Detailed Human Data Acquisition of Kitchen Activities: the CMU-Multimodal Activity Database (CMU-MMAC)

Fernando De la Torre
Carnegie Mellon University
5000 Forbes av.
Pittsburgh, PA 15232
ftorre@cs.cmu.edu

Jessica Hodgins
Carnegie Mellon University
5000 Forbes av.
Pittsburgh, PA 15232
jkh@cs.cmu.edu

Javier Montano
Carnegie Mellon University
5000 Forbes av.
Pittsburgh, PA 15232
javier.montano.martinez@gmail.com

Sergio Valcarcel
Carnegie Mellon University
5000 Forbes av.
Pittsburgh, PA 15232
sergio@3ces.com

Copyright is held by the author/owner(s).
CHI 2008, April 5 – April 10, 2008, Florence, Italy
ACM 978-1-60558-012-8/08/04.

Abstract

Over the past decade, researchers in computer graphics, computer vision, and robotics have begun to work with significantly larger collections of data. A number of sizable databases have been collected and made available to researchers: faces, motion capture, natural scenes, and changes in weather and lighting. These and other databases have done a great deal to facilitate research and to provide standardized test datasets for new algorithms, however, these databases are limited by the constrained settings within which they are collected. We propose a focused effort to capture detailed (high spatial and temporal resolution) human data in the kitchen while cooking several recipes. The database contains multimodal measures of the human activity of subjects performing the tasks involved in cooking and food preparation. Currently we record video from five external cameras and one wearable camera, audio from five balanced microphones and a wearable watch, motion capture with a 12 camera Vicon systems, and accelerometers, gyroscopes and magnetic sensors from five IMUs. Several computers were used for recording the various modalities. The computers were synchronized using the Network Time Protocol (NTP). Preliminary data can be downloaded from http://kitchen.cs.cmu.edu/, and it is currently used to solve problems of multimodal temporal segmentation of activities and activity recognition.

Keywords

Multimodal data, kitchen recording, audio, video, motion capture, inertial measurement units, wearable sensors.

Introduction

Over the past decade, researchers in computer graphics, computer vision and robotics have begun to work with very large collections of data to model human motion (e.g. [1]). These databases have been used to construct models of human ~~movement~~motion for which researchers have found many applications in sports science, medicine, biomechanics, animation of avatars in games or movies, surveillance, better strategies for humanoid robots, and human activity recognition among others. These databases have facilitated research and provided standardized test datasets for algorithms. However, many of these databases are limited by the constrained settings within which they are collected. For instance, current human motion capture databases typically capture the motion of professional actors, athletes, and artists who were brought into the studio for their specific talents, rather than a wide range of individuals performing everyday tasks. As a result, the motions in current motion databases are often performances—examples of finely honed skills, or clear caricatures of ordinary events.

Similar in spirit to recent work by Intille et. Al [2], the CMU-Multimodal Activity database (CMU-MMAC) aims to overcome some of the previous limitations by collecting high resolution spatial and temporal multimodal (audio, video, accelerations, motion capture) samples of human behavior. To capture human behavior in settings that are as natural as possible, we have installed an almost fully operable kitchen and captured the cooking of several meals from start to finish. The kitchen is a very important test bed because food preparation and eating are core elements of daily life and hence essential for a data collection that purports to represent the space of natural human activities. Moreover, the kitchen is key to a number of socially significant applications. For instance, an inability to reliably prepare a balanced diet is often a decisive factor in a move to assisted living for the elderly or cognitively impaired. Figure 2 illustrates the location of the sensors and several views of the kitchen constructed with working appliances.

Sensing modalities

This section explains the hardware and software components for each modality: video, audio, accelerometers/gyroscopes and optical motion capture.

Video: We used five FireWire cameras manufactured by Point Grey Research Inc (see Figure 1.a). Three of these cameras (FL2-08S2C) captured high resolution images (1024×768 pixels) at 30 fps. The other two cameras (FL2-03S2C) have lower resolution (640×480 pixels) but higher frame rate (60 fps) to capture faster motion. Both cameras were full native IEEE-1394b (FireWire b) standard compatible, allowing transmission speeds of 800 Mbits/s. The firewire interface enables full frame rate RGB image transmission at the maximum camera resolution. For more information

Figure 1. a) camera, b) microphones, c) IMUs.

A Firewire camera, FL2-08S2C (640 × 480 pixels), is attached to the mounting bracket from a headlamp is placed around the head of the subject. The video is

recorded in a laptop in the back, along with the time stamp for each individual frame. This laptop also recorded the data coming from the accelerometers and gyroscopes.

Multi channel audio: The audio modality was captured using five Behringer condenser B-5 Microphones (see Figure 1.b). These high quality Microphones offer different pickup patterns (cardioid or omnidirectional) in order to capture a specific source of sound or a general sample of sound quality and improve the noise immunity. Each Microphone was connected to an audio Pre-amplifier, which was then connected to a Professional M-Delta Audio PCI card that captured the audio in a regular computer. Audio was recorded and processed under a Unix system with a modified version of the audio editing software Audacity.
To capture audio we placed a total of six microphones around the kitchen. One was placed above the kitchen; the others were attached to cabinets, above the sink and by the refrigerator. All microphones were professional studio condenser microphones made by Behringer (model B-5) and capable of cardiod and omnidirectional polar patterns.

Inertial Measurement Units: Our third modality is captured with MicroStrain's 3DM-GX1 inertial measurement units. These units contain an accelerometer, gyroscope, and magnetometer. They combine these signals to measure absolute orientation, as well as angular velocity and instantaneous acceleration. All signals are gyro-stabilized and recorded at a regular rate (roughly 76 Hz). The data is captured using LabView software running on the laptop computer carried by the subject. Five sensors are placed on the subjects back, legs, and arms. After synchronous activation of all devices, the program cycles through each of the five sensors to capture a portion of their data stream.

Motion Capture: Our subjects' whole body and hand motions were recorded. Current state-of-the-art for whole body capture uses a set of 40-60 markers to approximate the rigid body motion of 15-22 segments, see Figure 2.b.. To the extent possible, the markers are placed on joint axes and bony landmarks so that they can more easily be used to find the motion of an idealized skeleton. The hands are often modeled as a single rigid link. We used biomechanical invariants to reduce the number of markers to less than the number required to fully specify the orientation of each rigid body segment. The system used was a Vicon motion capture (120 Hz) setup with twelve MX-40 cameras for the captures. The standard motion capture setup was refined to have additional markers as shown in Fig. 5.

Synchronization across modalities

Camera synchronization is achieved with MultiSync software. The MultySync software is designed to synchronize the image acquisition of multiple compatible point Grey cameras across different IEEE-1394b buses on the same computer and across separate buses on multiple computers connected by an standard Firewire cable.

Synchronization among all modalities is achieved with the Network Time Protocol (NTP), which synchronizes the clocks among computer systems over packet-switched, variable-latency data networks. NTP uses UDP port 123 as its transport layer. We use a NTP server from CMU (ac-ntp0.net.cmu.edu) with stratum two, which has a tency smaller than 2ms milliseconds. Timestamp files are generated for the cameras and for the IMUs. All the files contain the number, relative time between frames or samples and the local clock This local clock is obtained from the relative time between frames or samples of internal devices' clocks (camera trigger and IMU's tick counter) and the local clock of the computer at the first frame or sample, which is synchronized by NTP.

Synchronization with audio is achieved using the time when the recording stream is opened from the local clock of the computer (which is also synchronized by NTP) and taking into account the sampling frequency.

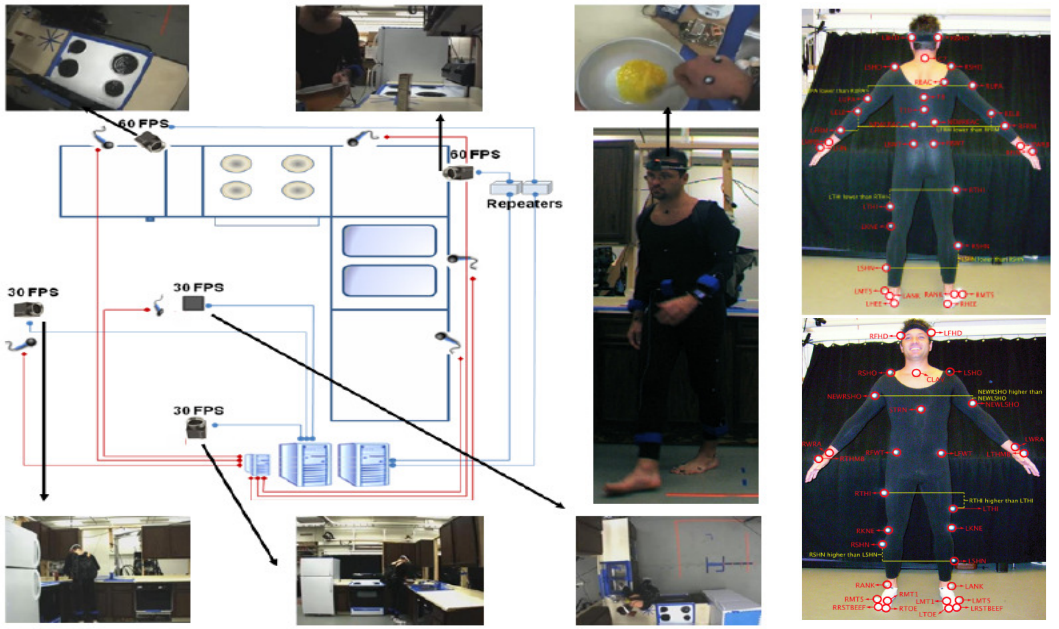

Figure 2: a) Sensor placement in the kitchen. b) Marker placement in motion capture data

Expected uses of the data

We expect the uses of the human motion database to span a number of disciplines. Because of the high spatial and temporal resolution, the natural quality of the database, and the use of several modalities, we expect that it will be used for the development of techniques for virtual rehabilitation, assistive technology, monitoring applications, build models of human actions, and as reference database for testing pattern recognition techniques (e.g activity recognition, temporal segmentation, biometric identification from video) and data fusion algorithms.

Acknowledgements

We thank Adam Bargneil, Xavi Martin, Alex Collado, Pep Beltran for preliminary version of the capture system. The data collection was funded in part by the National Science Foundation under Grant No. EEEC-0540865.

Citations

[1] Carnegie Mellon Motion Capture Database. http://mocap.cs.cmu.edu.

[2] S. S. Intille, K. Larson, E. Munguia Tapia, J. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson, "Using a live-in laboratory for ubiquitous computing research," in Proceedings of PERVASIVE 2006.

[3] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado and P. Beltran. *CMU-tech. report CMU-RI-TR-08-22*. April 2008.