technologies is the task of locating the lecturer’s head position, both in the three-dimensional (3D) space—in the form of head centroid world coordinates, as well as in the available two-dimensional (2D) frame views—as bounding boxes of visible faces [5]. Such location information can be further utilized in support of numerous audio-visual perception technologies: For example, 2D face information is useful for person identification [6], whereas 3D location coordinates can be employed in acoustic beamforming for far-field automatic speech recognition [7], as well as to obtain close-up presenter views based on steerable pan-tilt-zoom cameras or camera selection schemes [8,9]. The views can further assist identification [10] and audio-visual speech technologies [11], among others, with obvious utility in lecture indexing and understanding of the interaction.

It becomes clear that for the CHIL lecture scenario described above a visual system that combines face detection, tracking, and multi-camera processing is both feasible and desirable. This paper introduces such a system, developed to provide both 2D-face and 3D-head location information of a single person (the lecturer) in CHIL seminars. Like most 3D approaches, the proposed algorithm consists of a sequence of 3D initialization and tracking phases, with a tracking drift

Fig. 1 Overview of the CHIL lecturer video tracking task. Schematic diagrams of the smart rooms located at two CHIL project partners: a Universität Karlsruhe (UKA), Germany, and b Istituto Trentino di Cultura (ITC), Italy. The CHIL lecture corpus, used in our experiments for single-person (lecturer) tracking, has been collected at these two sites (see also Fig. 2)

Fig. 2 Examples of synchronous four camera views of the a UKA and b ITC data, part of the CHIL lecture corpus [2]. In such recordings, a standing subject presents a lecture in front of a small (mostly sitting) audience. Notice the highly overlapping fields of view of the four cameras, set up to ensure that at least two cameras capture the lecturer head at any given instant