

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 mgr inżyniera Xin Changa zatytułowanej „Human emotion recognition from image and speech using deep neural networks”. Praca powstała na Wydziale Elektroniki i Technik Informatycznych 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 zorganizowana 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 przez 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 torry 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.

Uwzględnivszy 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-doświadczałny.

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 przekonujący?

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 spłotowych 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 dziedzinę 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 wielosensorycznych, 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 rezyduального (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 uzupełniania 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 (rozd. 4.1), porównanie rozwiązań starszych opartych na cechach eksperckich z rozwiązaniami bazującymi na głębokich sieciach neuronowych (rozd. 4.2), analiza transferu wiedzy (rozd. 4.3), jak również metodyka rozpoznawania emocji w strumieniach wideo i z wykorzystaniem głębokich sieci neuronowych typu CNN oraz LSTM (rozd. 4.4) oraz zaproponowane wykorzystanie cech pochodzących ze spektrogramów sygnałów, gdzie jednym z głównych problemów jest zapewnienie synchronizacji zmiennych czasowo sekwencji wideo oraz audio (rozd. 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 przekonującego 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 przez brak, czy też 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 uwagi (*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 przynosi 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. Zaskoczonymu 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. Wnioskuje o jej przyjęcie oraz o dopuszczenie Pana magistra inżyniera Xin Changa do publicznej obrony.