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 zawartości 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 posiada 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 obszary 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 opisana 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 wideo 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 klasyfikacja 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 liczł 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. Wojciechowski