MINT.tools: Tools and Adaptors Supporting Acquisition, Annotation and Analysis of Multimodal Corpora

Spyros Kousidis[1], *Thies Pfeiffer*[2], *David Schlangen*[1]

[1]Dialogue Systems Group, Bielefeld University, Germany
[2]Artificial Intelligence Group, Bielefeld University, Germany
spyros.kousidis@uni-bielefeld.de

Abstract

This paper presents a collection of tools (and adaptors for existing tools) that we have recently developed, which support acquisition, annotation and analysis of multimodal corpora. For acquisition, an extensible architecture is offered that integrates various sensors, based on existing connectors (e.g. for motion capturing via VICON, or ART) and on connectors we contribute (for motion tracking via Microsoft Kinect as well as eye tracking via Seeingmachines FaceLAB 5). The architecture provides live visualisation of the multimodal data in a unified virtual reality (VR) view (using Fraunhofer *Instant Reality*) for control during recordings, and enables recording of synchronised streams. For annotation, we provide a connection between the annotation tool ELAN (MPI Nijmegen) and the VR visualisation. For analysis, we provide routines in the programming language Python that read in and manipulate (aggregate, transform, plot, analyse) the sensor data, as well as text annotation formats (Praat TextGrids). Use of this toolset in multimodal studies proved to be efficient and effective, as we discuss. We make the collection available as open source for use by other researchers.
Index Terms: Multimodal Corpora, Analysis Tools, Virtual Reality

1. Introduction

Acquisition of high quality, rich multimodal corpora of human-human or human-computer interaction is desirable for the study of interactive behaviour and communicative acts across several modalities, as well as for modelling embodied conversational agents (ECAs) geared towards natural interaction with human users [1]. However, collection and analysis of such data is challenging, as there are several criteria to be met: *Technical and resource challenges* include sufficient coverage of a given setting with multiple and diverse sensors, such as audio-visual (AV), motion capture [2], eye tracking [3], 3D scanners [4], time-of-flight cameras [5] and physiological sensors [6]; the overhead of appropriately managing these sources; and the high cost of manual segmentation and annotation of data, that often prohibits large data sets [7]. Ensuring *data quality* typically requires sophisticated, expensive equipment, frequently accompanied by proprietary software that has to be learned and poses file format and other compatibility problems. *Naturalness* of the content, which is highly desirable for studying human behaviour, is challenged by the presence and type of sensors [8], which also typically impose confinement to the laboratory or other secure indoor locations [9]. Finally, *re-usability* requires appropriate planning and additional resources to make a multimodal corpus suitable for different studies [7]. The availability of acquired data in compatible formats is an important issue.

In addition, analysis of multimodal data introduces further challenges: *synchronisation* of sensory information, specifically when sources are diverse in nature and recorded with different sampling rates; *aggregation* of these data in a single database, facilitating querying and retrieving information from the various sources, specifying regions in time and space [10]; simultaneous *visualisation* of multimodal data for the purposes of manual annotation, simulation or presentation [11]; and, finally *augmentation* of data coming from separate sources, which is currently possible for labelled segments [12], in order to derive spatio-temporal relationships among moving agents or objects (e.g. mutual gaze).

Existing approaches to multimodal corpus acquisition and analysis attempt to address the above issues in different ways: For example, synchronisation among sensors can be accomplished by recording a "clap" event visible/audible by all sensors [5], or a time server shared by all sensor workstations [6]. [13] used the audio from a turn-table that was artificially scratched to synchronise AV and motion capture sensors. On the analysis side, several approaches [2, 5, 12, 14] used widely adopted tools such as Praat [15] and Anvil [11], while others developed their own custom annotation and analysis tools [6, 16, 17]. Our methodology, which we present here, follows the paradigm of re-using widely tested freely available tools, using custom modules to interface them with each other.

We will present these modules and the general architecture in the next sections, separated into tools for recording and tools for annotation and analysis, using the concrete setup in our "multimodal interaction lab" (mintLab) for illustration.[1] An overview of this toolset is given in Figure 1. We close by describing some studies that we have already run using these tools.

2. Recording architecture (FAME.rc)

The recording architecture is centred around the *Instant Reality* package,[2] a VR engine with several distinct components. Two of these, which are freely available for private use, are utilised for recording mutimodal data: InstantIO, a data representation framework which supports many different data types such as numbers, strings, vectors or RGB images, is used to encode all sensory information that is available in "stream" form, e.g. data from motion capture equipment; and InstantPlayer, a 3D graphics browser compatible with the standard VRML and X3D for-

[1]The mintLab is jointly run by the Phonetics and Phonology Group (Wagner *et al.*) and the Dialogue Systems Group (Schlangen *et al.*) at the Faculty of Linguistics and Literary Studies at Bielefeld University.

[2]IGD Fraunhofer, http://www.instantreality.org

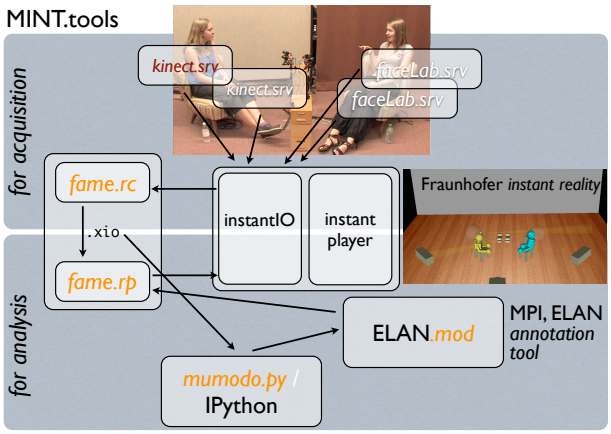

Figure 1: Overview of components of MINT.tools; our contributions labelled in italics. Top middle shows photograph of example lab setup; middle right shows corresponding VR scene, visualising motion capture and tracking of head posture, eye and gaze.

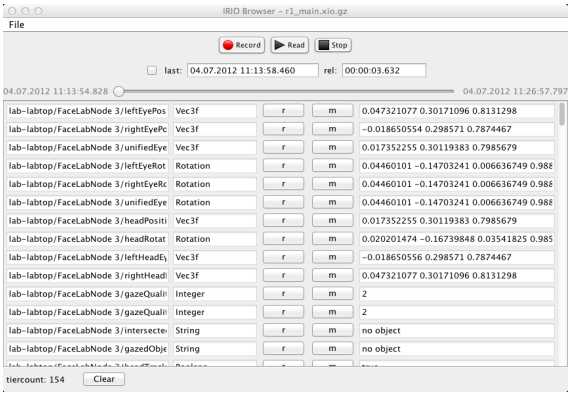

Figure 2: *Java logging and playback application.*

mats.[3] The combination of InstantIO and InstantPlayer allows real-time "pushing" of data into the VR scene (Figure 1).

Our contribution is a set of programs (we call this collection "FA^{3}ME" or short "FAME", Framework for Analysis, Annotation and Augmentation of Multimodal Experiments), that connect to sensor APIs and serve the sensor data in the format expected by InstantIO. The latter already provides support for a wide range of devices and it is easily extended to further sensors. We have added device connectors for motion capturing via Microsoft Kinect and eye tracking via Seeingmachines Facelab 5,[4] and provide example X3D scenes that visualise them (providing a skeleton for kinect data, for example, and a head and a gaze cone for the eye tracking).

This makes it possible to run each sensor and its dedicated software on separate workstations, while data is transmitted to the VR scene via LAN by means of the UDP protocol. In this way, all streamed data is collected at a single point during the actual recording, ensuring synchronisation without having to resort to time servers or any other solution. At the moment, we do not provide specific tools for synchronising audio and video to FAME.rc; in mintLab, we record video and audio on digital video tapes using synchronised cameras, linking to the FAME timecode simply by video-recording a monitor that displays it real-time. If cameras are available that expose their timecode, it should be possible to link the VR environment and the camera digitally; we have found, however, that our solutions works well for our purposes.

FAME.rc of our tool set is a Java application (Figure 2) that logs the streaming data to disk. It uses the InstantIO interface and automatically captures and logs any data (or only selected subsets of these) pushed into the scene via InstantIO. For writing to disk, an XML format is used which can be very easily parsed and is thus highly portable between applications.

[3] `http://www.w3.org/MarkUp/VRML`, and `http://www.web3d.org/x3d`, respectively.

[4] `http://www.microsoft.com/en-us/kinectforwindows/`, `http://www.seeingmachines.com/product/facelab/`, respectively.

The individual InstantIO sensors stream their data into a dynamic hierarchical tuple-space of typed data (data cloud). Adding support for new recording devices thus only requires the creation of suitable, lean InstantIO device connectors and no changes in the rest of the architecture. Adding additional devices to a recording session only requires the instantiation of one such connector per device. This makes the approach scalable to settings with many subjects or with complex spaces that require many trackers for full coverage. Since AV recording occurs in a separate circuit, the network bandwidth required by each device is minimal ($\sim$ 300 kbps), which renders network capacity a non-issue for practical purposes.

Apart from the aggregation and synchronisation benefits discussed above, the use of VR offers additional advantages: real-time *monitoring* of the recording process is facilitated by coding several input streams using the VR graphics. A straightforward example is the display of motion capture data using graphical skeletons, but further information can be displayed as well. For example, different levels of eye-tracking quality are colour-coded in the scene shown in Figure 1 (green, yellow, red). Information can also be printed in the scene in the form of text; however, the GUI of the logging application can also be used to inspect the values of data fields. We found that this live preview of the captured data reduced much of the uncertainty normally inherent in recording multiple sensor information streams, each of which may fail individually and silently. Finally, an important advantage of using VR is its utility in analysis, which is discussed in the next section.

3. Analysis (FAME.rp, mumodo)

The analysis tools that are part of MINT.tools comprise two distinct modules, the VR environment as a means of visualisation for the purpose of data annotation and augmentation, and a python package for reading in and manipulating (aggregating, transforming, plotting, analysing) recorded sensor data as well as text annotation formats (e.g. Praat TextGrids). Both of these are interfaced with the annotation tool "ELAN" [18], which has been chosen both for its functionality as well as the availability of its source code, which has made it possible for us to integrate it in the MINT.tools workflow.

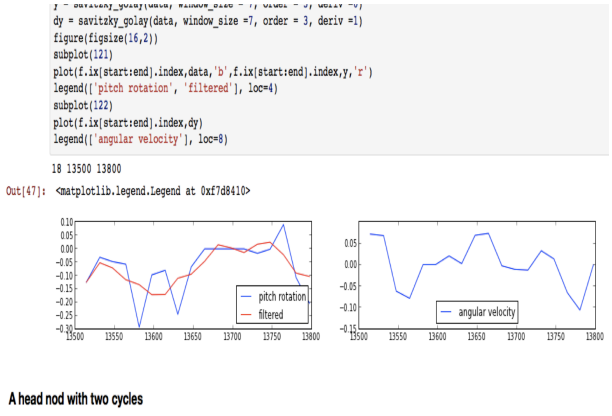

Figure 3: *IPython notebook visualising head gestures. The plots show the up-down rotation of the head (left) during a "nod' gesture, and the angular velocity (right).*

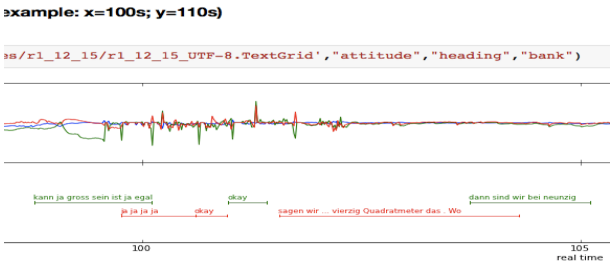

Figure 4: *Plot of tracking data and transcription text (only partly shown).*

3.1. Augmenting data using VR

Previous experience has shown the combination of VR with linguistic research to be very promising [19]. In particular, VR facilitates data *augmentation*, which allows automatic extraction of information about the interaction of modelled agents and objects. For example, human pointing behaviour has been assessed using such technology with the tool IADE [20] and, similarly, human-human interactions have been re-simulated in VR for the purpose of annotation of speech, gesture and gaze [21, 22]. Following this paradigm, we employ VR to augment data and detect events. For example, Figure 1 shows a dyadic interaction between two subjects who are being tracked by two Kinect and two Facelab eye-trackers. While the Facelab analysis environment provided by Seeingmachines only allows for the modelling of static objects for purposes of gaze intersection detection, the augmentation of motion capture data from one subject and gaze data from the other in VR enables an automatic annotation of whether either subject is gazing at the other. Such derived information can automatically be turned into symbolic annotation data (e.g., ELAN tiers), or can be made available for manual annotation.

In order to make such derived information from VR available during annotation, we have integrated the logging/playback application with ELAN, so that in effect the replay of the 3D scene acts in the same way as a video integrated into ELAN, and the 3D scene can be used as a basis for annotation. InstantIO also supports virtual sensor devices. For analysis, the intersection tests in the virtual reality scene implement a virtual InstantIO sensor and provide their results again within the FAME framework. On the data-level, there is thus a smooth transition between recorded data, manual annotations and automatic annotations, such as the mutual gaze detector, based on virtual sensors. Annotation using this toolchain offers new possibilities, as the immersive capabilities of VR (and the template scenes that we provide) literally allow the annotator to "walk around" in the scene, or even "look through the eyes" of a participant.

3.2. Data analysis enviroment

The final component of MINT.tools is a package written in the programming language Python (utilising a set of extensions for scientific computing, most importantly scipy, numpy, pandas), *mumodo.py* (for *MUltiMOdal DOcuments*). It is particularly well suited for use in interactive Python environments such as the IPython notebook, an example of which is shown in Figure 3.[5]

At the moment, *mumodo.py* offers (a) utility packages for importing and converting data from the XML format of the logger into the interpreter, (b) a set of functions that allow simple manipulation and visualisation in plots (e.g. Figure 4), and (c) capabilities to remote control ELAN from the python interpreter, in order to plot raw or derived data and simultaneously look at the relevant episode in the corpus, either on the video/audio or in the VR scene itself. We are currently developing a formal corpus management component that reads in corpus metadata and abstracts away from the way multimodal information is represented in the file system.

Our aim is to provide a set of functions to cover the functionality most often required in analyses of typical types of data such as annotated intervals or streams of timestamped events. However, the real power of this approach is the flexibility and power provided by the use of a full programming language (Python) which facilitates unlimited control over the data and the analysis functions. This requires basic programming skills on the part of the experimenter/analyst, which is however the case also for other popular solutions, such as Matlab or R.

4. Experiences so far

We have so far recorded four collections of multimodal data using the toolset described here. The first collection was performed as part of an evaluation of the recording architecture [23], and featured a listening task designed to elicit feedback head gestures. The latter were manually annotated, with the annotations validated by the tracking data.

The second collection, the *Dream Apartment Corpus (DAP)* contains 9 dyads engaging in spontaneous interaction, in the setting shown in Figure 1. The subjects were given the task to design a luxurious apartment in which they should co-habit, given a substantial amount of money was available to them. This corpus is rich in spontaneous speech, head and manual gestures. An analysis of 3000 communicative head gestures has already been performed and is published elsewhere [24].

The third collection features an almost identical setting but with different tasks and the addition of another, quite novel, sensor, namely breathing belts. The idea here is to study the breathing patterns of the subjects during different points in the

[5] An online read-only version of this notebook can be found on http://www.dsg-bielefeld.de/mint/mumodo_demo.html

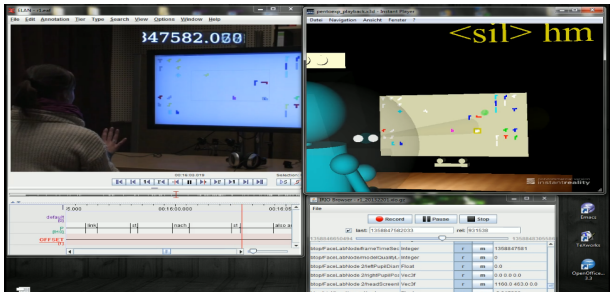

Figure 5: *Augmenting gaze data with moving objects in the PUT corpus.*

interaction, such as engaged periods, floor changes or attempts to take to the floor, intra-sentence pauses etc. In the first task, pairs of subjects are left alone in the lab and begin talking to each other by themselves after short periods of time. The second task is a negotiation similar to the DAP corpus, but involves naming five ingredients in order to (hypothetically) market a high-selling pizza. Analysis on this corpus is still in progress.

Finally, the fourth collection is a Wizard-of-Oz experiment aimed at modelling reference resolution using speech and gesture. Subjects are asked to achieve a target configuration on a Pentomino board (Figure 5, by giving verbal instructions but also pointing on the screen. The actions are performed by a confederate (wizard) who poses as the intelligent agent. The wizard has access only to ASR text and a camera view as shown on the left in (Figure 5, while it is possible to provide feedback to the participant or ask for clarifications using a TTS voice. During data collection, the board state which includes the positions of all objects at any given time (objects are dragged across the screen with the mouse) is also logged in the VR scene together with the tracking data. The same is true for the TTS and ASR texts. In the currently ongoing analysis, data from the pento board and the gaze tracker are augmented to determine which object is being looked at, at any given time.

It turned out to be very straightforward to adapt the lab resources to these different settings. Most parts of the architecture proved to be domain-independent, as was hoped. The only things that needed to be adapted to each recording situation were the scene descriptions (x3d files) used in the visualisation, as they need to represent the actual physical layout of the recording situation (mainly, the position of the sensors). But even within these scenes, the modularly structured template scenes provided guidance on what needed to adapt and what could stay constant. This was, most importantly, the definition of how the sensor data / InstantIO interfaces with scene elements such as the skeletons. With experience gained through each iteration of the architecture, adapting to novel settings becomes increasingly effortless.

5. Conclusions

We have presented a multimodal corpus collection, management and analysis methodology that seeks to answer the challenges that all such endeavours face. The main advantages are the modularity and scalability that make the toolset useful in a variety of experimental settings. Corpora collected using the methods described here were presented in order to showcase the analyses made possible by the existing available functionality. The different software modules are continuously being developed; releases can be found at the following url: `http://dsg-bielefeld.de/mint/`.

6. Acknowledgements

This research was partly supported by the Deutsche Forschungsgemeinschaft (DFG) in the CRC 673 "Alignment in Communication". The authors would like to thank Felix Hülsmann, Florian Hofmann, Michael Bartholdt, Katharina Jettka, and Jens Eckmeier, for implementing and testing most of the infrastructure presented here.

7. References

[1] Martin, J.-C and Devillers, L., "A Multimodal Corpus Approach for the Study of Spontaneous Emotions," in Affective Information Processing, J. Tao and T. Tan, Eds., Springer, 2009, pp. 267-291.

[2] Swift, M., Ferguson, G., Galescu, L., Chu, Y., Harman, C. , Jung, H., Perera, I., Song, Y. C., Allen, J. and Kautz H., "A multimodal corpus for integrated language and action," in Proc. of the Int. Workshop on MultiModal Corpora for Machine Learning, Beijing, China, 2012.

[3] Jokinen, K., Nishida, M. and Yamamoto, S., "Eye-gaze experiments for conversation monitoring," in Proceedings of the 3rd International Universal Communication Symposium, Tokyo, Japan, 2009.

[4] Fanelli, G., Gall, J., Romsdorfer, H., Weise, T. and Van Gool, L., "3D vision technology for capturing multimodal corpora: chances and challenges," in LREC Workshop on Multimodal Corpora, Valletta, Malta, 2010.

[5] Ntalampiras, S., Arsic, D., Stormer, A., Ganchev, T., Potamitis, I., and Fakotakis, N., "Prometheus database: a multimodal corpus for research on modeling and interpreting human behavior," in 16th International Conference on Digital Signal Processing, 2009 Santorini, Greece., 2009, pp. 1-8.

[6] Sumi, Y., Yano, M., and Nishida, T., "Analysis environment of conversational structure with nonverbal multimodal data," in International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, Beijing, China, 2010, p. 44.

[7] Knight, D., "The future of multimodal corpora," Revista Brasileira de Lingistica Aplicada, vol. 11, 2011, pp. 391-415.

[8] Cullen, C., Vaughan, B., and Kousidis, S., "Emotional Speech Corpus Construction, Annotation and Distribution," in Proceedings of the Language Resources and Evaluation Conference, Marrakech (Morocco), 2008.

[9] Oertel, C., Cummins, F. , Edlund, J. , Wagner, P. and Campbell, N., "D64: A corpus of richly recorded conversational interaction," Journal on Multimodal User Interfaces, 2010, pp. 1-10.

[10] Menke, P. and Cimiano, P., "Towards an ontology of categories for multimodal annotation," in Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR, 2012, p. 49.

[11] Kipp, M., "Spatiotemporal coding in ANVIL," in Proceedings of the 6th international conference on Language Resources and Evaluation (LREC-08), Marrakech, Morocco, 2008.

[12] Blache, P., Bertrand, R., and Ferré, G., "Creating and exploiting multimodal annotated corpora: the ToMA project," Multimodal corpora, pp. 38-53, 2009.

[13] Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S. , and House, D. , "Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture," in Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 2010, pp. 2992-2995.

[14] Butko, T., Nadeu, C. and Moreno, A., "A multilingual corpus for rich audio-visual scene description in a meeting-room environment," in ICMI workshop on multimodal corpora for machine learning: Taking Stock and Roadmapping the Future, Alicante, Spain, 2011.

[15] Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.42, retrieved 2 March 2013 from http://www.praat.org/

[16] Knight, D., Tennent, P. , Adolphs, S. and Carter, R. , "Developing heterogeneous corpora using the Digital Replay System (DRS)," in Language Resources Evaluation Conference Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, Valletta, Malta, 2010.

[17] Campbell, N., "Tools and resources for visualising conversational-speech interaction," Multimodal corpora, pp. 176-188, 2009.

[18] Lausberg, H., and Sloetjes, H., "Coding gestural behavior with the NEUROGES-ELAN system," Behavior research methods, vol. 41, pp. 841-849, 2009.

[19] Pfeiffer, T., "Using virtual reality technology in linguistic research," in IEEE Virtual Reality 2012, pp. 83-84, 2012.

[20] Pfeiffer, T., Kranstedt, A. and Lücking, L., "Sprach-Gestik Experimente mit IADE , dem Interactive Augmented Data Explorer," in Dritter Workshop Virtuelle und Erweiterte Realität der GI Fachgruppe VRAR, 2006, pp. 61–72.

[21] Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H. and Wachsmuth, I., "Deictic object reference in task-oriented dialogue," in Situated Communication, ed Berlin: Mouton de Gruyter, 2006, pp. 155–207.

[22] Pfeiffer, T., "Understanding multimodal deixis with gaze and gesture in conversational interfaces," PhD, Bielefeld, Bielefeld, 2010.

[23] Kousidis, S., Pfeiffer, T., Malisz, Z., Wagner, P. and Schlangen, D., "Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data", in the Interdisciplinary Workshop of Feedback Behaviors in Dialogue, Stevenson, WA, USA, 2012.

[24] Kousidis, S., Malisz, Z., Wagner, P. and Schlangen, D., "Exploring Annotation of Head Gesture Forms in Spontaneous Human Interaction", to be presented in the Tilburg Gesture Research Meeting (TiGer 2013), Tilburg, Netherlands, 2013.