processes

MDPI

Review

Multimodal Age and Gender Estimation for Adaptive Human-Robot Interaction: A Systematic Literature Review

Hussain A. Younis [1,2], Nur Intan Raihana Ruhaiyem [1,*], Ameer A. Badr [3], Alia K. Abdul-Hassan [4], Ibrahim M. Alfadli [5], Weam M. Binjumah [6], Eman A. Altuwaijri [7] and Maged Nasser [1]

1 School of Computer science, Universiti Sains Malaysia, Penang 11800, Malaysia; hussain.younis@uobasrah.edu.iq (H.A.Y.)
2 College of Education for Women, University of Basrah, Basrah 61004, Iraq
3 Department of Information Technology, Technical College of Management-Baghdad, Middle Technical University, Baghdad 10011, Iraq; ameer.badr@duc.edu.iq
4 Department of Computer Science, University of Technology, Baghdad 10011, Iraq
5 College of Computer Science and Engineering, Taibah University, Madina 42353, Saudi Arabia; ialfadli@taibahu.edu.sa
6 Applied Collage, Taibah University, Madina 42353, Saudi Arabia
7 College of Applied Studies and Community Service, King Saud University, Riyadh 145111, Saudi Arabia
* Correspondence: intanraihana@usm.my

Abstract: Identifying the gender of a person and his age by way of speaking is considered a crucial task in computer vision. It is a very important and active research topic with many areas of application, such as identifying a person, trustworthiness, demographic analysis, safety and health knowledge, visual monitoring, and aging progress. Data matching is to identify the gender of the person and his age. Thus, the study touches on a review of many research papers from 2016 to 2022. At the heart of the topic, many systematic reviews of multimodal pedagogies in Age and Gender Estimation for Adaptive were undertaken. However, no current study of the theme concerns connected to multimodal pedagogies in Age and Gender Estimation for Adaptive Learning has been published. The multimodal pedagogies in four different databases within the keywords indicate the heart of the topic. A qualitative thematic analysis based on 48 articles found during the search revealed four common themes, such as multimodal engagement and speech with the Human-Robot Interaction life world. The study touches on the presentation of many major concepts, namely Age Estimation, Gender Estimation, Speaker Recognition, Speech recognition, Speaker Localization, and Speaker Gender Identification. According to specific criteria, they were presented to all studies. The essay compares these themes to the thematic findings of other review studies on the same topic such as multimodal age, gender estimation, and dataset used. The main objective of this paper is to provide a comprehensive analysis based on the surveyed region. The study provides a platform for professors, researchers, and students alike, and proposes directions for future research.

Keywords: multimodal; age estimation; gender estimation; speech; image; dataset

Citation: Younis, H.A.; Ruhaiyem, N.I.R.; Badr, A.A.; Abdul-Hassan, A.K.; Alfadli, I.M.; Binjumah, W.M.; Altuwaijri, E.A.; Nasser, M. Multimodal Age and Gender Estimation for Adaptive Human-Robot Interaction: A Systematic Literature Review. *Processes* **2023**, *11*, 1488. https://doi.org/10.3390/pr11051488

Academic Editors: Adel Ali Ahmed, AbdulRahman Alsewari, Yousef Fazea and Waleed Ali

Received: 18 February 2023
Revised: 6 April 2023
Accepted: 10 April 2023
Published: 15 May 2023

1. Introduction

With the world's latest tremendous development, technological advancement, and information age, we started accessing this study which is referred to the main terms in the study, i.e., each of Multimodal Age Estimation [1,2]. Typically, it is more challenging to assume the age of a speaker based on their speech [3], Gender Estimation [4], and Human-Robot Interaction [5–7]. The first term, Multimodal, refers to the theory of communication between social auditors that represent communication between audiovisual, visual, and spatial resources. The second term is age and gender estimation. Therefore, many studies have presented Multimodal in meta-learning [8], English language [9], a comprehensive presentation of Vocal sacs [10], deep learning fields of vision, language, and speech [11–15].