Human Emotion Recognition from Image and Speech using Deep Neural Networks

Streszczenie. Rozpoznawanie emocji to ważny obszar badawczy dotyczący interakcji człowiek-komputer. Pomimo, że komputerowa analiza danych sensorycznych takich jak obraz twarzy i głos, osiąga spektakularne wyniki, w wielu przypadkach lepsze od wyników osiąganych przez ludzi, to automatyczna dwu-modalna analiza emocji na podstawie obrazu i dźwięku jednocześnie, tak jak to faktycznie realizuje mózg człowieka, daleka jest jeszcze od mierzalnego poziomu ludzkich możliwości. Niniejsza rozprawa doktorska jest próbą zbliżenia się do tej granicy. Przedstawione wyniki dotyczą czterech typowych scenariuszy badawczych stosowanych klasyfikacji emocji, dokonywanej na podstawie danych ekstrahowanych z: (a) pojedynczego zdjęcia twarzy, (b) nagrania wideo, tj. z temporalnej sekwencji obrazów, (c) nagrania audio, tj. nagrania mowy, (d) klipu filmowego, tj. zsynchronizowanego nagrania wideo i audio. W systemach rozpoznawania wyróżnia się komponenty służące wydobywaniu cech i komponenty klasyfikujące te cechy. W scenariuszu (a) w niniejszej pracy pokazano wyższość rozwiązania neuronowego nad klasycznym już podejściem, w którym cechy geometryczne i animacyjne modelu Candide-3, uzyskuje się na podstawie detekcji punktów szczególnych modelu FP68, a następnie klasyfikuje w modelu tzw. maszyny wektorów nośnych (SVM). Właściwa strategia uczenia się cech głębokich przez inne zadania związane z twarzami, tj. technika transferu modelu neuronowego sprawiła, że proponowany model jest skuteczny nawet przy stosunkowo ograniczonych zasobach zdjęć w zbiorze uczącym. W scenariuszu (b) zauważono, że temporalne urozmaicanie danych uczących znacząco poprawia skuteczność klasyfikatora emocji na podstawie sekwencji obrazu. Z kolei analiza sygnału mowy w scenariuszach (c) i (d) prowadzona jest na podstawie jego spektrogramu. Scenariusz (d), a więc dwu-modalna analiza emocji, z możliwym jej rozszerzeniem na przypadek wielo-modalny, jako najbardziej zbliżona do zachowań człowieka, zajmuje w pracy prominentne miejsce. Wykorzystując komponenty opracowane w realizacji scenariuszów (b) i (c), skupiono się na zagadnieniu fuzji rozwiązań jedno-modalnych. Zaproponowana architektura MRPN (Multimodal Residual Perceptron Network) eliminuje niedoskonałości rozwiązań stosujących tzw. późną fuzję i prowadzi do aktualnie najlepszych wyników osiąganych w klasyfikatorach emocji łączących dane wideo i audio, tj. na następujących, powszechnie stosowanych zbiorach danych testowych: RAVDESS, Crema-d, FER2013, RaFD, MUG oraz CK+.

Słowa kluczowe: rozpoznawanie emocji twarzy, rozpoznawanie mowy, rozpoznawanie emocji audio-wideo, multimodalna sieć neuronowa, głęboka fuzja funkcji

Human Emotion Recognition from Image and Speech using Deep Neural Networks

Summary. Recognizing emotions is an important research area of human-computer interaction. Although computer analysis of sensory data such as face image and voice achieves spectacular results, in many cases better than human results, automatic bi-modal analysis of emotions based on image and sound simultaneously, as is actually done by the human brain, is still far from a measurable level of human capacity. This doctoral dissertation is an attempt to get closer to this border. The presented results concern four typical research scenarios for the applied classification of emotions, made on the basis of data extracted from: (a) a single photo of a face, (b) a video recording, i.e. from a temporal sequence of images, (c) audio recordings, i.e. speech recordings, (d) a movie clip, i.e. synchronized video and audio recording. Recognition systems distinguish components for extracting features and components that classify these features. In scenario (a) in this paper, the superiority of the neural solution over the classic approach, in which the geometric and animation features of the Candide-3 model are obtained on the basis of the detection of special points of the FP68 model, and then classified in the model, the so-called support vector machines (SVMs). The proper strategy of learning deep features through other face-related tasks, i.e. the neural model transfer technique, made the proposed model effective even with relatively limited resources of images in the training set. In scenario (b) it was noticed that temporal augmentation of the training data significantly improves the effectiveness of the emotion classifier based on the image sequence. In turn, the analysis of the speech signal in scenarios (c) and (d) is carried out on the basis of its spectrogram. Scenario (d), i.e. the two-modal analysis of emotions, with a possible extension to the multi-modal case, as being the closest to human behavior, occupies a prominent place at work. Using the components developed in the implementation of scenarios (b) and (c), the focus was on the issue of fusion of one-modal solutions. The proposed MRPN (Multimodal Residual Perceptron Network) architecture eliminates the imperfections of solutions using the so-called late fusion and leads to the currently best results in emotion classifiers combining video and audio data, i.e. on the following commonly used test data sets: RAVDESS, Crema-d, FER2013, RaFD, MUG and CK+.

Keywords: facial emotion recognition, speech emotion recognition, audio-video emotion recognition, multi-modal neural network, deep feature fusion