Multimodal Machine Learning for Social Interaction with Ageing Individuals

Louis-Philippe Morency, Sakriani Sakti, Björn W. Schuller, and Stefan Ultes

3.1 Introduction

Multimodal machine learning (MMML) is a vibrant multi-disciplinary research field. It addresses some of the original goals of artificial intelligence (AI) by integrating and modelling multiple communicative modalities, including linguistic, acoustic and visual messages amongst other multisensorial information such as physiological signals or symbolic information, e.g., concerning contextual cues. With the goal of better understanding and modelling the behaviours of ageing individuals, this research field brings some unique challenges for multimodal researchers. This comes with the heterogeneity of the potentially mixed symbolic and signal-type data and the contingency often found between modalities. In this chapter, we identify four key challenges necessary to enable multimodal machine learning, in particular, considering ageing individuals:

(1) *multimodal*, this modelling task includes multiple relevant modalities which need to be represented, aligned and fused. Often, these are asynchronous, potentially in dynamic manner and/or operating on different time-scales;

L.-P. Morency
Carnegie Mellon University, Pittsburgh, USA
e-mail: morency@cs.cmu.edu

S. Sakti
Nara Institute of Science and Technology, Ikoma, Japan
e-mail: ssakti@is.naist.jp

B. W. Schuller
Chair EIHW, University of Augsburg, Germany & GLAM, Imperial College London, Augsburg, UK
e-mail: schuller@ieee.org

S. Ultes (✉)
Mercedes-Benz AG, Sindelfingen, Germany
e-mail: stefan.ultes@daimler.com

(2) *high variability*, this modelling problem expresses high variability given the many social contexts, large space of actions and possible physical or cognitive impairment;
(3) *sparse and noisy resources*, this modelling challenge addresses unreliable sensory data and the limitation and sparseness of resources that are specific for the special user group of ageing individuals; and
(4) *concept drift*, where two types of drift were identified, namely, on the group level (as the target group of usage is not fully known at the moment of development of according interfaces given that it is yet to age) and the individual level (given that ageing may lead to drifting behaviour and interaction preferences throughout the ongoing ageing effect).

These four challenges come together when we build an evaluation plan that enables, at the same time, a strategy to include the broader machine learning community in this effort. This research agenda will enable more effective and robust modelling technologies as well as development of socially competent and culture-aware embodied conversational agents for elderly care. In the following, we will address these four challenges in the named order one-by-one. Concluding, we will summarize the discussion of these.

3.2 Multimodal

An important aspect of ageing research is to develop foundational methods to analyze, model and represent multimodal information. The unique and challenging aspect of multimodal research is the heterogeneity of the multimodal data and the challenges in integrating and interpreting such heterogeneous data coming from multiple modalities. Creating computers able to understand this multimodal data brings many fundamental problems which can be grouped into five main classes, following the taxonomy of Baltrušaitis et al. [4]: representation, alignment, fusion, translation and co-learning.

Representation: Learning how a computer can represent numerically the heterogeneous data from multiple modalities. These computational representations should be designed for both efficient modelling and better visualization. For example, a joint representation of how a person looks and sounds when they are happy will allow computers to better recognize human emotions. These joint representations will be the most efficient when they can take advantage of the natural dependencies between modalities. Another objective is to improve the interpretability of multimodal data. By identifying commonalities and differences between multimodal data, a multimodal representation provides an avenue to bridge the gap between continuous versus discrete data and numerical versus symbolic data.

Alignment: The process of establishing spatial and/or temporal connections between events across modalities. For example, when reading the caption of an image, alignment is the process where words are linked to specific objects or groups

of objects in the image. Other examples of alignment include automatic video capturing and identifying the acoustic source in a video. One challenge in alignment is dealing with data stream with different sampling rates (e.g., continuous signals versus discrete events). Alignment may require defining a similarity metric between modalities to identify the connection points. The alignment can be temporal, as when we align the audio and images of a video, or it can be spatial alignment, as when we try to morph between two face images.

Fusion: The combination of information coming from two or more sources to uncover or predict a pattern or trait of interest. Examples of multimodal fusion include multimodal emotion recognition, audiovisual speech recognition and audiovisual speaker verification. The information can be redundant which helps increase robustness, or complementary which often helps increase accuracy. The challenges are multiple since the modalities do not need to be synchronized or even have the same sampling rate. Some modalities may be incomplete with missing information. Some modalities may provide continuous streams of data, while other modalities may be intrinsically discrete providing information about given events.

Translation: The transformation or mapping from one modality into another. Examples include speech-driven animation, and text-based image retrieval. The foundational methods learn the relationship between streams of data, capturing their dependencies. The goal of translation can be generative in nature, creating a new instance in one modality given information from another modality. It can also be descriptive models, where one modality is used to increase the characterization of another modality (e.g., describing the information conveyed in an image).

Co-learning: The transfer of knowledge from one modality to help with the prediction or modelling task in another modality. Co-learning examples are more technical in nature. One of the most popular examples these days is the use of language to help generalization of computer vision algorithms, specifically for object recognition. Co-learning aims to leverage the rich information in one modality, in the learning of another modality, which may have only limited resources (e.g., small number of examples with limited annotations or noisy input). Example of co-learning algorithms include co-training, zero-shot learning and concept learning.

Problems on multimodal processing may involve combination of these categories. Importantly, crosscutting research in these areas will open opportunities to better understand and interpret multimodal data across domains, serving as instrumental tool for the community. These tools can be generic, working across problems. They can also be specific to determined problems or modalities.

3.3 High Variability

Realizing multimodal machine learning in highly variable environments adds to the challenges described previously. Variations may impact the performance of machine learning models and may be categorized by

Variation on the input Deriving suitable knowledge from observed situations itself is already quite challenging as relevant information needs to be separated from irrelevant and redundant information. In a highly variable environment, this task is even more difficult, e.g., as some pieces of information is only relevant for a small amount of cases but irrelevant for the rest. Another aspect is the increased size that is necessary to model the input in a highly variable environment.

Variation on the output For classification or decision making tasks, the number of labels or actions is directly linked to the difficulty of the learning problem. With each added variation on theses sets, the number of classes or actions increases even further and thus poses a serious challenge on the learning task as these become more fine-grained or may even overlap in some cases. Thus, it is more difficult to discriminate between them.

Both types of high variability poses crucial challenges towards realizing multimodal machine learning approaches, especially within the context of social interaction with the focus group of ageing individuals. Some factors that cause this high variability are:

Many social contexts Elderly citizens engage in many social contexts—just as the non-elderly. They do volunteering, or are active members in local clubs in addition to regular participation in social activities like meeting with friends and family. Additionally, their living conditions might at the same time include elderly care homes or other care-taking facilities resulting in a broad range of social contexts.

Physical or cognitive impairment The existence of physical or cognitive disabilities becomes more prevalent with increasing age: "Seniors are almost twice as likely to have a disability as those of working age" [13, 14]. This comes along with a unique nature of communication on both sides of the communication channel: input to a social agent may come in many different ways and social agents must use communication means that are adequate for the given situation. Thus, a social agent needs to offer a wide range of options for communication.

Reluctance to adapt Seniors are often reluctant to adapt to new situations and exhibit stubbornness due to many reasons, e.g., feeling of losing their independence, fear of losing control of their lives, feeling depressed about the deaths of spouse, friends and/or family, feeling of being left out of the family, or fear of their own mortality [1]. Social agents must thus be able to provide interaction tailored to the specific needs of many different concrete users.

While high variability is not unique to multimodal machine learning setups, the additional modalities further increase the variability and extend it to multiple communication channels, thus adding to the challenges described in Sect. 3.2.

The high variability has a very simple yet very problematic effect on machine learning models: they either are less likely to generalize well or need an increasing amount of diverse and adequate data that contains all relevant situations and information. However, a simple increase in data does not necessarily lead to well-working models. The important part here is that the factors that cause the variability are part of the input feature set of the machine learning models. While approaches that learn

to extract the relevant information like deep learning model these factors implicitly, other approaches create the need for explicit modelling.

For applications that make use of multimodal machine learning to derive knowledge from observed situations, this either results in a huge input space or in separate models, either one for each category of variation[1] or arranged hierarchically. More individualized models are required that do not follow the "one-fits-all" paradigm but are able to grasp all the relevant information on the input side, e.g., [2, 23].

Artificial agents that interact with seniors additionally are faced with the challenge of exhibiting adequate behaviour for the given context. Hence, they need to understand the social context and they need to be able to react accordingly, e.g., the non-elderly might be more forgiving if the system behaviour is socially awkward.

3.4 Sparse and Noisy Resources

In multimodal machine learning, each modality may provide complementary information and cross-modality feature learning could help the model to understand more. However, to extract generalizable features in a supervised learning approach and ensure a robust perception of the overall information, multimodal machine learning requires a large-scale labelled training dataset. Unfortunately, such data is often unavailable, especially for multimodal resources needed to construct machine learning for social interaction with ageing individuals.

Several issues make data collection in ageing individuals become critical bottlenecks.

- Although the population of ageing individuals is growing in all regions of the world, it is still minimal compared to the general population. This makes it hard to collect sufficient amounts of their social interaction in real life.
- Cognitive or physical deficits can lead to the inability to perform some assessments, which leads to incomplete or partially-observed data.
- Imperfect sensory data are often unavoidable in real-world environments, resulting in unreliable measures and noisy data.
- Data preprocessing and manual labelling or annotation are expensive and time-consuming. If we cannot bear the cost, a large amount of data will remain unlabelled.

Consequently, to handle such sparse and noisy data, other approaches beyond traditional supervised learning fashion become necessary. Extensive researches are currently focused on a learning algorithm that can be performed without the need for expensive supervision. To date, several approaches have been proposed, including transfer learning, semi-supervised learning, self-supervised learning and active learning.

[1] A requirement for separate models is that the observed variability can be divided into distinct categories.

Transfer learning The study of transfer learning is motivated by the fact that humans can intelligently apply their knowledge learned from previous problems to solve new tasks faster or with better solutions. In transfer learning, the knowledge or information of an already trained machine learning model with a sufficient labelled training dataset is reused or applied in a new task with limited resources. Therefore, instead of constructing the model from scratch using a minimal dataset, we begin with patterns learned from solving a related task. Palaskar et al. applied transfer learning for audiovisual scene-aware dialogue [16]. Specifically, they developed a hierarchical attention framework to fuse contributions from different modalities and utilized the framework to generate textual summaries from multimodal sources (i.e., videos with accompanying commentary). Wolf et al. also introduced a transfer learning approach to generative data-driven dialogue in conversational agents [26]. The framework was called TransferTransfo, which is a combination of a Transfer learning scheme and a high-capacity Transformer model.

Semi-supervised learning A semi-supervised learning model aims to make effective use of all of the available data, not just the labelled data, but also unlabelled data. The primary method is to train a model with the labelled data, then let the model label the unlabelled examples. Then, finally, retrain the model with the additional training dataset produced by the model. In a multimodal framework, Effendi et al. proposed semi-supervised learning for cross-modal data augmentation via a multimodal chain, and addressed the problems of speech-to-text, text-to-speech, text-to-image and image-to-text [7]. A study by Tseng et al. showed that semi-supervised techniques could reduce the need for intermediate-level annotations in training neural task-oriented dialogue models [21].

Self-supervised learning Self-supervised learning provides a viable solution when labelled training data is scarce. The framework is commonly performed to learn the relations or correlations between inputs, such as predicting word context or image rotation, for which the target can be computed without supervision. In spoken dialogue research, Wu et al. utilized self-supervised learning for inconsistent dialogue order detection by explicitly capturing the conversation's flow in dialogues [27]. A study by Li et al. also performed a self-supervised method for a multimodal dialogue generation model [12]. Specifically, they adopted the multi-task learning, including response language modelling, video-audio sequence modelling and caption language modelling, to learn joint representations and generate informative and fluent responses.

Active learning Active learning is one way to address the problem of a limited annotation budget by actively enlarge the training datasets. The basic idea is to ease the data collection process by automatically deciding which instances an annotator should label to train an algorithm quickly and effectively. In the context of dialogue research, Hiraoka et al. proposed active learning in creating dialogue examples [10]. Gasic and Young [9] used active learning to speed reinforcement learning of the dialog system policy. A study by Rudovic et al. also offered a multimodal active learning approach, in which, deep reinforcement learning is used to find an optimal policy for active selection of the user's data [19].

3.5 Concept Drift

In a longitudinal human-machine interaction setting, one likely changes his/her characterizing attributes [24], preferences [5], behaviours and how these are realized and portrayed. In particular, when ageing, elderly may be gradually affected by increasing mental and physical limitations or confronted with abrupt life changes in life circumstances such as by moving to care-taking facilities. This renders them a user group of likely occurrence of concept drift in continued every-day multimodal social interaction scenarios.

In machine learning, concept drift describes a non-stationary learning problem over time. Hence, it primarily refers to the learning targets the model learns to classify or predict and their change over time [22, 25]. In a classification task, this could mean that the classes and their relation to the 'input' change. Similarly, in a regression problem, the numerical target would 'drift' in the sense of change over time. In particular, this becomes challenging, if this change is unknown and it has to be detected in the first place. In any case, such a drift demands for adaptation of the learnt models over time. This could, e.g., be realized in a life-long learning process [17] or by some suited means of incremental transfer learning to adapt to the changing target based on the knowledge previously required to best classify the former 'old' target. In a broader definition, concept drift could, however, also refer to input drift, such as when the input type changes over time. In addition, as outlined, one also potentially needs to detect the presence of concept drift in the first place [11, 22].

Coming back to dealing with (multimodal) interaction, an (elderly) user may as mentioned alter her behaviour over time, hence, also enforcing a model change in time (the concept drift) [8]. For our particular task of interest—multimodal machine learning for social interaction with ageing users—we see two main types of drift:

(1) drift on the group level. This is understood in the sense of the elderly as population or sub-groups thereof, such as clustered by gender, age groups, etc., as the target group including its characterizing attributes, preferences and behaviour patterns of interaction is not fully known at the moment of development of according interfaces given that it is yet to age, and
(2) on the individual level. The motivation for this level is given by the fact that ageing as a process may lead to drifting behaviour and interaction preferences throughout the ongoing individual ageing, potentially driven, e.g., by the named potential limitations in cognitive and/or physical abilities. Note that change of preferences and, in particular, change in (elderly) user behaviour can be both gradual or abrupt, given potential sudden occurrence of named potential limitations such as given by a gradual progressing age-induced disorder or following a fall or stroke or a change in living conditions. These two kinds of concept drift—abrupt or gradual—are generally known as the two types of occurrences of concept drift [22].

While the problem of concept drift has practically not been investigated in the field of social interaction with a human–computer interface or agent even for any

user group, the methods available to deal with the problem start to find interest in this and the broader application domain. As an example, life-long learning has been considered for emotion recognition [17]. Further, for recognition of behaviour of elderly in smart homes, the authors in [28] suggest to integrate activity duration into the learning process learning to cope with concept drift. More broadly looking into human behaviour modelling in general, the author in [15] names active learning an interesting option, given that incremental learning is often chosen as a mean to handle concept drift. This could include cooperation with the elderly individuals to label new most informative data points with their aid during usage.

However, in other domains of multimodal machine learning, the problem has been handled more frequently, such as in [18]. The authors see the problem of time-variant input/output relation as dynamic optimisation problem (DOP). They compare two methods that both improve robustness in the presence of concept drift: a time-window solution to train exclusively on most recent data, and a "time-as-covariate" approach modelling time as additional input variable. Considering a number of benchmark functions, they find the former better suited in case of higher-dimensional multimodal cases, and the later in the opposite case of lower dimensional multimodal tasks, as those make it easier to co-model time as input.

Further, in the general machine learning literature, one finds manifold solutions, such as in [6], where the authors suggest semi-supervised learning as a mean to handle concept drift. In addition, instance selection, instance weighting and ensemble learning [22] are popular 'traditional' methods for the detection and adaptation [3]. In particular for the recently popular deep learning approaches, online learning methods have been proposed such as hedge back-propagation [20]. For further insights, a number of overviews exist such as [29].

As a conclusion, concept drift will be a real-world challenge for multimodal social interaction over pro-longed usage—likely in particular for an ageing elderly user group. Methods to cope with the problem do exist, but yet have to be integrated and evaluated more in this application context.

3.6 Conclusion

Multimodal machine learning itself already contains hard challenges in general. These have been described in this chapter and have been embedded into the context of social interaction with ageing individuals. On top of *representation*, *alignment*, *fusion*, *translation* and *co-learning*, we identified *high variability*, *sparse and noisy resources* and *concept drift* as the most important topics. *High variability* is described as variation on the input or the output leading to an increase in complexity of the machine learning model. *Sparse and noisy resources* further require advanced techniques like transfer learning or active learning. Especially the latter is also proposed to alleviate some of the problems arising from *concept drift* where important properties of the problem change over time, and thus require special handling on the modelling side.

References

1. Assisted senior living: dealing with stubbornness. https://www.assistedseniorliving.net/caregiving/dealing-with-stubbornness/. Accessed 02 Oct 2019
2. AVEC '19: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. Association for Computing Machinery, New York, NY, USA (2019)
3. de Barros, R.S.M., de Carvalho Santos, S.G.T.: An overview and comprehensive comparison of ensembles for concept drift. Inf. Fusion **52**, 213–244 (2019)
4. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**2, 423–443 (2018)
5. Campigotto, P., Passerini, A., Battiti, R.: Handling concept drift in preference learning for interactive decision making. HaCDAIS **2010**, 29 (2010)
6. Dyer, K.B., Polikar, R.: Semi-supervised learning in initially labeled non-stationary environments with gradual drift. In: The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1–9. IEEE (2012)
7. Effendi, J., Tjandra, A., Sakti, S., Nakamura, S.: Listening while speaking and visualizing: improving asr through multimodal chain. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) pp. 471–478 (2019)
8. Esposito, F., Basile, T.M., Di Mauro, N., Ferilli, S.: Machine learning enhancing adaptivity of multimodal mobile systems. In: Multimodal Human Computer Interaction and Pervasive Services, pp. 121–138. IGI Global (2009)
9. Gašić, M., Young, S.: Gaussian processes for POMDP-based dialogue manager optimization. IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**(1), 28–40 (2014)
10. Hiraoka, T., Neubig, G., Yoshino, K., Toda, T., Nakamura, S.: Active Learning for Example-Based Dialog Systems, chap. Dialogues with Social Robots: Enablements, Analyses, and Evaluation, pp. 67–78 (2017)
11. Klinkenberg, R., Joachims, T.: Detecting concept drift with support vector machines. In: ICML, pp. 487–494 (2000)
12. Li, Z., Li, Z., Zhang, J., Feng, Y., Niu, C., Zhou, J.: Bridging text and video: a universal multimodal transformer for video-audio scene-aware dialog. AAAI2020 DSTC8 workshop (2020)
13. Morris, S., Fawcett, G., Brisebois, L., Hughes, J.: Canadian survey on disability reports: a demographic, employment and income profile of Canadians with disabilities aged 15 years and over (2017). https://www150.statcan.gc.ca/n1/pub/89-654-x/89-654-x2018002-eng.htm. Accessed 27 May 2020
14. Murman, D.L.: The impact of age on cognition. Semin. Hear. **36**(03), 111–121 (2015). https://doi.org/10.1055/s-0035-1555115
15. Padmalatha, E., Reddy, C., Rani, P.: Mining concept drift from data streams by unsupervised learning. Int. J. Comput. Appl. **117**(15) (2015)
16. Palaskar, S., Sanabria, R., Metze, F.: Transfer learning for multimodal dialog. Comput. Speech Lang. **64**, 101093 (2020). https://doi.org/10.1016/j.csl.2020.101093
17. Ren, Z., Han, J., Cummins, N., Schuller, B.: Enhancing transferability of black-box adversarial attacks via lifelong learning for speech emotion recognition models. In: Proceedings INTERSPEECH 2020, p. 5. ISCA, ISCA, Shanghai, China (2020)
18. Richter, J., Shi, J., Chen, J.J., Rahnenführer, J., Lang, M.: Model-based optimization with concept drifts. In: Proceedings of the 2020 Genetic and Evolutionary Computation Conference, pp. 877–885 (2020)
19. Rudovic, O., Zhang, M., Schuller, B., Picard, R.W.: Multi-modal active learning from human data: a deep reinforcement learning approach (2019). https://arxiv.org/abs/1906.03098
20. Sahoo, D., Pham, Q., Lu, J., Hoi, S.C.: Online deep learning: learning deep neural networks on the fly (2017). arXiv:1711.03705
21. Tseng, B.H., Rei, M., Budzianowski, P., Turner, R., Byrne, B., Korhonen, A.: Semi-supervised bootstrapping of dialogue state trackers for task-oriented modelling. In: Proceedings of the 2019

Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1273–1278. Hong Kong, China (2019)
22. Tsymbal, A.: The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin **106**(2), 58 (2004)
23. Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., André, E.: The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 831–834. ACM (2013)
24. Webb, G.I., Pazzani, M.J., Billsus, D.: Machine learning for user modeling. User Model. User-Adapt. Interact. **11**(1–2), 19–29 (2001)
25. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Mach. Learn. **23**(1), 69–101 (1996)
26. Wolf, T., Sanh, V., Chaumond, J., Delangue, C.: Transfertransfo: a transfer learning approach for neural network based conversational agents. NeurIPS 2018 CAI Workshop (2019)
27. Wu, J., Wang, X., Wang, W.Y.: Self-supervised dialogue learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 3857–3867. Florence, Italy (2019)
28. Zhang, S., McClean, S., Scotney, B., Chaurasia, P., Nugent, C.: Using duration to learn activities of daily living in a smart home environment. In: 2010 4th International Conference on Pervasive Computing Technologies for Healthcare, pp. 1–8. IEEE (2010)
29. Žliobaitė, I.: Learning under concept drift: an overview (2010). arXiv:1010.4784