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 methods. The research in this dissertation can be transferred to different tasks in Human-Computer Interaction (HCI) and Affective Computing (AC), as well as be implemented in various applications. These tasks and applications have enormous social and economical interests in the current society and the future. According to Maastricht University’s “Regulations for obtaining the doctoral degree Maastricht University” ¹, the scientific impact includes short-term and long-term contributions of the conducted research and its results to shifting insights and stimulating science, methods, results, theory, and applications. On the other hand, the social impact is the short-term and long-term contributions of the conducted research to changes in or development of social sectors and to social challenges. This paragraph is addressing the drafted four questions in the given regulations, which are related to the main objective of the research and its relevance, target group, and activity.

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 a fundamental research problem in Affective Computing (AC): Multimodal Emotion Recognition (MER) from audio and visual cues. It approaches the problem from different perspectives with various methods to enhance the performance of emotion recognition in video clips. Chapter 2 introduces the state-of-the-art approaches in Artificial Intelligence (AI) which are used in the fields of Human-Computer Interaction (HCI), Affective Computing (AC), and, in particular, in this dissertation. Furthermore, Chapter 3 presents state-of-the-art technologies, datasets, applications, modalities’ representations, learning schemes, and fusion techniques for MER. Besides, the subsequent chapters introduce the proposed methods, their findings, and conclusions as follows:

• In two studies, the research in Chapter 4 demonstrates the importance of multimodal features, their fusion, as well as an application of automatic emotion inference in educational settings. The first study, in this chapter, shows the impact of fusing different feature representations from audio and video modalities for emotion perception. It proves that these two modalities are complementary to each other, and their features improve performance significantly. The second study examines the link between the self-reported affective states by students and their interactions with learning materials. It provides an example of utilizing learning analytic technologies and machine learning techniques for understanding student’s

emotions, which is an important aspect in the future of e-learning. The study finds a correlation between students’ affects and their interaction parameters.

- The study in Chapter 5 focuses on improving the fusion algorithm to make use of audio-visual cues, efficiently. The proposed method, namely, Multimodal Emotion Recognition Metric Learning (MERML), shows the potential of powerful approaches such as metric learning for multimodal learning and fusion, which improved the latent space of audio-visual representations and subsequently the performance of emotion recognition.

- Chapter 6 follows the metric learning approach to produce joint multimodal and temporal feature representations for Audio-Video Emotion Recognition (AVER) using Deep Metric Learning (DML). This study exploits the temporal dynamics of audio and video signals using an end-to-end learning paradigm. In this chapter, the research demonstrates the importance of time information in the incremental presentation of audio-visual signals for emotion recognition. The proposed method was further adjusted to address another area in Affective Computing (AC), namely, personality recognition from bodily expressivities using motion and context information. Both studies prove the usability of Deep Metric Learning (DML), along with the proposed frameworks, for capturing multimodal data and modeling the temporal information in the field of Affective Computing (AC).

- Finally, Chapter 7 revisits the research problem of capturing the temporal dynamics of audio-visual modalities and further focuses on the idea of attending to informative time segments in these two modalities’ cues. The research in this chapter employs attention mechanisms to address the research objective. The results highlight the importance of automatically attending to informative time slices. Furthermore, the study introduces a meta-analysis, linking the research findings and propositions to research in psychology. The results offer an insightful perspective of the performance of the audio-visual cues over time, such as multimodal recognition speed on positive and negative emotions, the contribution of each modality in the multimodal fusion, the importance of their joint learning via the attention mechanisms, and the robustness of the proposed framework in more challenging environments.

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

The research in this dissertation aims at contributing to the goal of Affective Computing (AC), where the focus is to enhance emotionally incapable machines with emotional intelligence to improve human-machine interaction [164]. Each part of the conducted studies demonstrates the ability of the proposed solutions to perform Multimodal Emotion Recognition (MER), efficiently. Besides, the presented methods can be adapted to other tasks of Affective Computing (AC), as demonstrated in Chapter 6 with personality recognition, or other domains in Artificial Intelligence (AI) in general, where spatio-temporal and multimodal information persist, such as action recognition, person identification, and multimedia retrieval.
Moreover, many regard the current progress in Artificial Intelligence (AI) as a crucial part of the Fourth Industrial Revolution (Industry 4.0). The term Fourth Industrial Revolution was first referred to by Klaus Schwab in 2015, and published in [286]. It refers to the automation of manufacturing pipelines using the technological advances in the fields of Artificial Intelligence (AI), quantum computing, nanotechnology, the internet of things, etc. In this new era, where the focus is on machine-machine and human-machine communications, it is important to keep humans in the loop [287]. For this reason, within the advances of Artificial Intelligence (AI), a key factor to consider is emotional intelligence. Emotional intelligence refers to a set of skills that contribute to correct appraisal and expression of emotions, emotions modulation and regulation, and subsequently, the efficient usage of emotions in planning, working, and communications [288]. Emotional intelligence is essential in business, relationships, education, and other life aspects. Furthermore, the new era, brought by Industry 4.0, will lead to enormous consequences, that will shape our interactions with each other and the way we live, work, and learn [286, 287]. A key part of Affective Computing (AC) is emotion recognition. Indeed, obvious signals such as facial expressions and vocal utterances, which are addressed in this dissertation, can contribute to equipping machines with emotional intelligence. In this way, AI advancements are accompanied and supported with features to keep humans in new communication loops, namely, the machine-machine and human-machine communication channels.

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

The primary target groups of the studies in this thesis are researchers in Human-Computer Interaction (HCI) and Affective Computing (AC) fields, and the field of Artificial Intelligence (AI) in general. For example, the proposed multimodal and temporal architectures can be applied in other tasks, which include, but are not limited to gesture recognition, personality computing, action recognition, and multimedia retrieval. Moreover, as discussed in Section 3.4, the applications of AC and Multimodal Emotion Recognition (MER) can range from education [13, 18, 19], automatic vehicle driving [22–26], health-care [4, 20, 21], to entertainment [27–30]. These applications are of great interest in our societies with a tremendous social and economic impact.

For example, the developed techniques can be used in education where Technology Enhanced Learning (TEL) brings new kinds of educational and learning experiences [164]. According to R. W. Picard *et al.*, TEL systems should incorporate the emotional aspect of the learning process, in addition to the cognitive process [164]. As a result, human emotional needs are considered, beyond aspects that address merely productivity and efficiency. Indeed, in an educational context, emotions experienced by a learner directly affect the learning outcome [165, 166]. In fact, in the course of this dissertation, we considered emotion understanding from facial expressions, vocal utterances, and students’ interaction with learning materials. These three modalities can be part of an integrated learning platform that has affective capabilities to recognize learners’ emotions and to respond to their individual needs. In other words, accurate automatic multimodal emotion recognition can be useful in enhancing the learning outcomes by providing personalized and adaptive educational processes according to students’ emotions, as well
as other performance indicators related to productivity and cognitive skills. Another interesting area for the applications of Affective Computing (AC) and Multimodal Emotion Recognition (MER) is the recognition of drivers’ affective states in automatic vehicle driving. For example, the developed methods within this dissertation can be used to fuse various sensorial data to infer driver’s attention and stress-level. The sensorial data can include facial expressions, gaze, and physiological measurements. An affective system can ensure drivers’ and other people’s safety.

**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?*

The studies in Chapters 4, 5, 6, and 7 have been published in various peer-reviewed conference proceedings and high-impact journals. 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 methods and the conclusions of their findings have been presented in the respective scientific venues. Furthermore, the Convolutional Neural Network (CNN) representations of the study in Chapter 4 were part of an interactive demo that has been used to demonstrate its abilities to recognize facial expressions in real-time. Besides, the study of correlating students’ affective states with their interactions with the learning materials was one of the modalities in MaTHiSiS, which is a learning platform developed within the Horizon 2020 funded project MaTHiSiS (Managing Affective-learning THrough Intelligent atoms and Smart InteractionS)².

²[http://mathisis-project.eu/](http://mathisis-project.eu/)