RADA NAUKOWA DYSCYPLINY
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

zaprasza na

PUBLICZNĄ OBRONĘ ROZPRAWY DOKTORSKIEJ

mgr. inż. XIN CHANGA

która odbędzie się w dniu **30 maja 2022 roku o godzinie 12.00** w trybie stacjonarnym

Temat rozprawy doktorskiej:

,,Human emotion recognition from image and speech using deep neural networks"

Promotor: prof. dr hab. inż. Władysław Skarbek - Politechnika Warszawska

Recenzenci: prof. dr hab. inż. Bogusław Cyganek – Akademia Górniczo-Hutnicza w Krakowie

dr hab. inż. Paweł Forczmański – Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

prof. dr hab. inż. Konrad Wojciechowski – Politechnika Śląska

* Obrona odbędzie się w Sali nr 116. Osoby zainteresowane uczestnictwem w obronie proszone są o zgłoszenie chęci uczestnictwa w formie elektronicznej na adres sekretarza komisji: dr hab. inż. Grzegorz Pastuszak, prof. uczelni – email : Grzegorz.pastuszak@pw.edu.pl do dnia 27 maja 2022 do godz.12.

Z rozprawą doktorską i recenzjami można zapoznać się w Czytelni Biblioteki Głównej Politechniki Warszawskiej, Warszawa, Plac Politechniki 1.

Streszczenie rozprawy doktorskiej i recenzje są zamieszczone na stronie internetowej: https://bip.pw.edu.pl/Postepowania-w-sprawie-nadania-stopnia-naukowego/Doktoraty/Wszczete-po-30-kwietnia-2019-r/Dyscyplina-informatyka-techniczna-i-telekomunikacja-dziedzina-nauk-inzynieryjno-technicznych/mgr-inz.-Xin-Chang.

Przewodniczący Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej

dr hab. inż. Jarosław Arabas

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

WEAIiE Kancelaria
wpłynęło dnia 22.03.22r.
nr

Kraków, 10.03.2022

Prof. dr hab. inż. Bogusław Cyganek
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
Wydział Informatyki, Elektroniki i Telekomunikacji
Instytut Elektroniki

Recenzja rozprawy doktorskiej *„Human emotion recognition from image and speech using deep neural networks"* Pana magistra inżyniera Xin Changa

Wstęp

Recenzja dotyczy rozprawy doktorskiej Pana mgra inżyniera Xin Changa zatytułowanej *„Human emotion recognition from image and speech using deep neural networks"*. Praca powstała na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej, a promotorem pracy jest Pan prof. dr hab. inż. Władysław Skarbek. Praca została napisana w języku angielskim.

Recenzja została wykonana na zlecenie Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej. Jest ona zorganizowania w postaci odpowiedzi na pytania dotyczące rozprawy doktorskiej Pana Xin Changa.

1. *Jakie zagadnienie naukowe jest rozpatrzone w pracy i czy zostało ono dostatecznie jasno sformułowane prze autora? Jaki charakter ma rozprawa?*

Główne zagadnienie naukowe podjęte w pracy doktorskiej Pana magistra inżyniera Xin Changa dotyczy *opracowania oraz weryfikacji naukowej nowych metod rozpoznawanie emocji w systemach interaktywnych człowiek-komputer z użyciem* danych wielo-sensorycznych w postaci obrazu i dźwięku *i wykorzystujących nowoczesne sieci neuronowe.*

Rozpoznawanie stanów emocjonalnych, zadowolenia, lęku, gniewu, itd. jest charakterystyczną cechą wszelakich interakcji międzyludzkich. Stany te oceniamy nawet instynktownie i często w ułamku sekundy, korzystając z sygnałów odbieranych od innych osób, takich jak ich wygląd, głos, ale również wyraz twarzy czy też ułożenie rąk. Niezwykle ważny jest tutaj rodzaj kontaktu między ludźmi – ocena stanu jest przeważnie łatwiejsza przy kontakcie bezpośrednim, natomiast odbiór i ocena stanów emocjonalnych może ulec osłabieniu przy ograniczonej percepcji sygnałów, na przykład przy połączeniach multimedialnych w dobie CoVid19. Okazuje się, że rozpoznawanie tego typu stanów

emocjonalnych przez algorytm komputerowy jest zadaniem o wiele trudniejszym, nawet w dobie dynamicznego rozwoju sztucznej inteligencji oraz uczenia maszynowego.

Powyższe zagadnienie naukowe zostało w pracy jasno sformułowane, a następnie przeanalizowane, a głównym rezultatem tej pracy jest propozycja rozwiązania w postaci architektury wielomodalnych hybrydowych architektur głębokich sieci neuronowych działających jednocześnie na danych audio oraz wideo. Dzięki takiemu podejściu, integrującemu równoległe tory przetwarzania sygnałów pochodzących z różnych sensorów, możliwe było wyeliminowanie problemu tzw. późnej fuzji odpowiedzi torów jedno-modalnych, jak również tzw. uczenia całościowego sieci typu *end-to-end*. Użyteczność tego podejścia Doktorant wykazał wykorzystując ogólnie dostępne bazy danych testowych, na których osiągnął wyniki porównywalne lub nawet przewyższające rezultaty znane z literatury światowej osiągając w ten sposób poziom *state-of-the-art*. Wyniki te zostały również opublikowane w recenzowanym czasopiśmie o światowej renomie.

Oprócz powyższego Doktorant zbadał działania klasyfikatorów głębokich sieci neuronowych na różnych wersjach sygnałów jednomodalnych typu obraz, sygnał wideo, sygnał audio, jak również ich połączenia w postaci klipu filmowego zawierającego sygnał wideo oraz audio. Została też wykonana weryfikacja innych metod klasyfikacji.

Uwzględniwszy powyższe, uważam że problem badawczy podjęty przez Doktoranta Pana magistra inżyniera Xin Changa ma znaczenie zarówno w obszarze badań teoretycznych, jak i aplikacyjnych. Charakter pracy jest więc teoretyczno-doświadczalny.

2. *Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł (w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle) świadczący o dostatecznej wiedzy autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?*

Opracowanie oraz skuteczna realizacja celów naukowych Doktoranta, jak to zostało wspomniane w poprzednim punkcie niniejszej recenzji, nie byłoby możliwe bez dogłębnej analizy najnowszych źródeł i opracowań światowych. Dotyczy to głównie systemów wykorzystujących głębokie sieci neuronowe stanowiące swoisty *state-of-the-art* w dziedzinie klasyfikacji sygnałów cyfrowych.

W rozprawie zamieszczona została lista bibliograficzna zawierająca 117 pozycji literaturowych, w tym 6, w których Doktorant Pan magister inżynier Xin Chang jest jednym ze współautorów. Oceniam, że w rozprawie przeprowadzono w sposób właściwy analizę źródeł, w tym literatury światowej, stanu wiedzy i potencjalnych zastosowań, co świadczy o dużej wiedzy autora w szeroko pojętej dziedzinie uczenia maszynowego i sztucznej inteligencji ML/AI (ang. *Machine Learning and Artificial Intelligence*). Tym niemniej, pracę można by zredagować lepiej poprzez dołączenie do bibliografii i odniesienie się do prac przeglądowych w dziedzinie komputerowego rozpoznawania emocji, jak również architektur sieci neuronowych, które są używane w metodach opracowanych przez Doktoranta.

3. *Czy autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?*

Główne zagadnienie naukowe polegające na opracowaniu oraz weryfikacji naukowej metod rozpoznawanie emocji w systemach interaktywnych człowiek-komputer za pomocą jednoczesnego przetwarzania zarówno obrazu, jak i dźwięku, zostało przez Pana magistra inżyniera Xin Changa prawidłowo rozwiązane poprzez proces tworzenia kolejnych wersji klasyfikatorów, badanie ich działania, obserwacje występujących problemów, a następnie zaproponowanie kolejnej wersji eliminującej poprzednie niedogodności i prowadzącej do uzyskania lepszych rezultatów. Proces ten, wsparty intensywną analizą dostępnych rozwiązań, jest prawidłowym procesem badawczym i świadczy o dobrym opanowaniu przez Doktoranta metodyki prowadzenia badań naukowych. Do realizacji kolejnych wersji Pan Xin Chang użył zarówno klasyfikatorów starszych, typu SVM, które bazują na eksperckich modelach klasyfikacji oraz ekstrakcji cech, jak również najnowszych głębokich splotowych sieci neuronowych, które realizują paradygmat uczenia modelu wprost z danych. W ciągu ostatnich dziesięciu lat sieci tego typu wykazały znaczną przewagę nad poprzednimi rozwiązaniami bazującymi na modelach eksperckich, stając się *de facto* standardowym narzędziem przetwarzania danych. Tym niemniej, użycie sieci, a szczególnie umiejętność modyfikacji ich architektur w celu uzyskania lepszych wyników – co zostało zademonstrowane przez Doktoranta w prezentowanych rozwiązaniach – wymaga znacznej wiedzy oraz doświadczenia. Użyte przez Doktoranta metody uważam zatem za prawidłowo wybrane, o czym przede wszystkim świadczą bardzo dobre wyniki eksperymentalne uzyskane przez Doktoranta.

Przedstawione w rozprawie Pana magistra inżyniera Xin Changa zaawansowane metody rozpoznawania emocji za pomocą hybrydowych głębokich sieci neuronowych świadczą o jego głębokiej wiedzy w dyscyplinie Informatyka Techniczna i Telekomunikacja. W szczególności tematyka ta wpisuje się w dziedziny ML/AI, jak również widzenia komputerowego i technik multimedialnych. Opracowane przez niego metody są na wysokim światowym poziomie, a uzyskane wyniki eksperymentalne plasują je wśród ścisłej czołówki rozwiązań światowych, tzw. *state-of-the-art*.

4. *Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?*

Jak już wspomniano, badania naukowe oraz osiągnięcia Pana magistra inżyniera Xin Changa dotyczą opracowania oraz weryfikacji naukowej nowych metod rozpoznawanie emocji w systemach interaktywnych człowiek-komputer. Do tego celu Autor używa naturalnych danych wielo-sensorycznych, czyli obrazu i dźwięku, oraz odpowiednio skonstruowanych algorytmów rozpoznawania wzorców.

Główne osiągnięcia Doktoranta Pana Xin Changa w tej dziedzinie i zaprezentowane w rozprawie doktorskiej można podsumować następująco:

- Opracowanie architektury wielomodalnej sieci rezydualnej (ang. *Multimodal Residual Perceptron Network – MRPN*) działającej jednocześnie na danych audio oraz wideo, która eliminując problemy tzw. późnej fuzji odpowiedzi torów jedno-modalnych oraz uczenia całościowego *end-to-end*,

i która pozwala na ogólnie dostępnych bazach danych testowych typu *The Ryerson Audio–Visual Database of Emotional Speech and Song* (RAVDESS) oraz *Crowd-Sourced Emotional Multimodal Actors Dataset* (Crema-d) na osiągnięcie wyników porównywalnych lub nawet lepszych niż prezentowane w światowej literaturze naukowej osiągając tzw. poziom *state-of-the-art*.

- Zbadanie działania klasyfikatorów głębokich sieci neuronowych na różnych wersjach sygnałów jednomodalnych typu obraz, sygnał wideo, sygnał audio, jak również ich połączenia w postaci klipu filmowego zawierającego sygnał wideo oraz audio.
- Opracowanie metod uzupełniania danych w dziedzinie czasowej (ang. *time data augmentation*) oraz zbadanie ich wpływu na jakość klasyfikacji zarówno rozwiązań jedno-modalnych, jak i wielo-modalnych.
- Naukowa weryfikacja metod klasyfikacji z wykorzystaniem maszyn z wektorami wspierającymi (ang. *Support Vector Machine – SVM*) *versus* klasyfikacje z wykorzystaniem głębokich sieci neuronowych na podstawie cech geometrycznych i animacyjnych z modelu Candide-3 oraz detekcji punktów szczególnych modelu FP68.

Głównym problemem, na który natknęli się badacze zajmujący się trenowaniem sieci do operowania z sygnałami wielomodalnymi jest potencjalna możliwość spadku jakości całościowego modelu po połączeniu gałęzi jedno-modalnych. W tym kontekście na szczególną uwagę zasługuje pierwsza z wymienionych tu metod, tj. wielomodalna neuronowa sieć rezydualna MRPN, w której w oryginalny sposób Autor wykorzystał wielo-składnikową funkcję kosztu oraz technikę mieszania gradientów (ang. *gradient blending*), która umożliwiła wykorzystanie zarówno tzw. późnej fuzji (ang. *late fusion*) dwóch gałęzi przetwarzania różnych typów sygnałów wejściowych – to jest obrazu oraz dźwięku – jak również wykorzystanie tzw. uczenia całościowego, od końca do końca całej sieci, tzw. *end-to-end*. W szczególności, zastosowanie późniejszej fuzji sygnałów pozwala na przechowanie informacji z wysokim stopniem pewności, wyuczonej przez sieć ale niezależnie w każdym trybie jedno-modalnym (np. tylko obraz, albo tylko dźwięk). Powstałe w ten sposób gałęzie (pod-sieci) są dość odporne na wartości odstające ale w ich własnej domenie, co prowadzi do lepszej parametryzacji i w konsekwencji do usprawnienia właściwości generalizacyjnych tych jedno-modalnych gałęzi, a w rezultacie usprawnienia całego systemu, tj. po dalszym połączeniu obydwu (wielu) gałęzi w porównaniu np. do systemu uczonego w trybie *end-to-end*. Wychodząc z rozwiązań zaproponowanych w pracy Wang W. *et al.*, istotną i oryginalną modyfikacją jest zaproponowana przez Doktoranta ścieżka obejścia sygnału w module perceptronu rezydualnego (ang. *Residual Perceptron – RP*), która pozwala na wykorzystanie dobrych właściwości zarówno *late fusion* oraz *end-to-end* z jednoczesnym ominięciem słabych stron tych obydwu strategii uczenia. Dzięki temu podejściu w trybie predykcji (testu) aktywne są wyłącznie centralne moduły bloku RP z nieliniową funkcją sigmoidalną (a nie ReLU), podczas gdy boczne obejścia są aktywne wyłącznie w trybie uczenia. System ten można nazwać kontrolowaną fuzją modów, w którym zaproponowany moduł RP umożliwia uzyskanie lepszej parametryzacji cech właściwych danej gałęzi modalnej jeszcze przed fuzją z innymi gałęziami modalnymi. Jest to bardzo ciekawy pomysł, zarówno w aspekcie teoretycznym, jak również mający bezpośrednie przełożenie praktyczne tj. prowadzący do znacznie lepszych wyników działania tak połączonego systemu hybrydowego, szczególnie w przypadku napotkania danych niekompletnych i/lub zaszumionych, co nie jest rzadkością w przypadku zbiorów wielo-modalnych.

Nie mniej ciekawym osiągnięciem jest wykazanie, że zastosowanie uzupełniana danych w domenie temporalnej (ang. *time-dependent augmentation*) pozwala na zwiększenie dokładności klasyfikatorów stanów emocjonalnych, działających w trybach z różnymi modami. Wynik ten może być zapewne użyty w innych systemach tego typu.

Opracowana nowatorska metoda MRPN została przetestowana na zbiorach danych RAVDESS oraz Crema-d, wykazując w trybie uczenia *end-to-end* najlepsze rezultaty w porównaniu z innymi znanymi metodami opublikowanymi w światowym piśmiennictwie naukowym. Metoda ta została opublikowana również w czasopiśmie Sensors, o czym będzie mowa w kolejnym z punktów recenzji.

Również ciekawe i pomysłowe są pozostałe metody opracowane i przebadane przez Doktoranta, a które zostały opisane w rozdziale 4. Jest to m.in. metoda rozpoznawania emocji za pomocą analizy twarzy oraz modelowania 3D (rozdz. 4.1), porównanie rozwiązań starszych opartych na cechach eksperckich z rozwiązaniami bazującymi na głębokich sieciach neuronowych (rozdz. 4.2), analiza transferu wiedzy (rozdz. 4.3), jak również metodyka rozpoznawanie emocji w strumieniach wideo i z wykorzystaniem głębokich sieci neuronowych typu CNN oraz LSTM (rozdz. 4.4) oraz zaproponowane wykorzystanie cech pochodzących ze spektrogramów sygnałów, gdzie jednym z głównych problemów jest zapewnienie synchronizacji zmiennych czasowo sekwencji wideo oraz audio (rozdz. 4.5).

Wszystkie wyżej wymienione osiągnięcia świadczą o dużej dojrzałości naukowej Pana magistra inżyniera Xin Changa, który swobodnie porusza się zarówno w dziedzinie ML/AI, widzenia komputerowego, uczenia maszynowego, przetwarzania sygnałów multimedialnych, jak również w dziedzinie programowania systemów komputerowych.

5. *Jaka jest umiejętność poprawnego i przekonywującego przedstawienia uzyskanych wyników oraz jaki jest dorobek publikacyjny Doktoranta*

Rozprawa doktorska Pana magistra inżyniera Xin Changa liczy 95 stron. Napisana jest w języku angielskim. Praca podzielona jest na 6 rozdziałów głównych. Bibliografia liczy 117 pozycji literaturowych.

Oddzielnie wymienione jest 6 prac, w których współautorem jest Doktorant, a w 4 z nich jest on pierwszym z autorów. Struktura publikacji Doktoranta Pana Xin Changa jest następująca:

- Cztery prace współautorskie opublikowane w latach 2017-2020 w materiałach konferencyjnych SPIE Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments.
- Jedna publikacja wielo-autorska w bazie arxiv (https://arxiv.org/pdf/1901.11179.pdf).
- Jedna publikacja w czasopiśmie naukowym Sensors, 2021 (Impact Factor: 3.576), w której Doktorant jest pierwszym z autorów, natomiast drugim jest Promotor (z opisu w artykule udział autorów można oszacować jako 50:50). Jest to praca, w której opublikowano architekturę oraz wszystkie wyniki głównej metody MRPN. Prawie bez zmian została ona opisana również w rozdziale 5 rozprawy doktorskiej.

Wszystkie te publikacje dotyczą dziedziny uczenia maszynowego oraz widzenia komputerowego wraz z aplikacjami.

6. *Jakie są słabe strony rozprawy i jej główne wady?*

Rozprawa doktorska Pana Xin Changa zawiera opis głównych metod opracowanych przez Doktoranta, o których była już mowa w poprzednich punktach. Tym niemniej, poszczególne rozdziały nie są na jednakowo wysokim poziomie naukowym. Jestem przekonany, że znaczne skrócenie całej rozprawy, czy też jej organizacja w postaci przewodnika z dołączeniem artykułu Autora z czasopisma Sensors

znacznie by poprawiło jej czytelność. Poniżej zawarłem główne uwagi oraz pytania, które w znakomitej większości wynikają z przyjętej przez Doktoranta struktury pracy.

1. Struktura pracy oraz zamieszczone treści w poszczególnych rozdziałach są na bardzo różnym poziomie naukowym oraz z bardzo różnym poziomem szczegółowości. Moje największe wątpliwości budzi rozdział 2, zatytułowany „Artificial Neural Network" – nie bardzo rozumiem cel tego rozdziału, jak i zamieszczonych tam treści. Wygląda na to, że Autor podjął się próby przedstawienia krótkiego przewodnika dotyczącego działania sztucznych sieci neuronowych, jednakże próbę tę uznaję za nieudaną, gdyż tematyka jest zbyt obszerna a zamieszczone informacje są albo powierzchowne, albo wyrwane z głębszego kontekstu, a w rezultacie nieprecyzyjne i przez to mało użyteczne. Pisanie tego typu opisów być może miało sens w przypadku prac doktorskich pisanych kilkadziesiąt lat temu, kiedy dostęp do źródeł był bardzo utrudniony choćby prze brak, czy tez początki Internetu. Jednakże w dzisiejszych czasach, dostępności wielu portali, Wikipedii, artykułów naukowych często w trybie *open-access*, itd., znacznie lepsze rezultaty Autor osiągnąłby koncentrując się na swoich osiągnięciach przy jednoczesnym dostarczeniu odwołań do istotnych pozycji literaturowych i odpowiednim umiejscowieniu swojej tematyki badawczej.

   Zamiast tych zupełnie podstawowych wiadomości, takich jak perceptron, bardzo szkoda, że Autor nie wykorzystał swojej wiedzy i potencjału i nie opisał w zwięzły, ale i profesjonalny, sposób nowoczesnych architektur, których opis byłby przydatny dla zainteresowanego czytelnika a dotyczyłby zaawansowanych technik użytych przez Autora (takich jak zmodyfikowane moduły rezydualne, sieci 3D zamiast kerneli w 2.5.1, itd.) lub też takich jak *vision transformer*, jedynie wspomniany przez Autora w rozdziale 2.7, czy też fascynujący temat atencji (*attention and self-attention*) w sieciach neuronowych, itp.

2. Nie mniej ciekawe, i zapewne pouczające, byłoby przeczytać krótki przegląd po nowoczesnych światowych rozwiązaniach dotyczących wielo-modalnych systemów rozpoznawania emocji, z ewentualną dyskusją nowych i potencjalnie obiecujących kierunków. Zestawienie tego typu mogłoby posłużyć do stworzenia artykułu naukowego w rodzaju „scientific review", które jeśli dobrze skomponowane, znajdują dużą poczytność oraz liczne cytowania. Przykładem może tu posłużyć artykuł K. Patel *et al.*: *Facial Sentiment Analysis Using AI Techniques: State-of-the-Art, Taxonomies, and Challenges*, IEEE Access, 2020.

3. Z dużym zdziwieniem przeczytałem wstęp do rozdziału 5, a następnie cały rozdział, który właściwie jest kopią artykułu Doktoranta i Promotora w czasopiśmie Sensors z roku 2021. Artykuł ten jest typu *open-access*, a więc doskonale dostępny dla każdego i w każdej chwili, więc od razu nasuwa się pytanie po co tworzyć rozdział 5 zamiast co najwyżej krótkiego przewodnika i odwołania do właściwego artykułu w czasopiśmie Sensors. Rozdział ten miałby co prawda sens w przypadku gdyby Autor chciał istotnie rozszerzyć i zaprezentować czytelnikom Jego rozprawy doktorskiej, czy to wstęp, czy to samą metodę. Szczególnie to ostatnie byłoby bardzo wartościowe, tzn. gdyby rzeczywiście, a nie tylko hipotetycznie tak jak w rozdziale 5.2.3 i rysunku 5.5, pokazać że zaproponowana metoda MRPN jest ogólniejsza i może być zastosowana do innych, czy też większej liczby, różnie-modalnych sygnałów niż tylko do obrazu i dźwięku, jak to zaprezentowano zarówno w rozprawie, jak i we wspomnianym artykule w czasopiśmie Sensors. Warto by ten temat dalej przedyskutować.

4. Bardzo ciekawym kierunkiem badawczym jest również tzw. uczenie ze wzmocnieniem (ang. *reinforcement learning*) – szkoda, że Autor nawet nie wspomniał o tym kierunku badawczym, tym bardziej że zasługuje on na miano „sztuczna inteligencja" chyba w większym stopniu niż same sieci neuronowe. Przykładowo niedawny artykuł K. Zhang et al. *Real-Time Video Emotion Recognition Based on Reinforcement Learning and Domain Knowledge* IEEE Tr. On Circuits and

Systems for Video Technology, 2022 przedstawia to zagadnienie, jak również zawiera odwołania do wcześniejszych prac w tej dziedzinie.

5. Ciekawym kierunkiem jest też ostatnie z wymienionych osiągnięć Autora, tj. porównanie modelu opartego na rozpoznawaniu charakterystycznych punktów twarzy, a następnie klasyfikacja z ich wykorzystaniem. Uzyskane wyniki świadczą o znacznej przewadze głębokich sieci neuronowych, co raczej nie jest zaskoczeniem. Tym niemniej, bardzo ciekawe i co najmniej warte dyskusji, jest podejście łączące te obydwie domeny, tj. cechy eksperckie oraz cech wykryte za pomocą sieci. Czy takie podejście byłoby możliwe w przypadku systemu opracowanego przez Doktoranta? Jako przykład tego kierunku badań można podać artykuł G.V. Reddy et al.: *Facial expression recognition in the wild, by fusion of deep learnt and hand-crafted features*, Cognitive Systems Research, 2020.

6. Niezbyt dokładnie wyjaśniono pojęcie STFT, który to skrót nawet nie został wymieniony w zbiorczej tabeli symboli (str. 83). Chodzi tu zapewne o *Short-Time Fourier Transform* – o czym możemy się przekonać uważnie analizując rys 5.3 albo sięgając do publikacji Autora w czasopiśmie Sensors – tym niemniej, w tym kontekście ciekawe byłoby też wspomnienie o ewentualnej możliwości zastosowania ogólniejszych transformacji falkowych (ang. *wavelets*).

7. Wzór (1), czy bias b_i musi posiadać indeks i, tzn. czy jest różny dla każdej wagi? Czy właściwe jest stwierdzenie „In neurons, information is represented as a bias"? Z kolei już na rys 2.2 widoczna jest jedna wartość parametru „b". Jaka jest więc rola biasu w perceptronie? Bardzo dziwna wersja tego wzoru znajduje się również w linii nr 2 na stronie 22.

8. Jaką informację przenosi ilustracja na rys. 2.3? W tym samym rozdziale jest też bardzo uproszczona i powierzchowna próba wytłumaczenia procesu uczenia poprzez propagację wsteczną. Czy wszystkie rodzaje sieci neuronowych uczone są w ten sposób? Z kolei brak definicji funkcji kosztu (ang. *loss function*), o której mowa na str. 15.

9. Co to znaczy, że klasyfikator jest „all at once" (str. 15)? Czy SVM z metodą optymalizacji SMO można uznać za „all at once"?

10. W opisie rozdz. 2.5.2 jako „*classical convolutions*" można by raczej wymienić pierwsze prace autorstwa LeCun'a z lat dziewięćdziesiątych.

11. W pracy Autor dość często powtarza tę samą informację w różnych akapitach, np. na str. 32 „LSTM is designed ...", a już w kolejnym to samo, tylko nieco inaczej „The LSTM was created ...", z takimi samymi informacjami i odwołaniami [18, 19], itd. Sytuacja ta powtarza się również w innych częściach rozprawy np. str. 31 w odniesieniu do BPTT, czy też akapity „People also try..." na str. 41 i zaraz potem „Additionally, some people..." na str. 42.

12. Niejasna fraza na str. 33 „also suggested by the authors" – nie wiadomo jakich autorów miał Doktorant na myśli.

13. Znacznie lepszego wprowadzenie i wyjaśnienia wymaga notacja służąca do opisu struktur sieci neuronowych, wprowadzona na str. 51 i stosowana aż do rozdz. 5, po czym zaniechana właśnie w rozdz. 5. Zaskoczonemu takim formalnym zapisem czytelnikowi rozprawy Doktorant dopiero na str. 56 wspomina, że jest to formalny system reprezentacji sieci zaproponowany przez Pana Profesora W. Skarbka w pracy „*Symbolic tensor neural networks for digital media*", natomiast pewne wyjaśnienia symboli zamieszczone zostały dopiero na str. 64. Praca zyskałaby jednak na wartości gdyby Doktorant przybliżył ten formalny system zapisu również w rozprawie.

14. Proces augmentacji danych, zarówno wizyjnych jak i dźwiękowych, poprzez losowe przycinanie (ang. *random slicing*) nie jest w rozprawie dobrze opisany ani wyjaśniony (wspomniany w opisie

Table 5.1). Nieco lepsze wytłumaczenie znajduje się wyłącznie w artykule Autora w czasopiśmie Sensors, tym niemniej temat ten powinien zostać uzupełniony.

15. W opisie prac eksperymentalnych w rozdz. 4 przy opisie „Training strategy" (4.3.5) nie jest opisany sposób wyboru oraz użycie walidacyjnego zbioru danych.
16. Opis oraz analiza wyników eksperymentalnych w rozdz. 4.4.2 jest bardzo powierzchowna i niedokładna („By comparing Figure 4.17 and Figure 4.16...").
17. Brakuje informacji na temat platformy implementacyjnej, jak również na temat choćby średniego czasu trenowania (szczególnie w trybie *end-to-end*), jak również odpowiedzi proponowanych systemów. Czy system może działać w trybie real-time?
18. Wiele opisów w bibliografii nie zawiera pełnych informacji gdzie dany artykuł został opublikowany, konferencja, journal, *arxiv*? Np. [12]-[16], [18]-[21], [37],[40], [48], [61]-[63], itd.

Powyższe uwagi mają charakter bardziej polemiczny i dotyczą dalszego wyjaśnienia, bądź też rozszerzenia pewnych zagadnień natury naukowej poruszonych w rozprawie.

7. Jaka jest przydatność rozprawy dla nauk technicznych?

Metody opracowane oraz opisane w rozprawie doktorskiej Pana magistra inżyniera Xin Changa mają istotne znaczenie dla nauk technicznych, zarówno w aspekcie osiągnięć teoretycznych, jak i możliwości aplikacyjnych. Wartość ta jest pochodną głównych i samodzielnych osiągnięć Autora, o których była mowa w poprzednich punktach tej recenzji. W szczególności, podkreślić należy główne osiągnięcie, czyli opracowanie metody rozpoznawania emocji na podstawie analizy sygnałów audio oraz wideo za pomocą oryginalnej architektury wielomodalnych sieci rezydualnych MRPN. Charakteryzuje się ona dużą dokładnością odpowiedzi porównywalną ze światowymi rozwiązaniami *state-of-the-art*, jak również potencjalnie może być użyta do łączenia innych typów sygnałów.

Opracowane metody mogą znaleźć liczne zastosowania we wszelkiego typu systemach komputerowych wymagających rozpoznawania emocji, systemach interakcji człowiek-komputer, grach komputerowych, jak również ogólnie pojętej komputerowej analizie zachowania człowieka w różnych sytuacjach.

Wnioski końcowe

Podsumowując stwierdzam, że postawione w rozprawie zadania badawcze zostały rozwiązane. Uzyskane rezultaty stanowią oryginalny własny wkład Autora rozprawy Pana mgra inżyniera Xin Changa w rozwój dyscypliny naukowej Informatyka Techniczna i Telekomunikacja. Uzyskane wyniki świadczą, że Pan Xin Chang wykazał się dogłębną znajomością najnowszych metod i algorytmów uczenia maszynowego, sztucznej inteligencji ML/AI, metod przetwarzania obrazów i sygnałów audio, jak również dobrym opanowaniem warsztatu badawczego, a w rezultacie również wysoką dojrzałością naukową.

Recenzowaną pracę oceniam jako **spełniającą z nadmiarem** wymagania stawiane rozprawom doktorskim. Wnioskuję o jej przyjęcie oraz o dopuszczenie Pana magistra inżyniera Xin Changa do publicznej obrony.

WE...t Kancelaria
wpłynęło dnia 28.02.22r.
numer .....

Prof. dr hab. inż. Konrad Wojciechowski
Wydział Informatyki
Polsko-Japońska Akademia Technik Komputerowych

Bytom 20.02.2022

Recenzja rozprawy doktorskiej Pana mgr inż. Xin Chang
Human Emotion Recognition from Image and Speech using Deep Neural Networks

Recenzowana rozprawa liczy 87 stron i napisana jest w języku angielskim. Rozprawa podzielona jest na 6 rozdziałów zatytułowanych kolejno: *Introduction, Artificial Neural Network, Literature Review, Proposed Methods, Multi-modal Residual Perceptron Ntwork for AVER, Conclusion*. Rozprawę rozpoczynają wystarczająco szczegółowe dla zrozumienia treści rozprawy streszczenia w języku polskim i angielskim. Rozprawę kończą: *List of Symbols and Abbreviations, List of Figures, List of Tables, References, List of Xin Chang Publikations*. Podsumowując w ujęciu formalnym rozprawa jest ustrukturyzowana jednak z wartości niektórych rozdziałów są nieistotne z punktu widzenia całości rozprawy a niektóre powinny być poszerzone.

Ludzie komunikują sobie wzajemnie swój stan emocjonalny przez obraz/mimikę twarzy i głos. Warto podkreślić, że dla człowieka rozpoznanie stanu emocjonalnego na podstawie obrazu twarzy i głosu nie jest proste i jednoznaczne o czym świadczą wyniki z publikacji *CREMA-D Crowd-sourced Emotional Multimodal Actors Dataset*. Uzyskana z badań *croud sourcing* poprawna klasyfikacja stanów: neutralny, radość, gniew, wstręt, strach, smutek, wynosiła 58.2% w przypadku emocji odczytywanych z obrazu twarzy, 40.9% w przypadku emocji odczytywanych z głosu i 63.6% w przypadku emocji odczytywanych z twarzy i głosu łącznie. Tworząc system automatycznego rozpoznawanie stanu emocjonalnego człowieka jako próbę odtworzenia funkcjonalności mózgu człowieka należy brać przedstawione wyniki pod uwagę.

Rozprawa doktorska, jako element kariery naukowej ma pokazać, że osoba ubiegająca się o uzyskanie stopnia doktora umie poprawnie sformułować problem naukowy, postawić istotne dla tego problemu hipotezy, sformułować tezy a następnie stosując metody naukowe rozwiązać problem to jest zweryfikować hipotezy i dowieść prawdziwości postawionych tez.

W przypadku rozprawy teza/tezy to możliwość podniesienia względem wyników literaturowych wyrażonej przez *accuracy* jakości rozpoznawania emocji w systemach FER, SER i AVER. Oceniam, że tak sformułowana teza została potwierdzona w odpowiednich fragmentach rozprawy. Aczkolwiek jak piszę dalej dla jakości rozprawy korzystne byłoby ograniczenie się do systemu AVER co wynika z tytułu rozprawy i co Doktorant podsumował w konkluzji końcowej (rozdział 8).

Rozprawa doktorska jako dzieło powinna być skonstruowana jako system logicznie powiązanych fragmentów niezbędnych dla wykazania prawdziwości tez. Tak nie jest w przypadku recenzowanej rozprawy. O ile można zgodzić się, z tym, że niezależnie badane jest rozpoznawanie emocji na podstawie danych wizyjnych i audio to zbędne z punktu widzenia całości określonej przez tytuł jest porównywanie podejścia klasycznego do rozpoznawania emocji na podstawie obrazu twarzy z podejściem opartym na sieciach neuronowych.

W przypadku recenzowanej rozprawy widzę co najmniej dwa wystarczająco obszerne problemy naukowe które mogły by być tematami oddzielnych rozpraw: i) rozpoznawanie emocji na podstawie wizji w wersjach pojedynczego obrazu i sekwencji obrazów, ii) rozpoznawanie emocji na podstawie głosu w wersjach bez i z wstępnym przetwarzaniem, iii)rozpoznawanie emocji na podstawie

synchronicznych danych wideo i audio. Dodatkowo ograniczenie się tylko do wybranego problemu pozwoliłoby na jego pogłębienie względem publikacji na których rozprawa jest oparta.

Przyczyną redakcyjnych usterek rozprawy doktorskiej jest wybór formuły w której została zrealizowana. Doktorant po: iada 6 dobrych publikacji z których 5 ostatnich (odnosząc się do kolejności z wykazu na str. 87 rozprawy) jest spójnych tematycznie z zawartością rozprawy a tylko jedna z tytułem rozprawy. Dorobek Doktoranta uporządkowany chronologicznie to publikacje: 1) Rafał Pilarczyk, Xin Chang, Władysław Skarbek: Human Face Expressions from Images: 2d Face Geometry and 3d Face Local Motion versus Deep Neural Features, 2) Xin Chang and Władysław Skarbek: Multi-Modal Residual Perceptron Network for Audio–Video Emotion Recognition (Sensors), 3) Xin Chang, Władysław Skarbek: Multi-modal Residual Perceptron Network for Audio-Video Emotion Recognition (MDPI) zapoznałem się szczegółowo. Pierwsza z publikacji pokazuje, na podstawie testów numerycznych, że rozpoznawanie emocji obrazów twarzy z wykorzystaniem CNN zapewnia wyższą poprawność klasyfikacji niż podejścia klasyczne. Publikacje 2 i 3 odnoszą się do rozpoznawania emocji z równoczesnym uwzględnieniem danych wideo i audio co wymagało rozwiązania problemu fuzji danych. Istotnym nowym wynikiem jest kompozycja sieci w której na poziomie drugim zastosowano Residual Perceptron Network. Publikacje za wyjątkiem jednej są dwuautorskie i każdemu z Autorów przysługuje 50% udziału co zostało jawnie zapisane w publikacji którą przeczytałem. W tej sytuacji właściwą formą była w mojej opinii rozprawa z publikacji. Niestety Doktorant wybrał, moim zdaniem, drogę niewłaściwą kompilując rozprawę z tych publikacji. Konsekwencją wybranej formy rozprawy jest około 38% poziom autoplag atu.

Na tle wartościowego dorobku Doktoranta zawartego w analizowanych publikacjach rozprawa ma postać kompozycji słabo powiązanych ze sobą fragmentów wymienionych powyżej publikacji. Chce wyraźnie potwierdzić, że Doktorant miał jako współautor prawo do 30% wyników publikacji 1 (trzech współautorów) i po 50 % wyników (dwóch współautorów) publikacji 2, 3, jednak sposób w jaki to zrealizował jest niepoprawny pod względem logiki układu treści.

W pierwszym zdaniu rozdziału 1 Doktorant definiuje cel swoich badań jako stworzenie sieci typu „sieć w sieci" rozwiązującej problem *Audio-Video Emotion Recognition* (AVER) Jest zrozumiale, że przed fuzją danych wideo z danymi audio należało niezależnie zbadać rozpoznawanie emocji dla kanału wizyjnego na podstawie obrazu i sekwencji obrazów, (FER) i kanału audio na podstawie sygnału mowy, (SER). Dodatkowo rozpoznawanie emocji dla każdego z kanałów może być zrealizowane w podejściu konwencjonalnym (ekstrakcja wektora cech + klasyfikator) lub podejściu z wykorzystaniem CNN. Możliwe jest również podejście mieszane w którym dane surowe dla każdego z kanałów są przetwarzane wstępnie do wybranego poziomu. W rozprawie w przypadku kanału wizyjnego są to prostokąty ograniczające c)szary twarzy, czyli zastosowano płytkie ale racjonalne przetwarzanie wstępne. W przypadku kanału audio zastosowano wstępne przetwarzanie sygnału mowy do postaci *mel spectrogram*. Analiza literatury potwierdza takie podejście w przypadku klasyfikacji emocji w głosie/śpiewie jednak jego uzasadnienie nie jest tak proste jak w przypadku obszarów twarzy.

Wybrane uwagi szczegółowe

Termin wybrane został użyty dla zasygnalizowania, że większość dalszych uwag odnosi się do głównego, w mojej ocenie, wątku rozprawy to jest AVER. Pominąłem wątek w którym doktorant porównuje rozpoznawanie emocji w podejściu klasycznym z użyciem punktów charakterystycznych z podejściem z użyciem sieci neuronowych. Zapoznałem się z nim jednak i uwzględniłem w końcowej opinii o dobrej znajomości metod wizji komputerowej.

Rozpoznawanie emocji. W literaturze problemu rozpoznawania emocji na podstawie danych obrazowych występują dwa podejścia: i)klasyfikacja emocji na podstawie obrazu twarzy oraz ii)klasyfikacja emocji na podstawie spójnej czasowo sekwencji obrazów z częstotliwością rzędu 50 fps. Drugie podejście odpowiada intuicji, że emocja twarzy jest zjawiskiem przestrzenno-czasowym.

Rozwiązanie pierwszego problemu może być podstawą dla rozwiązania drugiego po warunkiem zapewnienia zachowania zależności czasowej. Przyjęta w rozprawie koncepcja polegała na wyznaczaniu wektora cech dla pojedynczego obrazu twarzy z wykorzystaniem sieci Resnet-18 i zapewnieniu korelacji czasowej tych cech z wykorzystaniem LSTM. Warto zastosować podejście w którym cechy przestrzenne i czasowe są rozdzielone a następnie integrowane czyli późne wiązanie ale w wątku rozpoznawania twarzy na podstawie sekwencji, zamiast LSTM. Podejście zaproponowane w rozprawie, polegające na wprowadzeniu sieci LSTM jest poprawne aczkolwiek możliwe są też inne. Jedno z nich pojawiło się w tabelce pod nazwą Transformer i opisie w punkcie 2.7.

Klasyfikacja. Problem rozpoznawania emocji jest problemem klasyfikacji wieloetykietowej *multi-label classification* czyli takim w którym do danego obrazu twarzy może być przyporządkowane wiele etykiet z różnymi prawdopodobieństwami. Czy zatem jest możliwe wprowadzenie miar jakości rozpoznawania emocji oceniających podobieństwa rozkładów prawdopodobieństw etykiet w klasyfikacji? Ocena wszystkich wyników klasyfikacji emocji uzyskanych w rozprawie oparta jest na mierze *accuracy* a w fragmencie AVER dodatkowo z wykorzystaniem macierzy pomyłek. Rozumie, że umożliwiło to Doktorantowi odnoszenie się do wyników literaturowych. Istnieje jednak wiele alternatywnych miar i należało je uwzględnić. Niezależnie w innych pracach Doktoranta była używana miara F1, dlaczego z niej zrezygnowano?

Sieci neuronowe. Głównym obszarem kompetencji Doktoranta są sieci neuronowe. Świadczą o tym rożne fragmenty rozprawy, niekiedy pojedyncze zdania jak np. zastosowanie kryterium *Cross Entropy* zamiast sumy kwadratów, *optymizera* AdamW zamiast Adam, sieci RP. W przypadku rozprawy sieci neuronowe stosowano docelowo do rozpoznawania emocji na podstawie wizji i audio. Wcześniej zaproponowano sieci dla rozpoznawania emocji na podstawie obrazu. Architektury CNN1, CNN2, CNN3 opisane w 4.2.2 są alternatywnymi. Istotna jest różnica w rozmiarze obrazu wejściowego. Dla CNN3 jest w pełni wystarczająca dla rozpoznawania twarzy w tym ich emocji. Sieć CNN1 akceptująca obszar twarzy o rozmiarze pikslowym 50x50 to za mało. Dla kamer stadionowych wymaga się co najmniej 80x80. Opisana dwuwariantowa koncepcja transfer learning jest poprawna. Ze względu na duże podobieństwo problemów rozpoznawania twarzy i ogólniej rozpoznawania obiektów za bardziej uzasadnioną uważam koncepcję douczania jedynie warstwy dense layers. Warto zauważyć, że Doktorant wraz z Promotorem zauważył wspomniane podobieństwa w pracy X. Chang and W. Skarbek, "From face identification to emotion recognition" Kolejna grupa sieci opracowanych przez Doktoranta uwzględniała czasowo-przestrzenną strukturę danych obrazowych z wykorzystaniem LSTM. W rozprawie eksponowana jest nowa docelowa struktura sieci użytej do rozpoznawania emocji na podstawie wideo i audio z późnym wiązaniem informacji z poszczególnych kanałów z fragmentem *Residual Perceptron* RP. Struktura RP odpowiada koncepcji uczenia się przez sieć odwzorowania z odjętą składową stałą, która jest uwzględniana/dodawana dopiero na wyjściu modułu RP. Uzyskane w testach wyniki potwierdzają przyjętą koncepcję. Nowa struktura została odniesiona do struktury przedstawionej w opiniotwórczej publikacji [97] i została opracowana i zrealizowana koncepcja porównania trzech sieci N0 N1, N2 dla problemu AVER. Struktura sieci wynika z ogólnej koncepcji i zawiera bloki realizujące przetwarzanie wstępne, wyznaczanie z użyciem Resnet cech dla poszczególnych modalności zapewnieniu ich spójności czasowej a następnie ich fuzji. Pozostaje pytanie jak głęboko powinny być wstępnie przetwarzane dane każdej z modalności przed poddaniem ich ekstrakcji cech i fuzji. Brakuje omówienia detektorów obszaru twarzy, pomimo, że w publikacjach w których Doktorant był współautorem było to badane. Praca nie jest zbyt obszerna i warto było włączyć ten opis dla kompletności rozprawy jako dzieła. Na rys 5.3 występuje wprawdzie blok *face detector* ale brak opisu. Osobnym problemem mógłby być wpływ detektorów obszaru twarzy na jakość rozpoznawania emocji. Podobnie dla audio. Oczywiście rozumie, że w jednej rozprawie nie da się nawet tylko poruszyć wszystkich problemów. W przypadku wstępnego przetwarzania sygnału mowy zdecydowano się na reprezentację częstotliwościową *mel spectrogram*. Należało wspomnieć o istnieniu innych reprezentacji zarówno w dziedzinie czasu jak i częstotliwości.

Augmentacja. W rozprawie położono silny nacisk na technikę augmentacji. Koresponduje ona z koncepcją pozyskiwania obrazów w niekontrolowanych warunkach *in wild* czyli personalizując chcemy uodpornić sieć na cechy nieistotne dla rozpoznawania emocji w obrazach lub sekwencjach. Pozwala ona zwiększyć liczbę danych w zbiorze uczącym w tym wyrównać liczebności danych w klasach. Znajomość znaczenia niezbalansowania zbiorów dla sensowności konwencjonalnych miar jakości dobrze świadczy o wiedzy ogólnej Doktoranta. W rozprawie augmentacja obrazu jest dobrze opisana i zilustrowana jakościowo brakuje jednak szczegółów jak zostały tak precyzyjnie wyznaczone przedziały zmienności dla poszczególnych parametrów, przykładowo dla rotacji +_- 8 stopni oraz z jakich rozkładów określonych na tych przedziałach są losowane wartości parametrów. Przykładowo może z jednostajnego. Idąc, już chyba za daleko, czy rozkłady dla poszczególnych parametrów były niezależne oraz czy badano zależność jakości klasyfikacji od przedziałów i rozkładów przyjętych dla augmentacji. Augmentacja strumienia o; isana jest dość ogólnikowo. Proszę o rozwinięcie sformułowania *"some random factors are applied to the whole frames of file instead of to the individual frames separately"* Domyślam się, że przykładowo wszystkie ramki sekwencji były obracane o ten sam kąt, ale proszę o potwierdzenie lub zaprzeczenie. Możliwe są bowiem inne bardziej zaawansowane podejścia oparte na augmentacji ruchu kamery nagrywającej sekwencję, przykładowo najazd kamery ze stałą prędkością. Nie znalazłem żadnej informacji o augmentacji danych audio. Jednak fragment: *This subsection discusses the benefits of our proposed framework and time-dependent augmentation. For this aim, two datasets, RAVDESS and Crema-d, are used*. Brak jest wyjaśnienia jak była realizowana czasowo przestrzenna augmentacja i których danych dotyczyła. W przypadku braku augmentacji danych audio czy do augmentowanych sekwencji wideo przyporządkowuje się ten sam fragment audio?

Bazy danych. Wszystkie testy zrealizowane w rozprawie wykorzystywały światowe ogólnie dostępne bazy danych. Doktorant wykazał się dobrą znajomością baz danych, aczkolwiek niektóre są z lat 2015 i nie wiem czy nie pojawiły się nowe lub czy nie zaktualizowano już istniejących. Podejście Doktoranta jest o uzasadnione tym, że mógł odnosić swoje wyniki do wyników innych badaczy osiąganych na tym samym zbiorze nagrań wid·o lub audio. Doktorant nie zaproponował własnej bazy jak również nie opisał metody jej tworzenia. Pierwszym krokiem w tworzeniu takiej bazy danych jest pozyskanie audio i wideo osób/aktorów odgrywających zadane emocje. W drugim kroku pozyskane nagranie wymaga adnotacji. Trywialnie adnotacją może być polecenie np. „smutek" które było wydane aktorowi podczas realizacji nagrania. W bardziej zaawansowanej wersji adnotacja tworzona jest techniką *croud sourcingu* W rozprawie pominięto te problemy. Doktorant świadomie operuje terminem *in wild* jednak nie odnosi go do baz danych na których prowadzone są testy. W przypadku rozprawy termin należy rozumieć nie tylko jako niekontrolowane warunki ale jako również niekontrolowane emocje. Dzięki takiemu doprecyzowaniu można zinterpretować uzyskany w rozprawie wynik rzędu 90% accuracy który odniesiony do danych z artykułu: *CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset* podającego rzędu 60% poprawnych rozpoznań przez ludzi jest zastanawiający. Czy oznacza to, że automat lepiej rozpoznaje emocje od człowieka, czy nauczył się jedynie rozpoznawania odgrywanych przez aktorów gestów mimicznych, zawartych w bazie, celowo nie używam terminu emocji. Sądzę, że w przypadku bazy RAVDESS rozpoznajemy właśnie klasy gestów mimicznych i fragmenty wypowiedzi aktorów opatrzone umownymi etykietami a nie prawdziwe emocje. Brałem udział w projekcie poświęconym metodzie detekcji mikroemocji związanych z kłamstwem opracowanej również jak kl: syfikacja emocji przez P. Ekmana. Można było poprosić aktora aby odegrał mimiką lub głosem "kłamstwo" Uzyskanie nagrania prawdziwego czyli sprowokowanie aktora do kłamstwa było kluczową trudnością badań. Podobnie może być z emocjami. Doktorant ma w tym zakresie nieco inną opinię "*The RAVDESS dataset's superior findings, in our view, are due to the inclusion of more crystal clear and genuine emotion information* " która w mojej opinii jest dyskusyjna.

Przykładowe drobne usterki redakcyjne.

1.Na stronie 63 w podrozdziale 5.2.1. Functional description of analyzed networks Doktorant przedstawia opis nowej opracowanej sieci. Opis zilustrowany jest rysunkiem na którym występują bloki z nazwami Resnet i SAC i dopiero na str. 69 w podrozdziale 5.3.3. *Data augmentation cannot generalize multi-modal feature patterns* dowiadujemy się, że we wszystkich eksperymentach stosowano Resnet-18 zaś jako SAC sieć LSTM i alternatywnie Transformer ale tu brak szczegółów. Świadczy to o braku uporządkowania treści.

2.Niska jakość niektórych rysunków, przykładowo 2.4, 3.8 spowodowana kopiowaniem.

3.Dwa podobne pod względem merytorycznym fragmenty napisane jeden za drugim z użyciem nieco innych słów:

a.HOG to plane mapping is defined via regression trees designed for all 68 fp68 as Figure 4.1 using cascade approach [73, 74]. The use of many small regression trees gives a more effective detector than using one large regressionmodel. The trees are built using stochastic gradient boosting of Friedman [75].

b.The HOG to plane mapping is created using regression trees constructed for every 68 fp68 as shown in Figure 4.1 utilizing a cascade method [73, 74]. The employment of a large number of tiny regression trees results in a more effective detector than the use of a single big regression tree. The trees are constructed using Friedman's stochastic gradient boosting technique [75].

4.Brak wyjaśnienia znaczenia koloru kropek na rys, 5.1.

5.Zastosowana konwencja zapisu sieci CNN autorstwa prof. W. Skarbka jest nietypowa. Należało dodać odnośnik lub wyjaśnić. Odnośnik został podany dopiero dalej przy kolejnym użyciu tej konwencji.

Podsumowanie. Istotnym wynikiem rozprawy doktorskiej jest oryginalne rozwiązanie dwóch problemów naukowych: i)rozpoznawaniu emocji widocznej na obrazie twarzy i emocji zrejestrowanej w nagraniu wideo ii)rozpoznawaniu emocji na podstawie nagrania wideo i synchronicznego nagrania audio co wymagało fuzji dwóch strumieni danych. Opracowana technika fuzji może być łatwo rozszerzona na większą liczb ę modalności.

Rozprawa doktorska prezentuje dobrą wiedzę teoretyczną Doktoranta w dyscyplinie informatyka techniczna i telekomunikacja w obszarach: i)wizji komputerowej, przetwarzania obrazu, wydobywania cech obrazowych i wzorców, ii)uczenia maszynowego w problemach klasyfikacji emocji, oraz bardzo dobrą w obszarze iii)sieci neuronowych, ich struktur i typów w szczególności dla potrzeb fuzji danych wielomodalnych. Doktorant wykazał umiejętność samodzielnego i logicznego prowadzenia pracy naukowej w poszczególnych fragmentach rozprawy.

Stwierdzam, że recenzowana rozprawa spełnia warunki stawiane rozprawom doktorskim w sformułowaniu Ustawy z dnia 20 lipca 2018 Prawo o szkolnictwie wyższym i nauce (Dz. U. 2018 poz. 1668 z późn. zm.) Wnoszę o dopuszczenie rozprawy do publicznej obrony.

K. Wiaderek

Szczecin, 18 lutego 2022r.

dr hab. inż. Paweł Forczmański, prof. ZUT
Katedra Systemów Multimedialnych
Wydział Informatyki
Zachodniopomorski Uniwersytet
Technologiczny w Szczecinie
ul. Żołnierska 52
71-220 Szczecin

RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY NAUKOWEJ DYSCYPLINY INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

Tytuł rozprawy: **Human Emotion Recognition from Image and Speech using Deep Neural Networks**

Autor rozprawy: **mgr inż. Xin Chang**

Przedmiotem niniejszej recenzji jest rozprawa doktorska zatytułowana "Human Emotion Recognition from Image and Speech using Deep Neural Ne'works", napisana w roku 2021, której autorem jest mgr inż. Xin Chang.

Promotorem pracy jest prof. dr hab. inż. Władysław Skarbek. Niniejsza recenzja została przygotowana na zlecenie prof. dr hab. inż. Michała Malinowskiego, Dziekana Wydziału Elektroniki i Technik Informacyjnych, zawarte w piśmie z dnia 20 grudnia 2021r., w związku z decyzją Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej z dnia 23 listopada 2021r.

W swojej recenzji odpowiem na następujące pytania, podzielone na dwie podstawowe grupy. W pierwszej, dotyczącej strony formalnej dysertacji, skupię się na pytaniach: czy treść pracy odpowiada tematowi określonemu w tytule, czy jej struktura i podział na rozdziały są poprawne, a także, czy zaprezentowane tezy są kompletne. W drugiej, przenalizuję merytoryczne aspekty pracy, tj. jak został sformułowany jej cel, czy i w jakim zakresie praca stanowi nowe ujęcie problemu, czy została odpowiednio umiejscowiona w dyscyplinie naukowej i czy Autor w odpowiedni sposób dokonał porównania osiągniętych rezultatów z wynikami opisywanymi w literaturze przedmiotu. W tym obszarze postaram się odpowiedzieć również na pytanie o praktyczne znaczenie uzyskanych wyników.

I. Ocena formalnej strony dysertacji

Przedstawiona do recenzji praca doktorska została wydana w formie książkowej i liczy 87 stron, z czego 73 strony to właściwa treść. Pozostała część pracy to spis oznaczeń, ilustracji, tabel i pozycji literaturowych. Praca napisana jest w języku angielskim i w większości jest to język poprawny i jednoznaczny. Strona redakcyjna rozprawy stoi na dobrym poziomie (z zastrzeżeniem uwag przedstawionych w dalszej części recenzji). Autor w swojej pracy prawidłowo prowadzi narrację i poprawnie konstruuje dowody przyjętych na początku tez. Tytuł odpowiada w znacznej mierze treści pracy, choć zauważyć należy, że zasygnalizowane rozpoznawanie dotyczy również, jak pokazano w samej

pracy, także obrazów ruchomych i w pewnej części również metod klasycznych (nie neuronowych).

Układ pracy jest poprawny i można o nim powiedzieć, że jest klasyczny. W pierwszych trzech rozdziałach Autor dokonał przeglądu literatury w obszarze sztucznych sieci neuronowych, rozpoznawania emocji na podstawie analizy obrazów statycznych, obrazów ruchomych, nagrań audio i nagrań audio-video. W dalszej części Autor dokonał pogłębionej analizy wybranych aspektów rozpoznawania emocji ludzkich na podstawie analizy danych audio-video.

Zawarty w pracy przegląd literatury jest dość obszerny i obejmuje okres od początku lat '70 XX w. do roku 2021. Pokryty zakres tematyczny jest bardzo szeroki i dobrze umiejscawia recenzowaną rozprawę. Cytowane publikacje, w liczbie 111, dotyczą zarówno klasycznych metod przetwarzania obrazów cyfrowych, metod głębokiego uczenia i opisu cech twarzy w kontekście odwzorowywanych emocji. Dodatkowo, w pracy znaleźć można odwołania do 6 publikacji Autora i Promotora.

Literatura została omówiona w kilku grupach zagadnień (sztucznych sieci neuronowych, wykrywania i rozpoznawania twarzy, klasyfikacji emocji na bazie obrazów i dźwięków). Z uwagi na oczywiste ograniczenia co do objętości, przegląd ten jest dość pobieżny, ale trzeba jednoznacznie stwierdzić, ze w dużej mierze kompletny. W pracy nie zabrakło również opisu benchmarkowych zestawów testowych, tj. FER2013, JAFFE i CK+.

Prawie połowa objętości pracy to opis autorskich metod rozpoznawania i klasyfikacji danych multimodalnych a także wyników badań eksperymentalnych na danych benchmarkowych.

Jeśli chodzi o ocenę formalną, to praca nie odbiega od standardu i ocena w tym aspekcie jest pozytywna. Autor wykazał się umiejętnością poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników w aspektach takich jak zwięzłość, poprawność redakcyjna i jasność przekazu.

2. Ocena merytoryczna

Praca dotyczy bardzo aktualnego problemu badawczego, tj. analizy emocji badanych osób wykorzystującej multi-modalne źródła danych, czyli obraz statyczny, sekwencje obrazów i nagrania audio. Jest umiejscowiona na styku widzenia komputerowego, przetwarzania sygnałów i (głębokiego) uczenia maszynowego. Jej tematyka skupia się wokół kilku związanych ze sobą zagadnień tj. klasycznych algorytmach lokalizacji punktów charakterystycznych na obrazach twarzy i powiązania ich z modelami ekspresji a także na metodach rozpoznawania emocji na bazie obrazu, sekwencji video, nagrań audio i sekwencji audio-video wykorzystujących głębokie uczenie.

W recenzowanej rozprawie można wskazać zarówno wątki teoretyczne, jak i eksperymentalne. Do pierwszej grupy można zaliczyć prezentację istniejących algorytmów i ich krytyczną analizę. Z przedstawionych w części teoretycznej wniosków wynikają kierunki dalszych prac, polegających na modyfikacji istniejących i opracowaniu nowych algorytmów. Autorskie algorytmy są w dużej mierze rozwinięciem istniejących wcześniej podejść, wykorzystujących znane koncepty.

Analizując literaturę Autor przedstawił standardowe koncepcje klasycznego i neuronowego przetwarzania danych oraz dokonał przeglądu najważniejszych metod z analizowanego obszaru badawczego. Wyszedł przy tym z założenia, że podstawą skutecznego rozpoznawania emocji na bazie obrazów jest wykorzystanie metod wykrywania i rozpoznawania twarzy ludzkich. W tym kontekście omówił wybrane zagadnienia dotyczące wykrywania punktów charakterystycznych

na twarzy, opisu mimiki twarzy, głębokich sieci neuronowych oraz uczenia transferowego.

W obszarze przetwarzania sygnałów akustycznych, Autor przyjął znane założenie, że informacja o stanie emocjonalnym może zostać wyekstrahowana z krótkoterminowego widma amplitudowego. Szkoda, że nie skupił się również na cechach przetworzonych takich jak współczynniki MFCC, czy też spektrogram MELowy.

Na bazie przeprowadzonego przeglądu literatury Autor prawidłowo zdiagnozował potencjalne problemy i możliwości, które wynikają z dotychczas stosowanych metod. Stwierdził, że największy potencjał tkwi obecnie w metodach głębokiego uczenia wykorzystujących splotowe sieci neuronowe CNN i długotrwałą pamięć krótkotrwałą LSTM.

Autor umieścił na początku pracy dość przewidywalną hipotezę badawczą mówiącą o tym, iż użycie sztucznych sieci neuronowych zamiast metod klasycznych poprawi adaptację systemów rozpoznawania emocji do nowych danych i pozwoli na zwiększenie ich odporności na warunki zewnętrzne. Tak postawiona hipoteza nie wydaje się być zbyt odważna, gdyż od dawna wiadomo, że systemy neuronowe, o ile dostarczy się im odpowiedni zbiór uczący i zastosuje adekwatną strategię uczenia, pozwalają na uzyskiwanie wyników niedostępnych dla metod klasycznych, szczególnie w kontekście dość złożonych danych multi-modalnych.

Szczegółowe cele pracy obejmują opracowanie lub też usprawnienie istniejących metod w celu zwiększenia ich skuteczności w stosunku do rozwiązań bazowych.

Po przeczytaniu rozprawy można dojść do konkluzji, że hipoteza została w sposób poprawny zweryfikowana a cele pracy zostały osiągnięte. Weryfikacja odbyła się na drodze eksperymentów numerycznych na standardowych bazach benchmarkowych.

Badania Autora zaprezentowane w dysertacji, w kontekście źródła danych, można podzielić na trzy podobszary:

1. analizę obrazów statycznych,
2. analizę obrazów ruchomych (sekwencji video),
3. analizę nagrań audio,
4. analizę nagrań audio-video (sekwencji video z synchronicznym dźwiękiem).

Jednocześnie, Autor w swojej pracy poruszył dwa wątki badawcze zorientowane na odmienne podejścia algorytmiczne, tj.

1. metody tzw. konwencjonalne,
2. metody neuronowe (w tym głębokie).

W dalszej części recenzji odniosę się do głównych osiągnięć Autora, pogrupowanych zgodnie z ww. wykazem. W tym kontekście w dysertacji zidentyfikować można opis pięciu różnych podejść.

Pierwsze z nich (Facial Emotion Recognition – FER) dotyczy konwencjonalnego algorytmu wykrywania punktów charakterystycznych twarzy na obrazie 2D i ich mapowania na trójwymiarowy model. Detektor punktów bazuje na bibliotece dlib i zwraca pozycje (x,y) 68 punktów (landmarków) na twarzy. Są one rzutowane na trójwymiarowy model 3D (Candide-3) i w dalszym kroku klasyfikowane zgodnie z systemem FACS za pomocą metody SVM. W dalszym kroku, Autor zaproponował zastąpienie klasyfikatora SVM klasyfikatorem neuronowym o niewielkim stopniu złożoności. Uzyskane rezultaty przemawiają na korzyść użycia klasyfikatora neuronowego. Rozumiem, że opis tej metody był konieczny jako przeciwwaga (i baza do

porównań) do opisanych metod neuronowych, gdyż sama metoda nie jest rozwiązaniem finalnym.

Drugie podejście to typowy mechanizm end-to-end, gdzie na wejściu do sieci splotowej podawany jest obraz wyekstrahowanej twarzy, natomiast w wyjściowej warstwie typu fully connected zwracana jest wartość odpowiadająca zidentyfikowanej klasie (zgodnie z systemem FACS). Opisane w tej części zostały trzy dość proste architektury sieci splotowych pracujące na obrazach o rozmiarach od 50x50 do 150x150 pikseli. Autor pokazał, że testowane sieci oferują skuteczność znacznie przewyższającą podejście konwencjonalne (ponad dwukrotnie większą), szczególnie w przypadku, gdy na wejściu pojawiają się twarze o znacznych zmianach orientacji przestrzennej. Opis tego podejścia był zapewne motywowany potrzebą pokazania tzw. baseline, gdyż użycie go w niekontrolowanych warunkach otoczenia może nie dawać satysfakcjonujących efektów.

W ramach tego samego podejścia, Autor zaproponował zastosowanie strategii uczenia transferowego (transfer learning), aby usprawnić uczenie ekstraktora cech. Wykorzystał do tego architekturę VGG-16 dodaną na początku potoku przetwarzania. W ten sposób dokonał zmiany dziedziny problemu z zadania wykrywania twarzy na zadanie klasyfikacji emocji. Powstała sieć, która pozwala na uzyskanie skuteczności porównywalnej do znanych metod, bez konieczności przeprowadzania żmudnego procesu uczenia wszystkich warstw. Jest to potwierdzenie znanego faktu, że wybrane architektury splotowe nauczone na obszernych zbiorach treningowych doskonale radzą sobie jako ekstraktory cech.

Trzecie podejście dotyczy rozpoznawania emocji na podstawie analizy strumienia video (Video Emotion Recognition - VER). Jak słusznie zauważył Autor, zastosowanie w tym przypadku klasyfikatora pracującego na pojedynczych klatkach (bez kontekstu czasowego) prowadzi do powstawania wielu błędów. Dlatego, jak pokazuje literatura w tym obszarze, lepiej jest stosować podejścia wykorzystujące dynamiczne sieci głębokie. W tym przypadku Autor zasugerował użycie architektury CNN+LSTM, gdzie początkowe warstwy splotowe wykorzystują sieć ResNet. Część pracy poświęcona temu podejściu jest bardzo skromna, ale wynika to zapewne z faktu, iż podobne podejścia są dokładniej opisane w dalszej jej części.

Czwarte podejście dotyczy rozpoznawania emocji na postawie analizy wyłącznie sygnału akustycznego (Speech Emotion Recognition - SER). W tym celu zaadoptowano architekturę ResNet+LSTM pracująca na spektrogramach sygnału mowy ludzkiej. Opis jest również bardzo skromny i dotyczy właściwie samej koncepcji (dość oczywistej) a nie konkretnych badań.

Ostatnie podejście, najbardziej rozbudowane zarówno w kontekście złożoności architektonicznej, jak i samego opisu, zostało umieszczone w końcowej części pracy. Widoczne jest, że jest ono najbardziej dojrzałe i zostało stworzone, aby poradzić sobie z problemami zidentyfikowanymi na wcześniejszych etapach prac badawczych. Podejście to dotyczy jednoczesnego przetwarzania danych wizualnych (sekwencji video) i danych akustycznych (nagrań audio) w celu wykrywania stanu emocjonalnego osoby badanej, czyli tzw. Audio-Video Emotion Recognition (AVER).

Autor zaproponował trzy architektury uwzględniające fuzje danych audio i video. Pierwsza architektura, nazwana N_0 pozwala na późną fuzję (konkatenację) danych, będących wynikiem niezależnej analizy audio i video. Druga architektura, nazwana N_1, rozszerza podejście N_0 o dodatkowe połączenia w dalszych warstwach dla niezależnych potoków audio i video. W tym podejściu fuzji podlega również proces obliczania funkcji kosztu. Ostatnia architektura, N_2, wykorzystuje koncepcję podobną do N_1, jednak jest bardziej złożona, m.in. poprzez wykorzystanie koncepcji połączeń rezydualnych (ResNet). Sieć N_2 nazwana została Multi-Modal Residual Perception Network (MRPN) i można ją uznać za najbardziej znaczące naukowe osiągnięcie Autora.

Poza opisem proponowanych metod praca zawiera także wyniki przeprowadzonych testów i porównań z innymi metodami na powszechnie uznanych danych benchmarkowych, które obejmują, poza wcześniej wspomnianymi bazami unimodalnymi FER2013, JAFFE i CK+, także multimodalne bazy RAVDESS i Crema-d. Opracowane podejścia wykazały się wysoką skutecznością, porównywalną, a często przewyższającą metody state-of-the-art.

W mojej opinii bardzo duże znaczenie ma przedstawiony dorobek eksperymentalny. Potwierdza on dobrą skuteczność opracowanych podejść, a odpowiednio dobrany zestaw danych testowych potwierdza wysoką wiarygodność uzyskanych wyników.

Wartościowe wnioski wyciągnięte przez Autora i kończące pracę obejmują kilka istotnych obserwacji:

1. Brak dostępu do odpowiedniej jakości danych obniża jakość procesu uczenia i efektywność działania systemów ER, tak więc użycie odpowiednich metod augmentacji danych pozwoliłoby na dalsze zwiększenie ich skuteczności;
2. Nie należy rozwijać wyłącznie metod multimodalnych, takich jak AVER, gdyż postęp w dziedzinie systemów uni-modalnych (ER) pozwala na wprowadzenie dalszych usprawnień w tych pierwszych;
3. Fuzja danych na dalszych etapach przetwarzania nie zawsze daje oczekiwane efekty, ale duże możliwości tworzenia głębokich architektur mogą tę sytuacje zmieniać, szczególnie w kontekście innych modów (także innych sposobów ekspresji).

Reasumując, najbardziej wartościowym osiągnięciem przedstawionym w rozprawie jest zaproponowany algorytm MRPN. Wynika to z oryginalnego podejścia sposobu przekazywanych informacji na kolejnych etapach uczenia sieci. Osiągnięcie to zostało zweryfikowane na wymagających danych benchmarkowych RAVDESS i Crema-d. Osiągnięcia poboczne, uzyskane niejako „po drodze", to oryginalne architektury dla problemów FER, VER i SER.

Wartym wspomnienia jest fakt, iż podczas swojej pracy naukowej, Autor opublikował (we współautorstwie z Promotorem) 6 artykułów bezpośrednio związanych z tematyką dysertacji. Pięć z nich zostało zamieszczone zostało w materiałach znaczącej konferencji z serii SPIE - International Society for Optics and Photonics, a jeden w czasopiśmie MDPI Sensors.

Jeśli chodzi o ocenę merytoryczną, to uzyskane wyniki można uznać za znaczące i konkurencyjne w stosunku do rozwiązań opisywanych w literaturze, tak więc ocena w tym aspekcie jest pozytywna.

3. Uwagi o charakterze dyskusyjnym i uwagi krytyczne

Przedstawiona dysertacja jest dowodem na rzetelnie wykonaną pracą koncepcyjną i eksperymentalną. Jednak, jak w każdej niemalże publikacji naukowej, tak i w recenzowanej pracy można znaleźć pewne błędy i niedociągnięcia. Moje uwagi podzieliłem na dwie grupy, odpowiadające kwestiom merytorycznym i redakcyjnym.

Uwagi merytoryczne

1. Brak jest szerszej dyskusji dotyczącej hiperparametrów odpowiedzialnych za proces uczenia sieci głębokich, szczególnie w aspekcie rozdzielczości obrazów wejściowych, częstotliwości próbkowania audio, liczby filtrów w warstwach splotowych i liczby neuronów w warstwach LSTM, a także użytych optymalizatorów. Rozumiem, że tworzenie modeli głębokich to bardziej „art" niż „science", ale taka dyskusja pokazałaby, że obrane rozwiązania nie są przypadkowe.
2. Jako, że mamy do czynienia z pracą naukową, oczekiwałbym więcej szczegółów dotyczących złożoności obliczeniowej opracowanych metod.
3. Wyniki przeprowadzonych badań eksperymentalnych pokazują, że potencjał wdrożeniowy opracowanych metod jest bardzo wysoki co jest dodatkowym, wartościowym i oryginalnym osiągnięciem twórczym. W dysertacji nie można niestety znaleźć informacji na temat czasu tworzenia modeli wykorzystujących opracowane metody. Wydaje się, że dokładniejszy opis tych elementów pozwoliłyby na jeszcze trafniejszą ocenę możliwości praktycznego zastosowania opracowanych metod.
4. Wprawdzie metody bazujące na detekcji punktów charakterystycznych opisane w dysertacji okazały się być niekonkurencyjne w stosunku do rozwiązań neuronowych end-to-end, jednak dla porządku powinno się wspomnieć o nowych rozwiązaniach z tego obszaru, oferowanych przez modele HRNet, BlazeFace i frameworki takie jak MediaPipe. Możliwe, że pełniejszy opis geometrycznych aspektów mimiki twarzy pozwoliłby na zwiększenie skuteczności klasyfikacji emocji.

Uwagi redakcyjne

1. Praca napisana jest w większości poprawnie, jednak znaleźć w niej można pewną liczbę błędów edycyjnych, tzw. literówek - poniżej lista tych, które udało mi się zauważyć:
    1. Str. 15: 'wrt' -> 'with',
    2. Str. 16: 'Figure ??' – brak automatycznego indeksu,
    3. Str. 16: błędnie dodane 'i' w funkcji sumowania,
    4. Str. 17: 'closed to zero' -> 'close to zero',
    5. Str. 24: 'Equation refbptt' – brak automatycznego indeksu,
    6. Str. 31: 'Figure reffigureauexample' – brak automatycznego indeksu
    7. Str. 38: 'neural' -> 'neutral' (to samo na Str. 42)
    8. Str. 41" 'peroformance' -> 'performance'
    9. Str. 54: 'pixel by frame' -> 'frame by frame'?
2. Symbole stosowane do opisu architektury sieci na Str. 44 są wyjaśnione dopiero na Str. 48.
3. Na Str. 27, na początku podrozdziału 3.1.1 - pierwsze dwa akapity są bardzo podobne, wręcz można je uznać za powtórzenie (z niewielkimi zmianami językowymi). Ta sama sytuacja ma miejsce na Str. 30 (podrozdział 3.2.1), na Str. 34 (ostatnie dwa akapity podrozdziału u 3.4) i na Str. 36 (czwarty i piąty akapit podrozdziału 4.1). Sytuacja ta wynika prawdopodobnie z

braku finalnej weryfikacji manuskryptu przed jego opublikowaniem)

4. Z uwagi na mały format publikacyjny, czytelność niektórych ilustracji jest bardzo niska, szczególnie dotyczy to ilustracji 4.19, 4.20, 5.4 oraz macierzy konfuzji 5.6 i 5.7. Utrudnia to ich dokładniejszą analizę.

Powyższe uwagi nie zmieniają jednoznacznie pozytywnego odbioru recenzowanej pracy, pozostawiając jednak pewien niedosyt w kontekście staranności jej przygotowania

4. Możliwość zastosowania uzyskanych wyników

Metody wykrywania emocji na obrazach twarzy i w strumieniach audio-video są bardzo złożonymi algorytmami. Istnieje wiele rozwiązań w tej dziedzinie, które rozwijane są od wielu już lat. Jak pokazuje przegląd literatury, część z elementarnych problemów nie stanowi już wyzwania naukowego. Z drugiej strony, Autor pokazał w swojej pracy, że uzyskanie bardzo wysokiej skuteczności jest cały czas problematyczne. Wobec dynamicznego rozwoju technik widzenia komputerowego, powstawanie prac w tym zakresie ma fundamentalne znaczenie dla współczesnych nauk technicznych, zwłaszcza w kontekście komunikacji człowiek-komputer. Praca nie zawiera niestety jawnie pokazanych możliwości praktycznego zastosowania opracowanych algorytmów, a wydają się one dość ciekawe. Wdrożenie przedstawionych algorytmów w urządzeniach codziennego użytku dać może impuls do rozwoju w całkiem nowych obszarach.

5. Wnioski końcowe

Podsumowując recenzję, mogę jednoznacznie stwierdzić, że mimo pewnych uwag dyskusyjnych o niewielkim znaczeniu, przedstawiona rozprawa mgra inż. Xin Changa spełnia wymagania ustawowe stawiane rozprawom doktorskim i wnioskuję o dopuszczenie Autora do dalszych etapów przewodu doktorskiego.