Emotion-aware cross-modal domain adaptation in video sequences

IMPACT PARAGRAPH

In this addendum, a discussion is presented to introduce the scientific and social impact of the conducted research in this dissertation, its results, and the proposed methodologies. The core research of this dissertation is domain adaptation, that is applied mainly in Human-Computer Interaction (HCI) and Affective Computing (AC). However, while the main experimentation was conducted in the spectrum of these fields, in principle, the applied methodologies could be easily transferred to a plethora of diverse applications where domain adaptation could be useful.

All these aforesaid applications have an enormous social and economic impact on society. On this ground, according to Maastricht University's "Regulations for obtaining the doctoral degree Maastricht University", dissertations should encompass an impact section which should include the "short-term" and "long-term" contributions of the conducted research and its results in relation to shifting insights and stimulating science, methodologies, results, theory, and applications. On the other hand, the social impact relates to the short and long-term contributions of the conducted research to changes in the development of social sectors and to social challenges. This paragraph addresses the drafted four questions in the doctorate regulations, which are related to the main objective of the research and its relevance, its target groups, and activities.

> **Research:** *What is the main objective of the research described in the thesis and what are the most important results and conclusions?*

The main objective of this dissertation is to address an important research problem in machine learning, that is: performing domain adaptation from audio and visual cues. It approaches the task from different perspectives with various methodologies with the end goal of enhancing the performance of Emotion Recognition (ER) when it is gauged in one modality by leveraging information from the other. For instance, the task can be to improve Audio Emotion Recognition (AER) by leveraging information from the face modality. In particular:

- Chapter 1 introduces the task under study and the state-of-the-art approaches in the fields of domain adaptation, emotion recognition (FER and AER) with the focus on the ones that widely inspired this dissertation. Furthermore, Chapter 1 presents state-of-the-art technologies, datasets, applications, modalities' representations, and learning schemes.

- In Chapter 2, the domain adaptation study is performed from the Distance Metric Learning (DML) perspective. In this case, a proof-of-concept algorithm is developed to model the audio-visual relations and study whether face modality can help improve AER. This approach is composed of several modules such as: feature

extraction and selection, clustering and the core DML projection. From the experimental phase, it is shown that it is indeed possible to transfer knowledge from face to audio modality.

- In Chapter 3, a deep learning direction is pursued. A study on Generative Adversarial Networks (GANs) is performed, with the purpose of discovering the correlation between face and audio modalities. Several methods are studied with the aim to build the proper architecture for the GANs network, and a proper way to tune the networks is also performed.
- As a follow-up research, a method to perform temporal analysis and study the temporal connection between face and audio modalities is applied in Chapter 4. This methodology makes use of 3d extracted features from face modality and attention mechanisms. A way to improve the training procedure of GANs architecture is also suggested.
- Finally, Chapter 5 studies the inverse task which is the improvement of FER using the audio modality. It is shown that it can be possible to increase the performance of the face modality by leveraging audio.

From the experimental phase, from all chapters it is clear that domain adaptation can be successfully applied to improve the performance of the audio or face modality by leveraging the other modalities and improving the AER and FER correspondingly. Chapter 2 provided a compact framework to perform domain adaptation, however, when we employed more sophisticated deep learning architectures (as in Chapters 3 and 4) we managed to outperform our initial results from Chapter 2. Moreover, we observed that it is really crucial to study the temporal relations between the two domains, which can lead to a more efficient "transfer" of knowledge between them. Finally, in Chapter 5 we performed a preliminary study on domain adaptation for the face modality. In this case, it was proven that it is possible to improve FER by employing audio information.

The next question that this chapter addresses is the following:

> **Relevance:** *What is the (potential) contribution of the results from this research to science, and, if applicable, to social sectors and social challenges?*

Our current era is mainly shaped from the so-called "digital revolution" in which the fields of data and computer science play a leading role. In the last decades, we are experiencing a constantly increasing interest in the fields of machine learning and deep learning in academia but also in industry. The "corporate world" has shown a great interest in investing in these fields and most of the big IT companies have already created their own dedicated "artificial intelligence" research and development department. This interest of the corporate world is also redeemed in our society since it equates to a shift in everyday life, which is now shaped through the omnipresence of modern technology. From mobile phones, smart TV's and electronic devices, to wearable health sensors and software that analyzes data for companies, states or individuals. In the near future, it seems likely that this tendency will increase and continue changing our lives. This tendency also fuels the popularity of research in machine learning and artificial intelligence in academia.

However, one of the notorious shortcomings of this cutting edge research is the so-called "lack of generalization". The developed machine learning algorithms need an enormous amount of datasets to learn how to perform a specific task, while, at the same time, they lack the flexibility to be employed in related tasks with slightly different characteristics and input datasets. Hence, in this dissertation we investigated a remedy for this notorious drawback which can be drawn from the research of the domain adaptation field.

Each chapter of this dissertation demonstrates the ability of the proposed solutions to perform DA efficiently between two inherently different modalities. This methodology can be used in a broader context, by applying it to different modalities. The importance of domain adaptation in research and as a consequence in society is beyond doubt.

By employing domain adaptation our purpose is to develop a more efficient framework that is able to combine data of a different nature to generate efficient models. In this scheme, we can leverage a big amount of data from different cues, which is crucial when developing a deep learning model since using only narrow data distributions is not really possible. Several popular deep learning and machine learning algorithms (object detection, language translation, face recognition, and so forth) can be benefited from this application.

The second contribution of this study is related to emotion recognition where the focus is to enhance emotionally incapable machines with emotional intelligence to improve human-machine interaction. Particularly, when the task is to perform emotion recognition in modalities for which we do not possess plenty of data. While the main focus of this dissertation is to perform domain adaptation, the task under study in each chapter is emotion recognition. We are modelling whether it is possible to perform domain adaptation with the purpose of enhancing the classification performance of a modality for instance, audio, by leveraging information from another modality. Hence, Chapters 2, 3 and 4 provide methodologies for performing emotion recognition from audio. While, in Chapter 5 we provide methodologies for performing face emotion recognition.

Target group: *To whom are the research results interesting and/or relevant? And why?*

The conducted research concerns developers, practitioners, and researchers in the fields of "machine learning" and affective computing. In this work, we provide several frameworks for performing "heterogeneous" domain adaption between two inherently different modalities such as face and audio. We hope that researchers in the same and in similar fields will be inspired to continue research in this direction and will expand our research and ideas in new and interesting paths.

As aforementioned, this work was tested on the domain of affective computing and concerns the study of audio-visual relations. However, it can be easily transferred to different tasks and different modalities by performing the necessary modifications in the corresponding parts of the approach. For instance, in the case of performing a different classification than emotion recognition, we will need to change and retrain all the involved classifiers. One simple example is the following: performing person identifica-

tion from audio that lacks annotated datasets by leveraging the huge availability of face recognition datasets. Towards this end, we can make use of the introduced techniques in this dissertation to generate audio samples by giving as input face samples, and leverage these generated samples to perform person identification.

Furthermore, our work can be useful to industry and developers that would like to develop robust classifiers in domains that lack large annotated datasets. In particular, they can leverage our approach by transferring knowledge from "close-related" domains to enhance the performance of the classifier at hand. A real-life tool for domain adaptation (part of the conducted research of this dissertation) was developed for the European Horizon research project called "MaTHiSiS" [1]. The scope of this tool was to improve emotion recognition performance from cues for which we do not have access to large datasets. Mainly, this tool was tested for performing audio emotion recognition by leveraging information from the face modality. However, this tool provides a friendly interface that can handle easily different modalities other that face and audio.

Some other applications where domain adaptation can be applied in industry is "the task of language translation", image classification for unseen objects, in gaming, in education applications and many more.

> **Activity:** *In what way can these target groups be involved in and informed about the research results, so that the knowledge gained can be used in the future?*

This thesis is article-based, where the studies in Chapters 2, 3 and 4 are published in various conferences and journal proceedings. At the beginning of each chapter, the papers which are parts of the corresponding chapter are listed. Moreover, throughout the course of the Ph.D. research, the proposed methodologies and the conclusions of their findings have been presented in the respective scientific venues. Besides, a tool for performing domain adaptation was developed for European Horizon2020 project called MaTHiSiS and was part of the whole learning framework.

[1] http://mathisis-project.eu/