Automatic Synchronization of Wearable Sensors and Video-Cameras for Ground Truth Annotation – A Practical Approach

Thomas Plötz[1], Chen Chen[2], Nils Y. Hammerla[1], Gregory D. Abowd[2]

[1] Culture Lab, School of Computing Science
Newcastle University, Newcastle upon Tyne, UK
{*thomas.ploetz,nils.hammerla*}@*newcastle.ac.uk*

[2] School of Interactive Computing
Georgia Institute of Technology, Atlanta, USA
{*cchen85,abowd*}@*gatech.edu*

Abstract

The common practice of manual synchronization of body-worn, logging accelerometers and video cameras is impractical for integration into everyday practice for applications such as real-world behavior analysis. We significantly extend an existing technique for automatic cross-modal synchronization and evaluate its performance in a realistic experimental setting. Distinctive gestures, captured by a camera, are matched with recorded acceleration signal(s) using cross-correlation based time-delay estimation. PCA-based data pre-processing makes the procedure robust against orientation mismatches between the marking gesture and the camera plane. We evaluated five different marker gestures and report very promising results for actual use.

1. Introduction

Similar to any other domain that involves sensor data analysis, wearable computing typically requires ground truth annotations for sample data at certain stages of the development process. Manual annotation of data recorded by, e.g., body-worn accelerometers, is usually impractical just by analyzing the raw sensor readings.

Common practice for such annotation tasks is to videotape the particular recording sessions and then to let the human annotators label the sensor data based on watching the video footage. For frame-precise annotation all sensors and cameras need to share a common temporal basis. The most straightforward approach to synchronization consists of physically connecting all sensors and cameras to the same computer, which would provide a “world clock”. However, this approach is not feasible for logging devices, i.e., sensors and cameras that store the recorded data locally without immediately transmitting them to some central entity. In this case differences in timestamps of the wearable devices and the video cameras are inevitable.

For the development of a system for automatic behavior analysis of individuals with developmental disabilities based on limb-attached acceleration loggers, we were seeking for a practical solution to the aforementioned synchronization problem. Practicality here refers to a procedure that is simple for lay users, i.e., practitioners who use the sensing system in their everyday research or therapeutic practice. At the same time the synchronization of accelerometers and video cameras needs to be accurate and reliable since it is the basis for adjustments of patient treatments.

Based on a literature survey we developed and implemented a synchronization procedure where wearers of accelerometers perform certain marker gestures in front of the camera. The synchronization method automatically detects these marker sequences in both video and accelerometer data and calculates normalized cross-correlation values, which are the basis for the estimation of inter-modality offsets and their confidence scores. We evaluated the synchronization procedure regarding its feasibility for practitioners, i.e., non-experts in wearable computing, and regarding its accuracy of time-delay estimation. The effectiveness of five different marker gestures was evaluated in a multi-person user study where we found that simple waving gestures are most convenient for users of the system whilst providing best synchronization accuracy. Furthermore, we evaluated our synchronization procedure with respect to its robustness regarding orientation changes of the gesturing person in front of the observing camera. The procedure is very reliable up to an effective angle of ±40 degrees between main orientation of the marker gesture and the camera plane.

This work has in part been funded by the RCUK Research Hub on Social Inclusion through the Digital Economy (SiDE).

1550-4816/12 $26.00 © 2012 IEEE
DOI 10.1109/ISWC.2012.15

2. Motivation for Cross-Modal Synchronization

We are interested in a reliable and easy to use automatic synchronization procedure for the following use-case. Psychologists of a behavior clinic need to assess the severity and frequency of problem behaviors of their patients. The patients wear bracelet-mounted accelerometers that log behavior data for about one week. The automatic analysis system needs to be personalized, i.e., the underlying activity recognizer needs to be adapted to the patients' idiosyncratic behavior. Therefore, functional assessments [4] are regularly performed in the clinic where an examiner interacts with the patient and any occurring problem behavior is manually labeled based on surveillance video footage. This ground truth annotation is the basis for adapting and validating the behavior analysis system.

Functional assessments are conducted by clinic staff members who typically are non-experts with respect to the technical details of the analysis system. For every session the recording system needs to be synchronized. Hence, the procedure should be as simple and reliable as possible.

2.1. Background

A number of applications of cross-modal sensor combination have been described with relation to the wearable computing community, which by definition rely on synchronization. For example, inertial measurement units (IMU) are used to support computer vision for tracking [1, 7], or for localization and identification of objects or persons [8].

Given the evident need for synchronization across modalities we found it surprising that there are only a few publications that describe *automatic* approaches. Manual synchronization is the pre-dominant approach, which is undesirable for reasons already explained.

The general idea for either manual or automatic synchronization lies in capturing certain specific activities — marker sequences— by all modalities involved. Temporal alignment would then be achieved by shifting the particular signals such that the markers share a common temporal anchor point. To some extent this idea was first exploited for pairing two mobile phones in an authorization scenario by means of shaking both devices while holding them together and analyzing the recorded accelerometer data [5]. Two phones are considered being paired if both recorded marker sequences matched without shifting.

Automatic synchronization techniques are typically either computationally demanding, or require cumbersome or laborious procedures. For example, Gu and Tomasi introduced a theoretical framework where the phase disparity between two data streams is explicitly modeled as a so-called Ornstein-Uhlenbeck random process, which is then used to calculate a discriminative measure of synchrony [3]. The computational complexity of the procedure is substantial and it is not clear how well the approach generalizes beyond the reported preliminary results of audio-visual synchronization. Teixeira *et al.* matched walking trajectories of persons captured by CCTV with their mobile phones' accelerometers using hidden Markov models [8]. Again the procedure is computationally very demanding and requires rather long trajectories for the analysis.

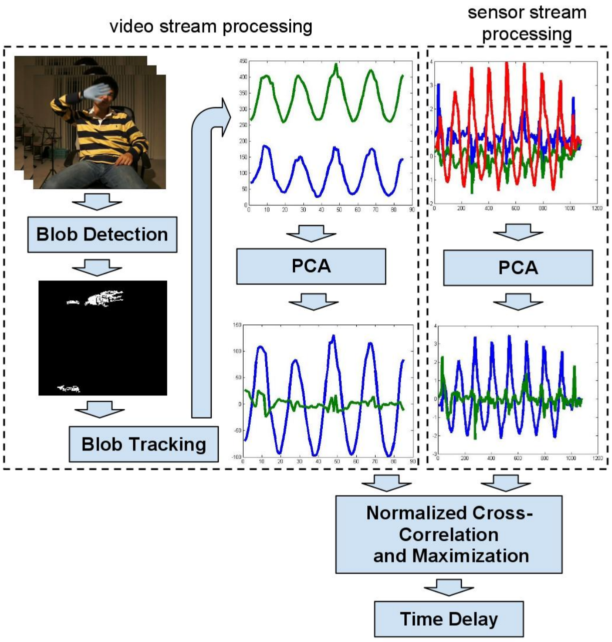

Figure 1. Automatic synchronization of video and accelerometer data – system overview

2.2. Automatic Synchronization

Our synchronization approach is inspired by a technique for automatically identifying an accelerometer-equipped moving object (such as a phone) in CCTV footage [6], which we significantly improve. The moving objects and the camera have separate, un-synchronized clocks and the identification approach utilizes estimates of time-delays between actual object accelerations and those calculated from visual tracking. By maximizing the cross-correlation between such signals, each moving object can be paired to the most likely visually-tracked blob.

We enhanced this identification approach for our synchronization procedure, summarized in Fig. 1. Before attaching the sensor bracelets to a patient, the examiner holds them in her hand and gestures in front of the camera. After the recording session, the examiner roughly indicates the beginning and ending of this gesture marker sequence in the video and then runs the synchronization procedure. The procedure aligns actual acceleration signals, $\mathbf{a}$, recorded by the limb-worn sensing platform and sub-sampled to match the camera's sampling rate, with those calculated from moving objects that are tracked within video footage, $(\mathbf{v}_b)$.

For the latter we first extract all moving blobs from the beginning of the video using Gaussian Mixture Models. By means of a CamShift tracker [2], these blobs are then tracked over the course of the video and acceleration signals are calculated as the first derivative of blob velocities.

For signal comparison, the orientations of both recording devices (camera and accelerometer) need to share some common basis, i.e., they have to be either identical or at least transformable into each other. While the orientation of the static camera is known, this is not the case for the limb-worn sensors as they only contain accelerometers (to maximize battery lifetime). In order to compensate for the unknown orientation of the accelerometer, and thus for its potential mismatch with the camera plane, we use a principle component analysis (PCA) technique. For each modality, we estimate PCA transformations (Φ and Ψ) and project the particular raw data onto 2D subspaces, which maximizes the variance of the movements resulting in transformed samples $\hat{\mathbf{v}}_b(i) \in \mathbb{R}^2$ and $\hat{\mathbf{a}}(j) \in \mathbb{R}^2$:

$$\hat{\mathbf{v}}_b(i) = \Phi^T(\mathbf{v}_b(i) - \boldsymbol{\mu}^v) \;\; b = 1 \dots B\,,\, i = 1 \dots N^{v_b} \quad (1)$$

$$\hat{\mathbf{a}}(j) = \Psi^T(\mathbf{a}(j) - \boldsymbol{\mu}^a) \;\; j = 1 \dots N^a, \quad (2)$$

with $N^{\{\mathbf{v}_b,\mathbf{a}\}}$ denoting signal lengths and B being the number of blobs. That way the acceleration data are projected onto a plane and thus in principle become comparable to 2D camera data, which is in contrast to the standard approach of processing magnitudes. For the latter one would need to estimate the direction of gravity to make both signal types comparable. By projecting onto 2D PCA-feature spaces we practically eliminate the gravity component. We proceed with two-dimensional data as this is, according to our experience, more robust than processing magnitudes only.

In order to compensate for changing sensor orientations over the course of the marker gesture, the transformations are continuously updated on a sliding window containing n consecutive samples, where n is typically the length of the performed gestures as it has roughly been marked by the human annotator. The mean vectors $\boldsymbol{\mu}^v$ and $\boldsymbol{\mu}^a$ are calculated over these sliding windows.

For every time step t, we calculate normalized cross-correlation (NCC) values $\mathcal{N}_t(b)$ between all generated, PCA-transformed blob acceleration signals $\hat{\mathbf{v}}_b$, and the PCA-transformed acceleration signal $\hat{\mathbf{a}}$:

$$\mathcal{N}_t(b) = \frac{\sum_{n=0}^{N_{\text{ws}-1}} \hat{\mathbf{a}}(n)\hat{\mathbf{v}}_b(n)}{\sqrt{\sum_{n=0}^{N_{\text{ws}-1}} \hat{\mathbf{a}}(n)^2}\sqrt{\sum_{n=0}^{N_{\text{ws}-1}} \hat{\mathbf{v}}_b(n)^2}}, \quad (3)$$

where $N_{\text{ws}} = \min(N^{\mathbf{v}_b}, N^{\mathbf{a}})$. Identification of the corresponding pair of blob-trajectory and actual accelerometer signal is then conducted by maximization of NCC values:

$$\left(\hat{b}, \hat{t}_{\hat{b}}\right) = \arg\max_{b,t_b} \mathcal{N}_t(b). \quad (4)$$

This maximization not only identifies the blob trajectory $\hat{b}$ that corresponds to the actual acceleration signal, but also unveils its temporal offset $\hat{t}_{\hat{b}}$, which is used for synchronization by shifting the signals towards matching.

	avg. abs. error	avg. abs. dev.
gesture	[frames]	[frames]
vertical	0.6	0.48
horizontal	2.2	2.32
diagonal	3.0	1.6
circular	12.8	11.36
wave-like	28.6	24.56

Table 1. Gesture-dependent time-delay estimation (averaged over 5 participants each).

3. Experiments

Two main assessment criteria are relevant for our synchronization approach: (i) the accuracy of automatic time-delay estimation across modality boundaries; and (ii) the practicality of the procedure for clinical scenarios where lay users would perform the synchronization. We evaluated these criteria by means of practical experiments where participants wore our bracelet-mounted sensing system and gestured in front of a camera. The experimental setup is intended to match the routine in the behavior clinic, so we record continuously, resulting in realistic data streams that contain marker sequences in addition to all other movements as they are performed in such settings. To stabilize visual tracking results, and to match practices we have observed in a clinical setting, participants wore examination gloves.

We measure the effectiveness of our synchronization approach as the number of frames the automatic time-delay estimation differs from manual ground truth synchronization. We pursued two kinds of experiments. First, five participants performed five different hand gestures for synchronization (Fig. 2): (i) vertical movement; (ii) horizontal movement; (iii) diagonal; (iv) circular; and (v) wave-like movements. Apart from asking the participants to gesture roughly in parallel to the camera, no further constraints were given. Average accuracy values for all five gestures are given in Tab. 1. The automatic synchronization works reasonably well for the less-complex gestures (vertical, horizontal, and diagonal back and forth hand movements). In contrast, substantial offset errors occur between predictions and ground truth for both circular and wave-like gestures.

Additionally, we evaluated the accuracy of our procedure as the orientation of the marker gesture was intentionally changed relative to the camera plane. Ideally, the marker-gestures should be performed parallel to the camera plane because then the visual observation of the movement would maximally match the trajectory recorded using the accelerometer. However, in real-world scenarios, we would expect some error from the practitioners. They would per-

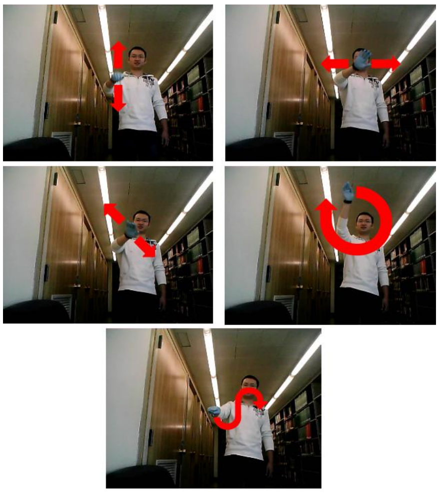

Figure 2. Gestures performed for evaluation of the automatic synchronization technique.

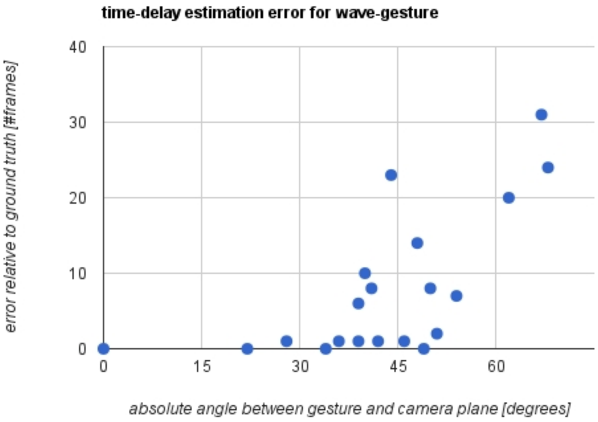

Figure 3. Effect of orientation change of horizontal gesture on time-delay estimation.

form marker gestures for synchronization before an actual examination task and often do not have the time for carefully adjusting themselves or simply are not aware of this technical constraint. Note that the PCA-based sample transformation effectively solves the dimension mapping problem. There is, however, no way of overcoming the missing data problem if a gesture is hardly observable because of its adverse orientation with respect to the camera.

One of our participants performed one of the less complex marker gestures (horizontal movement) in different orientations in front of the camera. We expected the synchronization procedure to suffer the most from orientation mismatches when performing horizontal movements, which is why that marker gesture was chosen for this experiment. The subject gestured 20 times while gradually rotating his whole body counter-clockwise after each back-and-forth gesture. Based on the footage from an overhead (bird's eye) camera we estimated the main orientation of each back and forth gesture with respect to the camera plane. For every back-and-forth gesture, we ran the automatic synchronization and compared the prediction to manual ground truth annotation. Fig. 3 shows time-delay estimation errors in relation to the effective angle between gesture orientation and camera plane. Up to an angle of ±40 degrees very reasonable results can be achieved (max. 1 frame offset).

4. Summary

We have developed an automatic procedure for the synchronization of wearable accelerometers and video cameras. The approach provides frame-precise estimations of time delays with minimal user interaction.

The procedure will be extended in at least two ways. We will integrate further modalities, most importantly audio, and generalize the approach to synchronization of multiple cameras and microphones. For the former we will investigate ways of producing time-varying audio signals, like waving a rattle, which do not require substantial additional efforts for the practitioner but allow for precise and reliable synchronization. Automatic selection of the camera to be used for synchronization will increase the practical value of the approach for multi-camera setups. We are currently working on integrating the system into the open source annotation software ELAN to make it publicly available [9].

References

[1] M. Aron, G. Simon, and M.-O. Berger. Use of inertial sensors to support video tracking. *Computer Animation and Virtual Worlds*, 18(1), 2007.

[2] G. R. Bradski. Computer Vision Face Tracking For Use in a Perceptual User Interface. *Intel Techn. Journal*, 2, 1998.

[3] S. Gu and C. Tomasi. Phase diffusion for the synchronization of heterogenous sensor streams. In *Proc. ICASSP*, 2009.

[4] B. A. Iwata and A. S. Worsdell. Implications of Functional Analysis Methodology for the Design of Intervention Programs. *Exceptionality*, 13(1):25–34, 2005.

[5] R. Mayrhofer and H. Gellersen. Shake Well Before Use: Intuitive and Secure Pairing of Mobile Devices. *IEEE Transactions on Mobile Computing*, 8(6):792–806, 2009.

[6] O. Shigeta, S. Kagami, and K. Hashimoto. Identifying a moving object with an accelerometer in a camera view. In *Proc. Int. Conf. Intell. Robots and Systems*, 2008.

[7] Y. Tao, H. Hu, and H. Zhou. Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation. *Journal of Robotics Research*, 26(6):607–624, 2007.

[8] T. Teixeira, D. Jung, and A. Savvides. Tasking Networked CCTV Cameras and Mobile Phones to Identify and Localize Multiple People. In *Proc. UbiComp*, 2010.

[9] P. Wittenburg, H. Brugman, A. Russel, and H. Sloetjes. ELAN: a Professional Framework for Multimodality Research. In *Proc. Int. Conf. Lang. Resources and Eval.*, 2006.