A Semi-Automatic Multimodal Annotation Environment for Robot Sensor Data

Konstantinos Tsiakas
Department of CS & Engineering
University of Texas at Arlington
Arlington, TX, USA
Email: konstantinos.tsiakas@mavs.uta.edu

Theodoros Giannakopoulos
Computational Intelligence Lab
Institute of Informatics & Telecomm.,
NCSR 'Demokritos', Athens, Greece
Email: tyiannak@gmail.com

Stasinos Konstantopoulos
Software and Knowledge Engineering Lab
Institute of Informatics &
Telecomm., NCSR 'Demokritos'
Email: konstant@iit.demokritos.gr

***Abstract*—In this paper, we present RoboMAE, a multi-modal sensor data annotation environment that allows humans to concentrate on high-level decisions producing full frame-by-frame annotations. Multi-modal annotation tools focus on interpreting a scene by annotating data on separate modalities. In this work, we focus on the cross-linking of the same object's recognition across the different modalities. Our approach is based on exploiting spatio-temporal co-occurrence to link the different projections of the same object in the various supported modalities and on automatically interpolating annotations between explicitly annotated frames. The backend automations interact with the visual environment in real time, providing annotators with immediate feedback for their actions. Our approach is demonstrated and evaluated on a dataset collected for the recognition and localization of conversing humans, an important task in human-robot interaction applications. Both the annotation environment and the conversation dataset are made publicly available.**

Keywords-multimodal annotation; robotic sensors; human computer interaction

I. INTRODUCTION

Manual annotation is already a laborious, but essential, task in the development of any multimedia analysis system that attempts to assign human-interpretable labels to data; treating *multimodality* makes the annotation task harder, as the alignment of the projections of the same object in the different modalities needs to also be marked. In other words, fully annotating multimodal data requires more effort than the sum of the effort needed for the individual modalities since it is also necessary, for example, to link the speaker recognized in the audio modality with a human figures present in the visual channel.

Several general-purpose multi-modal annotation tools have been designed in the past. For example, ANVIL [1] is one of the most widely used and advanced free video annotation tools, mostly used in the area of multimodal communication research and usually focusing on the modality of speech. In [2] ANVIL has been used for creating a multimodal corpus of particular human actions. Lately [3], ANVIL has been extended by Kinect-based motion analysis procedures. In addition, VisSTA (Visualization for Situated Temporal Analysis, [4]) also focuses on natural multi-modal language annotation.

In this work, we pursued an approach towards a semi-automatic annotation tool for robot sensor data that turns the tables and makes an *advantage* out of the need to simultaneously annotate multiple modalities. We emphasize the need to both *internally represent* and *graphically visualize* the data in a manner that stresses the space and time each individual object and event occupies. In this representation, we exploit spatio-temporal coincidence in order to automatically infer initial annotations and cross-modality object correspondences. Human annotators confirm or correct the automatic annotations in any of the visualized modalities they find more convenient and the cross-modality correspondences carry these over to all modalities. To give a concrete example: if in a scene it is easier to tell who is speaking given his/her voice, then the annotator should only annotate the audio modality and let that carry over to that person's appearance in the other modalities; if in another, more noisy scene it is easier to tell who is speaking from lip movement, then the annotator should only annotate the image modality and let that carry over to that person's appearance in the other modalities.

This achieves a more judicious allocation of annotation effort allowing human annotators to concentrate on high-level decisions regarding the interpretation of a scene, while at the same time producing full frame-by-frame annotations with the same object's appearances across the different modalities cross-linked. In order to make the above more concrete, let us consider the task of scene interpretation for a robot featuring a fairly common sensor inventory: (a) *camera* for obtaining RGB images (b) a passive *stereoscopic camera* or an active *structured light sensor* for obtaining depth images (c) a *microphone* and (d) a *laser range finder* for obtaining planar range scans. Creating a unified perception from these modalities presents us with both an opportunity and a challenge: the opportunity to exploit straightforward, unambiguous recognitions in one modality in order to annotate another and the challenge of how to best represent annotations across modalities and the link between the appearances of the same real-world object in the different modalities.

There will be different levels of natural overlap that can be exploited in order to align modalities into this unified perception. Our particular mixture of modalities exemplifies all of full, partial, and no overlap. More specifically, *full overlap*, as in aligning RGB with depth data, is straight-forward since both modalities are typically recorded from sensors on the same device and are analyzed into objects that almost fully overlap in their shared frame of spatial reference. Compare, for

 ISBN: 978-1-61208-320-9

example, the RGB and depth image in the center of Figure 3. *Partial overlap* occurs in aligning the above with data from the laser range finder: range data is the planar contours of objects at a low height from the ground, typically used for obstacle avoidance. Mapping range data to the RGB-D frame (or vice versa) and looking for overlapping objects is not straightforward and often these contours are outside the field of vision of the RGB-D sensors. Compare, for example, the RGB-D images in the center with the range data in the bottom left of Figure 3: the three pairs of curves in the latter are the contours of the legs of the three people seen in the former, but at a height below what is visible in the RGB-D images. Finally, aligning data from a different space altogether, such as the *audio signal* that only has a temporal dimension and cannot be positioned at all in space. Even using microphone arrays to localize sound would only give us a rough angular position, which cannot be used to geometrically calculate spatial overlap between the sound source and the objects in the RGB-D images or range data.

In the remainder of this paper, we first present the use case and the data collection procedure (Section II) and the RoboMAE multi-modal sensor data annotation environment we have developed (Section III). We then proceed to evaluate our environment (Section IV) and conclude with discussion and future research directions (Section V).

II. Use Case and Data Collection

Our use case is the interpretation by the robot of a *human conversation* scene, an important task in any human-robot interaction application. In order to support the development and evaluation of the relevant sensor data analysis components, we envisaged a graphical tool that facilitates the following cycle:

- the different modalities are visualized simultaneously and in synchronization, including initial automatically derived annotations also presented visually
- the human annotator edits annotations in any individual modality as well as the linking across modalities
- manual edits are used to improve the automatically derived annotations

The cycle repeats until the annotator is satisfied with the quality of the annotations, so that they can be exported for training and testing the robot's recognition components.

The data has been recorded using Sek (Figure 1), a custom-made robot at NCSR 'Demokritos' that has all four sensing modalities mentioned above.RGB and depth from an Xbox 360 Kinect, audio from an Andrea microphone, laser range data from a Hokuyo 30LX laser range finder. The laser scanner is placed almost 10cm above the ground, while the height of the Kinect sensor's position is around 80cm. (For more details please see http://roboskel.iit.demokritos.gr/personnel/sek) We have made nine different recordings, with a total run-time of almost 25 min, where ten volunteers were asked to play out different conversation scenarios of varying difficulty for automatic recognition.

The recorded modalities are synchronized by global timestamps and formatted as follows: audio is 1 sec-long WAV files,

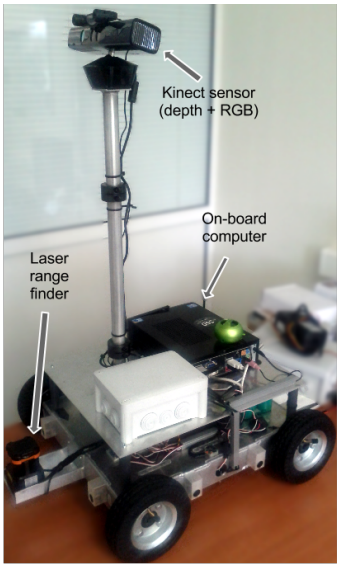

Figure 1: SEK. The robot platform used to record the data.

RBG frames are JPEG files, kinect depth frames are raw binary files, and laser scan data is in a single text file.

III. Semi-Automatic Multimodal Annotation

A. Graphical user interface and manual functionalities

The annotation tool focuses on providing a user-friendly interface for multi-modal annotation of audio, visual and laser data and a set of semi-automatic methods that will utilize the annotation process. It visualizes the different modality data recorded from the robot's different sensors. The user is able to see the data frame from each sensor at any time, as a synchronization procedure of different modality data is embedded.

In Figure 3, we present a screenshot of the implemented annotation tool. The slider control at the bottom of the GUI is used to select the time frame. The upper left display presents a 2-second window of audio, that can be played back. The annotator can zoom in and out of the display to change the size of audio window size. The upper right display is the visual modality while the bottom right display is the depth modality, visualized as gray-scale video.

Finally, the range data display on the bottom left visualizes a planar laser scan. This display can be toggled between two alternative visualizations, showing either the raw polar coordinates or their Cartesian transform.

In case the annotator performs a fully manual annotation task, the annotation can be divided in three main tasks: visual, audio and laser track–depth image mapping. The user has to annotate all frames by using the respective controls. Regarding the labeling of the annotated humans, either the default names (Speaker1, Speaker2, etc.) or any other name can be used. There are two ways to complete a face annotation task. Either by drawing bounding boxes on each frame or by using an interpolation procedure as an assisting tool. For simple cases where the positions of the face bounding boxes do not dramatically change for a particular time period, the annotator

 ISBN: 978-1-61208-320-9

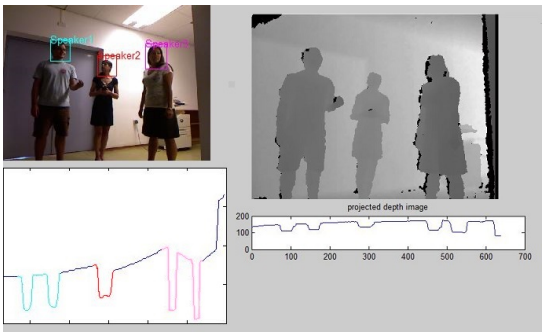

Figure 2: Depth image projection and laser scan mapping.

can use an interpolation procedure applied on each face's position over time. Specifically, the user can annotate some certain frames and use cubic interpolation to automatically annotate the intermediate frames. In this case, the annotator should check the 'Interpolation' choice in the Face Annotation panel for 'Faces', annotate some certain frames and select them as 'Interpolation Points'. After executing this method, the intermediate frames will be annotated. The accuracy of this method depends on the frames' selection.

As far as the sound annotation sub-process is concerned, the user can interactively choose and play a specific audio segment and finally match it to any of the annotated faces. In this way, audio annotation is linked to the RGB annotation described above. The annotator is able to see if a segment is annotated by checking the corresponding text box next to the audio segment figure. Moreover, the user is given the choice of a speaker color-coded view of audio segments.

Regarding the depth image information, each depth image value depicts the distance from the sensor to the specific point. Our purpose in the context of a unified multimedia annotation tool is to associate this depth image information with the laser scan output. This is achieved through a projection calculation of the bottom third of the depth image to the horizontal axis. In the sequel, taking advantage of the similarity between the laser scanner and the depth sensor projection, we define a mapping function that assigns each point of the projected depth image to the laser scan curve. The user can click either on the projection, the depth image or the laser scan curve and select an area of interest with equivalent meaning (Fig.2).

B. Automatic annotation

The safest way (in terms of annotation accuracy) to complete an annotation task is to follow a fully-manual annotation procedure. That means, for example, that the user needs to draw bounding boxes on each frame, annotate each audio segment and each laser scan plot. As this is a tedious process, apart from the manual annotation functionalities in the GUI, RoboMAE integrates recognition techniques, such as face detection, face tracking, speaker diarization, image projection.

1) Visual: Instead of defining face bounding boxes for every single frame (or simply use the interpolation procedure described before), users can employ a face detection approach based on the Viola-Jones algorithm [5]. This can be used as an initial estimate of the face positions in each frame. Apart from the automatic face detection approach, we have also used the Mean Shift algorithm for automatic *face tracking* [6]. The annotator must choose 'Tracking' in the Face Tracking Panel, choose 'Faces' and annotate speakers in a particular frame, either by drawing bounding boxes or by using the Face Detector. The user can then complete the face annotation by choosing 'Tracking frames' to point to the last frame to track and executing the tracking method. Naturally, accuracy depends on the accuracy of the individual manually annotated faces.

2) Audio: Speaker diarization partitions an audio stream into segments denoting speaker identity. In other words, a speaker diarization algorithm answers the question 'Who speaks when?' [7], [8], [9]. Most of the proposed methods on speaker diarization are only based on audio information, however there are also multimodal approaches [10], [11]. Here, we employ semi-supervised learning in order to cluster the audio segments into speakers [8]. The idea is to have the user annotate speaker identity in a small part of the audio signal and then use this information to 'guide' the semi-supervised speaker diarization algorithm. In other words, the user annotates a small number of speech segments and the semi-supervised algorithm returns a fully-annotated stream.

3) Laser track and depth image mapping: Laser scan data is currently annotated fully manually or by interpolation. The user can choose 'Laser' in the 'Face Tracking' panel and—choosing certain annotated frames—to execute the cubic interpolation method described previously.

IV. Usability and Annotation Performance

We have evaluated the *usability* of the implemented tool in terms of the time needed for identically annotating the same data using either the fully manual or semi-automatic approaches. The average annotation time was reduced by 60%, dropping from $562\,\mathrm{min}$ for the fully manual annotation to $219\,\mathrm{min}$ for the semi-automatic annotation.

In order to measure how close the initial automatic annotations were to the fully manual ones, we measured the performance of the *face tracking* and *speaker diarization* modules assuming the fully manual annotations as ground truth. In this experiment, *face tracking* achieves an $\mathrm{F}_{\beta=1}$ measure of 68% and *speaker diarization* a *cluster accuracy rate* of 74%.

V. Conclusion

We have presented RoboMAE, a visual playback and annotation editing environment for multi-modal sensor data. The major innovation in our tool is that it exploits sparse manual annotations in order to *interpolate* a complete frame-by-frame annotation and to *transfer* object recognitions across modalities. By interacting with the visual environment in real time, the backend facilitates starting out with a sparser and effortless annotation that only delves into details where necessary in order to converge to a satisfactory result. Our contribution comprises the complete MATLAB code for RoboMAE and the annotated dataset used in the experiments described here, both publicly available at http://roboskel.iit.demokritos.gr.

We are currently integrating more advanced pattern recognition methods over the laser range data [12], in order to

 ISBN: 978-1-61208-320-9

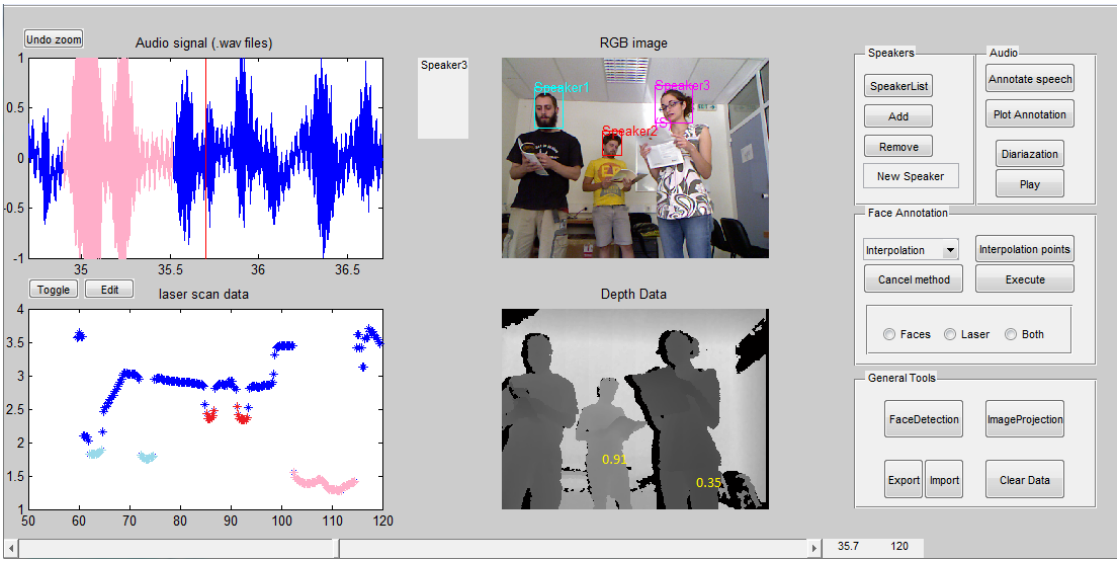

Figure 3: RoboMAE supports manual and semi-automatic techniques to help the user complete multimodal annotation tasks accurately and efficiently.

enhance the automatic annotations that currently only rely on interpolation (cf. Section III-B3). Furthermore, we are extending the heuristics used to transfer annotations across modalities, e.g. by experimenting with skeleton models extracted from depth and used to guide face tracking.

Longer-term plans include taking advantage of the experience gained by developing this first prototype to re-design the architecture of RoboMAE. The further aim is that RoboMAE is not tied to any particular sensor type and automatic recognition method, but to define generic interfaces for the recognition tools used in the back-end.

We will also develop annotation quality metrics that will assist the users decide whether the current annotations are 'good enough' for their purposes or further refinement is needed. One idea is to support a cycle where a small, 'ground truth' portion of the material is annotated in detail and checked thoroughly. As annotation over the rest of the material progresses, this is used to re-train recognition tools and test them over the ground truth, providing an indication of the quality of the current annotations in the larger portion of the data.

ACKNOWLEDGEMENTS

The work described here was partially carried out at the 2013 International Research-Centred Summer School (http://irss.iit.demokritos.gr). and in the context of *Roboskel*, the robotics activity of the Institute of Informatics and Telecommunications, NCSR 'Demokritos' (For more details please see http://roboskel.iit.demokritos.gr).

We would also like to gratefully acknowledge the participation of colleagues and IRSS students in the data collection.

REFERENCES

[1] M. Kipp, "ANVIL: The video annotation research tool," in *The Oxford Handbook of Corpus Phonology*. Oxford University Press, 2014, to appear, pre-print at http://www.anvil-software.org.

[2] M. Swift, G. Ferguson, L. Galescu, Y. Chu, C. Harman, H. Jung, I. Perera, and et al., "A multimodal corpus for integrated language and action," in *Proc. Workshop on MultiModal Corpora for Machine Learning*, 2012, held at LREC 2012, Istanbul, Turkey, 22 May 2012.

[3] M. Kipp, "Annotation facilities for the reliable analysis of human motion," in *Proc. 8th Intl Conf. on Language Resources and Evaluation (LREC), Istanbul, Turkey*, 2012, pp. 4103–4107.

[4] Y. Shi, T. Rose, and F. Quek, "A system for situated temporal analysis of multimodal communication," in *Proc. Workshop on Multimodal Corpora*, 2004, held at LREC-08, Lisbon, Portugal, 25 May 2004.

[5] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[6] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2000, pp. 142–149.

[7] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557–1565, 2006.

[8] T. Giannakopoulos and S. Petridis, "Fisher linear semi-discriminant analysis for speaker diarization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 1913–1922, 2012.

[9] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *INTERSPEECH*, vol. 5, 2005, pp. 2441–2444.

[10] G. Friedland, H. Hung, and C. Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4069–4072.

[11] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization," in *Proc. of the 10th Intl Conf. on Multimodal Interfaces*. ACM, 2008, pp. 257–264.

[12] T. Varvadoukas, I. Giotis, and S. Konstantopoulos, "Detecting human patterns in laser range data," in *Proc. 20th European Conference on Artificial Intelligence (ECAI 2012)*, 2012.

 ISBN: 978-1-61208-320-9