A Novel Distance Learning for Elastic Cross-Modal Audio-Visual Matching

Publisher: IEEE

Rui Wang; Huaibo Huang; Xufeng Zhang; Jixin Ma; Aihua Zheng

Abstract:

In this work we propose a novel network formulation for joint representation of cross-modal audio and visual information base on metric learning. We employ a distance learning framework as a training procedure. For this purpose we introduce an elastic matching network (EmNet) and a novel loss function to learn the shared latent space representation of multi-modal information. The elastic matching network is capable of matching given face image (or audio voice clip) from diverse number of audio clips (or face images). We quantitatively and qualitatively evaluate the purposed approach on the standard audio-visual matching evaluation dataset, the over-lap of VoxCeleb and VGGFace by both multi-way and binary audio-visual matching tasks. The promising performance comparing to the existing methods verifies the effectiveness of the proposed approach, which yields to a new state-of-the-art for cross-modal audio-visual matching.

Published in: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)

Date of Conference: 08-12 July 2019

Date Added to IEEE *Xplore*: 15 August 2019

ISBN Information:

Electronic ISBN:978-1-5386-9214-1

USB ISBN:978-1-5386-9213-4

Print on Demand(PoD) ISBN:978-1-5386-9215-8