Multimodal Multisensor Activity Annotation Tool

Michael Barz
German Research Center for
Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, 66123
Saarbruecken
michael.barz@dfki.de

Markus Weber
German Research Center for
Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, 66123
Saarbruecken
markus.weber@dfki.de

Mohammad Mehdi Moniri
German Research Center for
Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, 66123
Saarbruecken
moniri@dfki.de

Daniel Sonntag
German Research Center for
Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, 66123
Saarbruecken
sonntag@dfki.de

Abstract

In this paper we describe a multimodal-multisensor annotation tool for physiological computing; for example mobile gesture-based interaction devices or health monitoring devices can be connected. It should be used as an expert authoring tool to annotate multiple video-based sensor streams for domain-specific activities. Resulting datasets can be used as supervised datasets for new machine learning tasks. Our tool provides connectors to commercially available sensor systems (e.g., Intel RealSense F200 3D camera, Leap Motion, and Myo) and a graphical user interface for annotation.

Author Keywords

multimodal; multisensor; data capture; data annotation

ACM Classification Keywords

H.5.m. [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous

Introduction

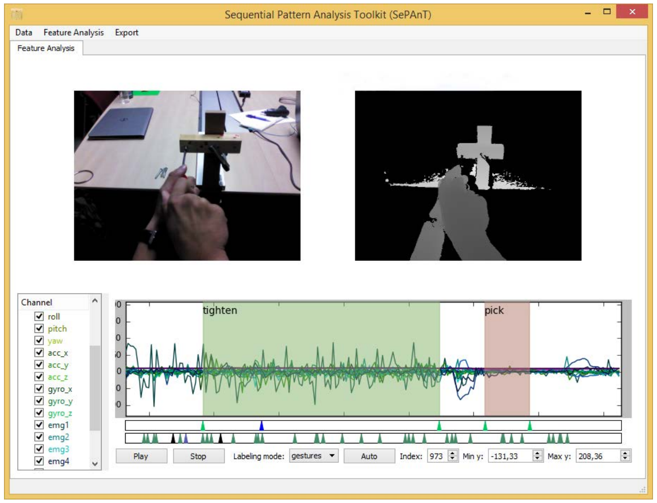
Humans learn models of the three dimensional world and a native model of physics at a very young age. Today, one can embed such knowledge into intelligent user interface applications as wearable sensors and systems are becoming ubiquitous. Including physiological data should enhance humans and computers in such a way that the

Ubicomp/ISWC'16 Adjunct, September 12-16, 2016, Heidelberg, Germany
ACM 978-1-4503-4462-3/16/09.
http://dx.doi.org/10.1145/2968219.2971459

resulting interactive experience is improved. Resulting interfaces can, for example, contribute to industry 4.0 settings [6] or to smart environments in health care [5].

However, available tools for data acquisition and annotation are mostly limited to a particular sensor. The *Video Image Annotation Tool* [1] allows to manually annotate regions of MPEG video files in a frame by frame manner. ANVIL [2] offers a multi-layered annotation for gesture research based on 2D video input. Our tool extends the idea of video annotation by providing access to depth information and signals of body-worn sensors. Dasiopoulou et al. present an overview of the state-of-the-art in image and video annotation tools [1]. Two new directions are prominent: first, recent works take advantage of highly capable devices such as smartphones and tablets that embrace novel interaction paradigms [4]; we generalise this to multi-modal multi-sensor annotation. Second, the tool *LabelMovie*, that has been opted for videos (spatio-temporal annotation), offers crowd-sourcing and machine learning options for quality assurance or automated evaluation, respectively [3].

With this work we provide a tool that allows a user to generate multimodal supervised dataset for pervasive settings. First, it enables a user to capture multiple sensor streams at once and provides support for state-of-the-art body-worn devices and depth cameras (see Table 1). Second, it enables researchers to efficiently scan and annotate these data streams (see Figure 1). Our tool presented here accounts for the need of additional seed annotations for running machine learning based annotation tools on new multi-modal multi-sensor data, which, in most circumstances, also needs to be collected individually for domain applications.



☑ roll
☑ pitch
☑ yaw
☑ acc_x
☑ acc_y
☑ acc_z
☑ gyro_x
☑ gyro_y
☑ gyro_z
☑ emg1
☑ emg2
☑ emg3
☑ emg4

Labeling mode: gestures
Index: 973
Min y: -131,33
Max y: 208,36

Figure 1: User interface of the annotation component.

Activity Annotation Tool

Our tool enables experts to record and annotate data streams of state-of-the-art sensors in order to create high quality supervised machine learning datasets. It comprises two main components, one for capturing sensor streams and one for annotating these data. The capture component facilitates to synchronously collect data from multiple sources at once either to record and annotate them or for real-time classification. The annotation component (see Figure 1) is responsible for exploring and labelling record sessions.

Capturing of Sensor Streams

The *Device Module* is responsible for initializing and capturing data from an input device plug-in (see Figure 2). Devices and its individual parameters (e.g., framerate) are configured beforehand. The *Event Manager* collects all raw

[1] https://sourceforge.net/projects/via-tool/

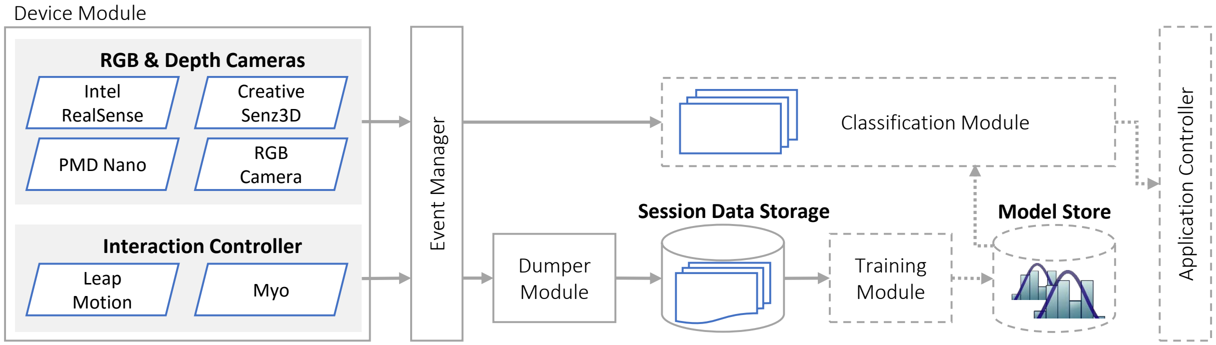


Figure 2: Software architecture of the capture component. Device modules interface the native SDKs of the individual camera or interaction controller. The Event Manager collects and synchronises incoming data streams to either write them to the disk or to analyse them in real-time.

Cameras

Device	*Output*	*Platform*
Webcam (UVC)	RGB	Windows, Linux
Intel RealSense F200	RGB, depth, point cloud	Windows
Creative Senz3D	RGB, depth	Windows, Linux
PMD Nano	RGB, depth, amplitude	Windows, Linux

Interaction Devices

Device	*Output*	*Platform*
Leap Motion	hand tracking data	Windows, Linux
Myo	IMU, EMG data	Windows

Table 1: List of supported devices, the corresponding output data and their platform compatibility.

data events triggered by active input modalities and synchronizes them according to the capture timestamps delivered by each device module. For capturing sensor streams the *Dumper Module* serializes all raw data events to the *Session Data Store*. It contains the raw data as well as annotations which will be manually added in an annotation process. The *Training Module* as well as the *Classification Module* are still target of future work, but can easily be integrated by, e.g., utilising an existing machine learning framework [3]. The *Classification Module* will share the same interface as the *Dumper Module*. Finally, the *Application Connector* will notify any connected application on classification events.

Annotation of Session Data

Annotating the raw data in the *Session Data Store* requires a suitable visualization of the sensor data. Our user interface facilitates three viewports for this purpose: a line-graph view to visualise 1D-signals and two image-views to show the 2D and 3D slice for the selected point in time.

Channels of interaction devices (1D) can be hidden or shown via selection in a separate list, for example, if accelerometer data is meaningful only. All interactions related to navigation and annotation are based around the 1D view-port: the lower navigation bar shows start- and stop-markers of annotations (green) and allows to zoom into the 1D-signal (black markers), the upper navigation bar shows the position marker (blue) and label markers for the zoomed region aligned with the 1D view-port.

Conclusion

We presented a new multimodal-multisensor annotation tool which also supports 3D data sources and additional annotation layers. The tool enables us to generate multi-channel supervised machine learning datasets for intelligent interactive systems. We already support some of the most prominent devices available on the consumer market. However, the plug-in structure of our tool makes it possible to include further devices for observation, modelling, and prediction of user behaviour (e.g., mobile eye tracking equipment or smart watches). In future work, we will integrate suitable machine learning toolkits for seamless training and semi-automatic labelling.

Acknowledgements

We thank our project partners in Kognit, http://kognit.dfki.de, funded by the Federal Ministry of Education and Research (BMBF) and CPS for smart factories (EIT Digital).

REFERENCES

1. Stamatia Dasiopoulou, Eirini Giannakidou, Georgios Litos, Polyxeni Malasioti, and Yiannis Kompatsiaris. 2011. Knowledge-driven Multimedia Information Extraction and Ontology Evolution. Springer-Verlag, Chapter A Survey of Semantic Image and Video Annotation Tools, 196–239. `http://dl.acm.org/citation.cfm?id=2001069.2001077`

2. Michael Kipp. 2008. Spatiotemporal Coding in ANVIL. In *Proceedings of the 6th international conference on Language Resources and Evaluation*. ELRA.

3. Z. Palotai, M. Lang, A. Sarkany, Z. Töser, D. Sonntag, T. Toyama, and A. Lörincz. 2014. LabelMovie: Semi-supervised machine annotation tool with quality assurance and crowd-sourcing options for videos. In *12th International Workshop on Content-Based Multimedia Indexing*. 1–4. DOI: http://dx.doi.org/10.1109/CBMI.2014.6849850

4. Klaus Schoeffmann, Marco A. Hudelist, and Jochen Huber. 2015. Video Interaction Tools: A Survey of Recent Work. *ACM Comput. Surv.* 48, 1, Article 14 (2015), 14:1–14:34 pages. DOI: http://dx.doi.org/10.1145/2808796

5. Daniel Sonntag. 2014. ERmed – Towards Medical Multimodal Cyber-Physical Environments. In *Foundations of Augmented Cognition. Advancing Human Performance and Decision-Making through Adaptive Systems*. Springer, 359–370. DOI: http://dx.doi.org/10.1007/978-3-319-07527-3_34

6. Thomas Stiefmeier, Daniel Roggen, Georg Ogris, Paul Lukowicz, and Gerhard Tröster. 2008. Wearable Activity Tracking in Car Manufacturing. *IEEE Pervasive Computing* 7, 2 (April 2008), 42–50. DOI: http://dx.doi.org/10.1109/MPRV.2008.40