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 Networks", 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.