RADA NAUKOWA DYSCYPLINY
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

zaprasza na

PUBLICZNĄ OBRONĘ ROZPRAWY DOKTORSKIEJ

mgr inż. Weroniki Gutfeter

która odbędzie się w dniu **18 maja 2023 roku,** o godzinie **9:15** w trybie hybrydowym

Temat rozprawy:

„Identyfikacja twarzy na podstawie obrazów wieloujęciowych z zastosowaniem głębokich sieci agregujących”

Promotor: prof. dr hab. inż. Andrzej Pacut – Politechnika Warszawska

Recenzenci: prof. dr hab. inż. Khalid Saeed – Politechnika Białostocka

prof. dr hab. inż. Krzysztof Ślot – Politechnika Łódzka

Obrona odbędzie się **w sali nr 116** Gmachu Elektroniki Politechniki Warszawskiej. Osoby zainteresowane uczestnictwem zdalnym na platformie **MS Teams** proszone są o zgłoszenie chęci uczestnictwa w formie elektronicznej na adres sekretarza komisji: dr hab. inż. Jacka Misiurewicza, – email: jacek.misiurewicz@pw.edu.pl do dnia 17 maja 2023 godz. 17:00.

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-do-30-kwietnia-2019-r/Dyscyplina-informatyka-techniczna-i-telekomunikacja-dziedzina-nauk-inzynieryjno-technicznych/mgr-inz.Weronika-Gutfeter

Przewodniczący Rady Naukowej Dyscypliny
Informatyka Techniczna i Telekomunikacja
Politechniki Warszawskiej
dr hab. inż. Jarosław Arabas, prof. uczelni

Streszczenie

W pracy zaproponowano metody rozwiązania problemu tworzenia wzorców biometrycznych twarzy na podstawie zdjęć wieloujęciowych. Przez zdjęcia wieloujęciowe rozumiane są zbiory obrazów twarzy jednej osoby, w których głowa jest ujęta pod różnymi kątami, przy czym zestaw kątów jest zdefiniowany arbitralnie. Przykładem tego typu danych są zdjęcia sygnalityczne wykonywane przez fotografów policyjnych oraz obrazy uzyskane z modeli 3D głowy poprzez ich projekcję na płaszczyznę. Potrzeba rozwiązania tak zdefiniowanego problemu została umotywowana obserwacjami rzeczywistych systemów identyfikacji - zauważono, że wiele z algorytmów biometrii twarzy jest zoptymalizowana pod kątem scenariusza identyfikacji na podstawie zdjęć frontalnych i nie wykorzystuje dostatecznie ujęć skrajnych, takich jak zdjęcia profilowe.

W ramach badań zanalizowano oraz zaimplementowano szereg podejść do identyfikacji zbiorów obrazów twarzy, w których występują zdjęcia niefrontalne, w tym poprzez agregację deskryptorów pojedynczych ujęć (wzorce wytworzone przez modele single-view), poprzez modele agregujące (wzorce wytworzone przez modele multi-view) oraz automatyczną korektę pozy w obrazie. Na potrzeby eksperymentów zaadaptowano modele agregujące stosowane dotąd do klasyfikacji obiektów 3D, w tym model MVCNN *(multi-view convolutional network*) oraz model RotationNet. W kolejnym kroku zaproponowano własny model agregujący z mechanizmem uwagi o nazwie SygnaT. W pracy pokazano, że wszystkie 3 modele agregujące (multi-view) uzyskały lepsze wyniki niż agregacja modeli pojedynczych (single-view). Najwyższe współczynniki dokładności identyfikacji uzyskano dla modelu SygnaT. Była to różnica o ponad 6% dla współczynnika Rank-1 w porównaniu z uśrednionymi deskryptorami single-view oraz o 18% dla współczynnika Rank-1 wyliczonego dla identyfikacji na podstawie zdjęć profilowych.

Wszystkie rozważane podejścia były oparte o głębokie sieci splotowe i zakładały budowę modułową. Jako rdzeń analizowanych rozwiązań przyjęto ten sam enkoder single-view oparty na architekturze ResNet-50 i wytrenowany na bazie VGGFace2 przez autorów pracy. Przyjęcie wspólnego rdzenia ułatwiło miarodajne zestawienie różnych metod ze sobą, a przez modularną strukturę pokazano, że ta część sieci może być modyfikowana lub zamieniona na inną wg preferencji projektantów systemu biometrycznego. Przy omawianiu wyników wzięto pod uwagę różne scenariusze identyfikacji twarzy, w tym również w scenariuszu ze zbiorem otwartym, w którym możliwa jest rejestracja nowych osób bez konieczności ponownego treningu modeli.

Słowa kluczowe:

identyfikacja twarzy, biometria, sieci głębokie, sieci agregujące, sieci splotowe, MVCNN, RotationNet, multi-head attention, atencja, rozpoznawanie twarzy 3D, fotografia sygnalityczna

Abstract

The main goal of the thesis is to solve the problem of face identification in the multi-view face images. Datasets that consist of multi-view face images can be obtained in various scenarios. One of the typical usage is the police booking photography, where the specialists are obligated to acquire a set of face images from strictly defined angles for each suspect.

Another scenario is also one of the methods employed in the 3D face recognition. As the 3D models are quite heavy and computationally demanding, a typical solution to that problem is to transform 3D data into a set of 2D images showing the object from different views. This approach can be named as multi-view object classification.

After preliminary experiments we observed that most of the state-of-the-art algorithms are optimized to recognize frontal images and there is a significant drop of accuracy when identification is made on sets containing extreme views like full profile pictures. In our work we conducted a survey of methods that can be used for creating biometric templates from image collections with non-frontal faces. We analyzed various approaches like 3D face alignment, single-view templates aggregation and multi-view networks. For our experiments we adapt methods that have been proposed for multi-view object recognition, namely MVCNN and RotationNet. However we should emphasize that there exists a difference in these two domains. Multi-view object recognition is typically based on a closed dictionary of labels and objects are represented by a relatively large number of views while faces are organized in the sets of three or five probes and the possibility of new identifier registration is crucial for the system. For the purpose of the experiments it was necessary to collect face data that is structured in the following manner and to define adequate testing scenarios. We also introduce a new multi-view model in which we apply multi-head attention for data flow aggregation. We named it SygnaT. SygnaT solves some of the limitations of the previous models, as it does not require a strict number of probes or specific order in the view set.

All of the proposed multi-view networks gain higher results than the single-view approaches. For the SygnaT network the Rank-1 accuracy is higher by 6% in a full identification test and by 18% in a test on profile pictures.

Each of the analyzed solutions was built using deep convolutional networks and there is a common backbone for all of them. The backbone is implemented with ResNet-50 architecture. Therefore weights can be transferred from the baseline single-view model which was trained on VGGFace2 to make all the models more unified and easily comparable. We showed that the multi-view networks work in various face identification scenarios, we tested all solutions in closed and open-set tests. Moreover we conducted an experiment with SygnaT on an unstructured dataset (face recognition benchmark IJB-C) and proved that also in this configuration it is better than the single-view model approach.

Keywords:

face biometrics, facial identification, deep learning, multi-view classification, multi-view networks, convolutional neural networks, MVCNN, RotationNet, multi-head attention, self-attention, face 3D recognition, police photography

Prof. dr hab. inż. Khalid Saeed
Wydział Informatyki
Politechnika Białostocka
ul. Wiejska 45A, 15-351 Białystok
Tel. (+48-85) 746 9196
k.saeed@pb.edu.pl

Białystok, 6.03.2023 r.

RECENZJA rozprawy doktorskiej

mgr inż. Weroniki Gutfeter

z Wydziału Elektroniki i Technik Informacyjnych
Politechniki Warszawskiej

zatytułowanej "Identyfikacja twarzy na podstawie obrazów wieloujęciowych z zastosowaniem głębokich sieci agregujących"

Promotor:

Prof. dr hab. inż. Andrzej Pacut
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Niniejszą recenzję przygotowałem na zlecenie zawarte w piśmie, które otrzymałem od Profesora Jarosława Arabasa przewodniczącego Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej. Recenzję dokonałem na podstawie dostarczonej mi rozprawy doktorskiej dnia 9.01.2023.

I. Omówienie zawartości rozprawy

Głównym celem pracy doktorskiej mgr inż. Weroniki Gutfeter było zastosowanie metod klasyfikacji obiektów wieloujęciowych do identyfikacji twarzy człowieka jako cechy biometrycznej. Innowacyjne podejścia polegały na opracowaniu algorytmów pozwalających rozwiązać problem identyfikacji twarzy na podstawie zdjęć wieloujęciowych, gdzie obrazy twarzy jednej osoby były zbierane pod różnymi kątami głowy. Doktorantka zaadoptowała metody stosowane dotąd do klasyfikacji obiektów 3D. Zagadnienie to zostało jasno sformułowane przez autorkę pracy, a trzy odpowiednie tezy, które starała się udowodnić, zostały opracowane poprawnie. Rozprawa ma charakter

statystyczno-doświadczalny i zawiera 130 stron. Składa się z pięciu rozdziałów oraz materiałów dodatkowych i bibliografii.

Rozdział pierwszy poświęcony jest przeglądowi wybranych zagadnień związanych z problemem rozpoznawania twarzy. Wybrana literatura do analizy stanu wiedzy jest poprawna, pomimo niedosytu aktualnych pozycji.

W rozdziale drugim autorka przedstawiła metody zastosowania sieci głębokich w identyfikacji twarzy człowieka. Tu szczegółowo pokazane są sposoby generacji wzorców twarzy z użyciem sieci splotowych oraz podstawowy model jednoujęciowy będący wyjściowym punktem do prac autorki nad identyfikacją twarzy w systemach obrazów wieloujęciowych. W rozdziale tym przedstawiono analizę aktualnego stanu wiedzy dotyczącego metodologii i podejść w kontekście obrazów wieloujęciowych. Autorka zbadała i omówiła wpływ pozycji głowy na dokładność systemu identyfikacji twarzy na podstawie modelu jednego ujęcia (single-view model).

W rozdziale trzecim rozprawy autorka przedstawia wybrane podejścia stosowane w identyfikacji obrazów wieloujęciowych twarzy. Doktorantka zajmuje się implementacją bardziej skomplikowanych struktur i modułów do agregacji.

W rozdziale czwartym autorka przedstawia szczegółowy opis modeli własnych do agregacji obrazów wieloujęciowych twarzy. Doktorantka proponuje własny model agregujący SygnaT wykorzystujący mechanizm *atencji* w strukturze sieci. Mechanizm ten został zaproponowany w pracy *„Attention is all you need"* opublikowanej w Curran Associates w 2017, gdzie autorzy zastosowali go do tłumaczenia języka naturalnego.

Część merytoryczną zamykają wnioski końcowe w rozdziale piątym - *Podsumowanie* oraz Dodatek - *Materiały dodatkowe*. Całość pracy zamyka literatura - *Bibliografia*, która zawiera 73 pozycje, a 3 z nich są współautorstwa doktorantki. Niestety, wśród 73 wybranych pozycji bibliografii, tylko 8 datowanych jest 2020 lub nowsze wydanie.

II. Opinia i ogólna ocena pracy

Rozprawa doktorska mgr inż. Weroniki Gutfeter pt. "*Identyfikacja twarzy na podstawie obrazów wieloujęciowych z zastosowaniem głębokich sieci agregujących*" stanowi oryginalne rozwiązanie problemu oraz wkład własny doktorantki w rozwój metod identyfikacji twarzy. Autorka wykonała założone cele pracy, zaprezentowała, rozwiązała problem identyfikacji twarzy na podstawie obrazów wieloujęciowych - implementacja algorytmów, rysunki i tabele zostały odpowiednio opracowane. Przykładowe podejścia innych autorów podobnej tematyki zostały wybrane umiejętnie. Doktorantka opracowała systemy identyfikacji twarzy, w których wzorce twarzy są rejestrowane na podstawie zbioru obrazów wieloujęciowych i pokazała, że ten system jest lepszy niż system oparty na zdjęciach frontalnych. Opracowała również system identyfikacji

twarzy wykorzystujący modele wieloujęciowe (multi-view) i pokazała, że wskaźniki jakości identyfikacji nie są gorsze niż system oparty o modele jednoujęciowe (single-view). Udowodniła eksperymentalnie i przedstawiła statystycznie, że modele agregujące stosowane dotąd do rozpoznawania obiektów 3D można dostosować do zadania identyfikacji twarzy.

Reasumując należy uznać, że autorka wykazała w swojej rozprawie doktorskiej dobrą znajomość technik identyfikacji twarzy. Z sukcesem zrealizowała postawione zadania pokazując eksperymentalnie wartość proponowanych metod i modeli. Pracę napisała poprawnie, zaś istotne dla tematyki pracy zagadnienia zostały omówione przez autorkę wyczerpująco.

Według mojej oceny uważam, że Pani mgr inż. Weronika Guttfeter osiągnęła wyznaczony cel rozprawy doktorskiej, który wnosi nowe aspekty do nauk technicznych w zakresie informatyki. Wymieniona rozprawa spełnia warunki ustawy o stopniach naukowych.

Uwagi merytoryczne oraz usterki w pisowni pracy

Jak nadmieniłem powyżej, praca napisana jest rzetelnie, ale praktycznie niemożliwe do uniknięcia są drobne usterki merytoryczne lub błędy edycyjne, których przykłady zostały podane poniżej:

Uwagi merytoryczne

- Ogólny stan wiedzy powinien być umieszczony w całości w jednym rozdziale na początku rozprawy.

- Brakuje ogólnego opisu zawartości rozdziałów rozprawy. Doktorantka powinna podać strukturę pracy na początku rozprawy i omówić zawartość każdego rozdziału i pokazać lukę badawczą w literaturze. Autorka taką strukturę pokazała w podsumowaniu. Z drugiej strony, podsumowanie powinno zawierać konkluzje - jakie problemy zostały rozwiązane, opis wyników, a szczególnie na ile jej oryginalne rozwiązania są lepsze od istniejących podejść. Przydatna byłaby analiza porównawcza pod kątem wyników pracy ze stanem wiedzy.

- Rysunek 2.7: W przypadku obrotów o kąty ±60°, ±30°, ±15° wyniki można uznać za "symetryczne". W przypadku obrotu osób o kąt -90° obserwuje się większy błąd niż w przypadku obrotu o kąt + 90°. Nie wyjaśniono dlaczego.

- Rysunek 3.9: Doktorantka opisuje: "*Na wykresie 3.9 można zobaczyć że dokładność systemu rosła wraz ze wzrostem liczby próbek w zestawie, co jest zgodne z intuicją*". Wyraźny skok dokładności widoczny jest w początkowych zmianach liczby podpowiedzi. Dla prawej części wykresu oraz zmiany liczby z 11 do 12 widzimy "nasycenie" oraz marginalne zmiany. Oznacza to, że wniosek autorki nie jest do końca właściwy. Proszę o wyjaśnienie.

- Rysunek 4.3: Rysunek przedstawia miary klasyfikacji, które nie były wcześniej omówione tj. dokładność top-1. Warto to wyjaśnić.

- Tabela 4.4: Przedstawiona tabela budzi pewne wątpliwości - różnice między otrzymanymi wynikami są stosunkowo niewielkie (szczególnie warianty SygnaT (token) oraz SygnaT (avg). Prosiłbym o wyjaśnienie.

Usterki w pisowni

- Streszczenie, jeśli ma być abstraktem (tak jak dalej podano po angielsku) nie powinno zawierać cytatów.

- Rys. 2.3: Doktorantka nie wyjaśniła „współczynnika dokładności". Najprawdopodobniej jest to miara skuteczności w procesie nauki, ale powinno się to zdefiniować.

- Rysunki 2.8 oraz 2.9 pokazują metryki z wykorzystaniem miary podobieństwa cosinusowego. Natomiast wskaźnik kolorów dobrany został dość pobieżnie. Przede wszystkim kolory dobierane są w zakresie od 0.0 do 0.9 (górna granica raczej powinna wynosić 1.0), a liczba przedziałów to 15. W związku z tym podział kolorów nie odpowiada liniowo wartościom. Rysunki zyskałyby na czytelności, gdyby zastosowano 10 kolorów ze skokiem 0.1 lub normalizację koloru do wartości obserwowanych.

- Rysunek 3.1 zastosowanie linii łączących "bloki" w diagramie sugeruje występowanie pewnych zależności (np. kolejności) pomiędzy nimi, ale chyba tak nie jest. Proszę o wyjaśnienie.

- Na stronie 63 czytamy "*Warto jednak zauważyć, że metody tego typu powodują duży narzut na obliczenia i tworzenie bazy wzorców. Wiele z nich jest zoptymalizowanych pod kątem wizualnie akceptowalnych wyników, a nie przenoszenia cech potrzebnych do identyfikacji - następuje wygładzenie obrazu w czasie projekcji i jak zauważają autorzy [10] utrata istotnych cech o wysokiej częstotliwości*". Jest to bardzo dobry wniosek podsumowujący metodę, ale nie jest we właściwym miejscu. Sugerowałbym przeniesienie go do kolejnego podrozdziału pt. "Podsumowanie eksperymentów z metodami automatycznej korekcji pozy".

III. Merytoryczne osiągnięcia doktorantki oraz jej publikacje

Pani mgr inż. Weronika Gutfeter jest współautorem 2 referatów konferencyjnych oraz 1. artykułu opublikowanego w czasopiśmie IEEE Intelligent Systems (140 punktów na liście Ministerstwa), co świadczy o znaczeniu osiągniętych wyników pracy naukowej doktorantki. Brakuje jednak procentowego udziału doktorantki w najważniejszej publikacji wieloautorskiej. Oświadczenie autorki byłoby całkowicie wystarczające. Na podkreślenie zasługuje również fakt, że doktorantka ma aktywną działalność naukowo-projektową. Wybrane projekty badawcze z udziałem doktorantki jako współwykonawczyni w ramach jej pracy jako asystent naukowy w Instytucie Badawczym NASK to:

– APAKT: automatyczna analiza zagrożeń w sieci, w tym treści pokazujących wykorzystanie seksualne dzieci.

- BIOWIZ: system komputerowy wspierający funkcjonariuszy policji i innych służb mundurowych w identyfikacji osób na podstawie wizerunku utrwalonego na zdjęciu lub materiałach wideo.

- BioWalidator: walidacja fotografii twarzy do dowodów osobistych i innych dokumentów tożsamości.

IV. Wnioski końcowe

Wystawiam ocenę pozytywną rozprawie doktorskiej mgr inż. Weroniki Gutfeter pt. "*Identyfikacja twarzy na podstawie obrazów wieloujęciowych z zastosowaniem głębokich sieci agregujących*" oraz stwierdzam, że praca spełnia wymagania i warunki nakładane przez ustawę o stopniach naukowych. Rozprawa stanowi oryginalne rozwiązanie problemu naukowego oraz osobisty wkład Doktorantki w rozwój metod identyfikacji twarzy człowieka z obrazów wieloujęciowych.

Na tej podstawie wnioskuję o dopuszczenie Autorki wymienionej rozprawy doktorskiej do jej obrony w celu uzyskania stopnia doktora nauk technicznych w dziedzinie nauk inżynieryjno-technicznych w dyscyplinie informatyka techniczna i telekomunikacja.

Khalid Saeed

Łódź, 28.02.2023

Prof. dr hab. inż. **Krzysztof Ślot**
Instytut Informatyki Stosowanej
Politechnika Łódzka

Recenzja rozprawy doktorskiej Pani mgr inż.
Weroniki Gutfeter
pt.

Identyfikacja twarzy na podstawie obrazów wieloujęciowych z zastosowaniem głębokich sieci agregujących

1. Tematyka i struktura rozprawy

Przedstawiona do recenzji rozprawa doktorska dotyczy tematyki biometrycznego rozpoznawania twarzy, a szczególnym obszarem tej obszernej dziedziny, któremu poświęcone są prace Doktorantki jest problem optymalnego wykorzystania informacji zawartej w tzw. obrazach 'wieloujęciowych' - różniących się pozą zdjęciach tej osoby, będących w dyspozycji algorytmu analizy. Podjęty przez Doktorantkę wątek prac jest ważny i aktualny – poprawna identyfikacja lub weryfikacja twarzy ma ogromne znaczenie praktyczne, a istniejące podejścia do realizacji tego zadania w żaden usystematyzowany sposób nie próbują wykorzystać informacji zawartej w wieloelementowych zbiorach przykładów danej kategorii. Podjęty przez Doktorantkę problem jest zarazem nietrywialny, a jego rozwiązanie wymaga wykazania się znajomością zaawansowanych metod analizy danych, posiadaniem eksperckiego rozeznania w obszarze tematycznym biometrii twarzy oraz pomysłowością i kreatywnością. Obszar merytoryczny badań wymaganych do osiągnięcia zakładanych celów rozprawy lokuje się bez wątpienia w obrębie dyscypliny naukowej informatyka techniczna i telekomunikacja, w której prowadzony jest przewód Doktorantki.

Zawartość rozprawy odzwierciedla chronologię prac Doktorantki w podjętym obszarze. Obszerna jej część zawiera bardzo kompetentny, szeroki i wnikliwy przegląd rozmaitych podejść do problemu uwzględnienia wpływu pozy na poprawność rozpoznawania, uzupełniony prezentacją dobrze dobranego materiału eksperymentalnego (baz danych) stanowiącego podstawę dla weryfikacji efektów prac, a także wskazaniem metryk, używanych w ilościowej ocenie wyników. Aby uzyskać obiektywny pogląd co do istotności różnych zaproponowanych koncepcji, Doktorantka dokonuje doświadczalnej weryfikacji jakości wybranych metod, reprezentatywnych dla różnych nurtów prac, korzystając z dostępnych procedur lub dokonując własnych implementacji tych procedur. Metody te są uszeregowane według rosnącego poziomu złożoności, począwszy od

narzucających się prostych koncepcji na rozwiązanie postawionego zadania, do dużo bardziej skomplikowanych algorytmów. Ostatnia część pracy poświęcona jest prezentacji wkładu własnego Doktorantki, która przedstawia odpowiednie adaptacje metod opracowanych dla rozwiązywania innych kategorii problemów, uzyskując satysfakcjonujące rezultaty przeprowadzanych analiz. Metodyka prezentowanych prac nie budzi zastrzeżeń – proponowane koncepcje są dobrze uzasadnione, podlegają starannej weryfikacji eksperymentalnej, podsumowanej z wykorzystaniem właściwych miar ilościowych i zwieńczonych jakościową dyskusją otrzymanych rezultatów.

2. Cele i tezy rozprawy

Celem badań Doktorantki jest opracowanie strategii rozpoznawania biometrycznego, pozwalającej na redukcję wrażliwości analizy na potencjalnie dowolną pozę rejestrowanej twarzy poprzez optymalne wykorzystanie informacji zawartych w dostępnych zbiorach ujęć twarzy danej osoby. Powodem podjęcia przez Nią prac w rozważanym obszarze były względy praktyczne – sformułowanie odpowiedniej metody pozwoliłoby na zwiększenie poprawności rozpoznawania osób w systemach policyjnej analizy materiału pozyskanego z monitoringu przestrzeni publicznych, dla których to osób dysponuje się zestawem kilku ujęć twarzy, różniących się pozą (kątem obserwacji twarzy przez kamerę o osi optycznej skierowanej prostopadle do osi twarzy), określanych w pracy terminem ‘obrazów wieloujęciowych’. Konsekwencje osiągnięcia założonego celu wykraczają oczywiście poza wspomniane konkretne zastosowanie i stanowią rozwiązanie o ogólnej przydatności.

Doktorantka formułuje trzy tezy, wykazanie słuszności których określa zakres jej prac:

1. Możliwe jest zbudowanie systemu identyfikacji twarzy, który będzie uwzględniał informacje o zmienności przestrzennej twarzy i różnicy w wyglądzie w zależności od pozy bez konieczności zastosowania specjalizowanych urządzeń do akwizycji, jak skanery 3D. Można pokazać, że system, w którym wzorce twarzy są rejestrowane na podstawie zbioru obrazów wieloujęciowych będzie lepszy niż system oparty na zdjęciach frontalnych.

2. Zbiory wieloujęciowe obrazów twarzy wymagają dedykowanych rozwiązań, a składanie informacji z poszczególnych próbek można zrealizować za pomocą modułu agregującego zintegrowanego w strukturze sieci neuronowej w postaci tzw. modelu agregującego multi-view. System identyfikacji wykorzystujący modele multi-view powinien wykazać wskaźniki jakości identyfikacji nie gorsze niż system oparty o modele jednoujęciowe tzw. single-view

3. Modele agregujące stosowane dotąd do rozpoznawania obiektów 3D należących do zamkniętego słownika obiektów (takie jak MVCNN i RotationNet) można dostosować do zadania identyfikacji biometrycznej twarzy.

3. Merytoryczna ocena pracy

Przedstawiona rozprawa ma charakter eksperymentalny – Doktorantka formułuje heurystyczne hipotezy, które następnie, korzystając z adekwatnych zbiorów danych, poddaje weryfikacji eksperymentalnej. Punktem wyjścia dla prac Doktorantki jest ocena poprawności rozpoznawania twarzy w warunkach nienadzorowanej akwizycji obrazów, z użyciem najlepiej spisujących się obecnie algorytmów bazujących na głębokich sieciach konwolucyjncych i z zastosowaniem standardowej strategii uczenia, gdzie obrazy twarzy dla docelowych kategorii osób są używane w

procedurze dostrajania parametrów wstępnie wytrenowanej architektury. Jako narzędzie przeprowadzenia analizy stosuje pretrenowaną na bazie zdjęć VGGFace2 głęboką sieć neuronową ResNet50, powszechnie uznawaną za jedną z najlepszych architektur do klasyfikacji obrazów. Aby umożliwić klasyfikację próbek kategorii nieznanych algorytmowi w fazie uczenia, a więc, przygotować się do budowy algorytmu klasyfikacji działającego na zbiorze ‘otwartym’, Doktorantka przyjmuje jako podstawę klasyfikacji podobieństwo kosinusowe wektorów generowanych przez przedostatnią warstwę sieci ResNet. W efekcie przeprowadzonych, bardzo wartościowych w mojej opinii eksperymentów, Doktorantka wskazuje na istotne upośledzenie wyników analizy spowodowane zmianami pozy obserwowanej twarzy. Dodatkowo, Doktorantka wskazuje na możliwość osiągnięcia nieznacznej poprawy wyników poprzez włączenie do zbiorów treningowego i testowego obrazów wieloujęciowych, wykazując tym samym potrzebę podjęcia prac nad próbą optymalnego wykorzystania zawartych w nich informacji. Sieć wykorzystana przez Doktorantkę w celu oceny wpływu pozy na jakość rozpoznawania została przez Nią użyta w dalszych pracach jako model odniesienia dla analiz porównawczych oraz jako komponent proponowanych przez Nią, bardziej złożonych architektur, służący do ekstrakcji reprezentacji obrazu twarzy.

Sformułowanie własnych koncepcji rozpoznawania biometrycznego twarzy o zredukowanej wrażliwości na zmienność pozy zostało poprzedzone, jak wcześniej wspomniano, eksperymentalną oceną jakości działania wybranych, zaproponowanych w literaturze strategii wykorzystania materiału wieloujęciowego, uzupełnioną kilkoma autorskimi pomysłami i podsumowaną ciekawą dyskusją uzyskanych wyników. Efektem tych prac jest wskazanie najbardziej obiecującego kierunku badawczego – modyfikacji metod zakładających użycie głębokich modeli agregujących informacje pochodzące z wielu widoków twarzy tej samej osoby, który stał się głównym obszarem zainteresowania Doktorantki. W ostatniej części rozprawy Doktorantka kolejno opisuje swoje pomysły na adaptację trzech metod rozpoznawania, z których dwie: ‘wielowidokowa’ sieć konwolucyjna (Multi-View Convolutional Neural Network – MCNN) i sieć ‘rotacyjna’ (RotationNet), stanowiły architektury zaproponowane do rozpoznawania obiektów innych niż twarze, na podstawie informacji zawartej w kilku dostępnych widokach tych obiektów, zaś trzecia stanowi adaptację metody ‘transformerów’ opracowanej dla przetwarzania języka naturalnego.

3.1. Weryfikacja istniejących podejść do klasyfikacji zdjęć wieloujęciowych

Pierwszy grupa metod zmierzających do zwiększenia poprawności rozpoznawania twarzy o dowolnej pozie, której reprezentatywny przykład został wdrożony i zweryfikowany przez Doktorantkę, bazuje na ciekawej koncepcji wykorzystania posiadanego zbioru wieloujęciowego do budowy kompletnego modelu 3D, z którego następnie można generować dowolny widok referencyjny, dopasowany do pozy analizowanego zdjęcia. W efekcie przeprowadzenia szeregu wartościowych eksperymentów, Doktorantka negatywnie weryfikuje biometryczną przydatność tego podejścia, jednocześnie formułując szereg istotnych wskazówek, których uwzględnienie w konstrukcji algorytmu może poprawić jego atrakcyjność dla innych potencjalnych zastosowań.

Kolejną testowaną przez Nią koncepcją jest pomysł agregacji deskryptorów obrazów różnych widoków twarzy, zmierzającej do uzyskania reprezentacji integrującej wiedzę o niezależnym od kąta obserwacji wyglądzie twarzy. Doktorantka rozważa dwa zaproponowane podejścia do agregacji: dokonywaną na poziomie deskryptorów widoków i dokonywaną na poziomie wyników indywidualnych porównań widoków z nieznaną próbką. Doktorantka rozważa kilka prostych metod

agregacji (uśrednienie, mediana, a w przypadku fuzji dla wyznaczanych odległości – średnia, odległość minimalna/maksymalna, soft-min). Doktorantka podsumowuje swoje doświadczenia konkluzją o niewielkim lub żadnym zysku płynącym z przedstawionych metod agregacji dokonywanej w odniesieniu do próbek galeryjnych (tworzących bazę wiedzy o klasach). Ten wniosek nie wydaje się być niezgodny z intuicją – raczej hipoteza, że uśrednianie deskryptorów (podobnie jak selekcja wartości minimalnych lub maksymalnych dla odpowiadających sobie pozycji) lub wyników porównań powinno zwiększyć poprawność klasyfikacji, jest w moim odczuciu niepoprawna. Dobry deskryptor odzwierciedla wiele aspektów treści zawartej w danych wejściowych, ale założenie, że informacje o tych aspektach są ‘rozplątane’ i mają ciągłe i skupione rozkłady dla każdej z cech (wtedy proponowane agregacje mają sens), jest w moim odczuciu kompletnie nieuzasadnione. Oprócz agregacji dokonywanej po stronie próbek galeryjnych, Doktorantka sprawdza również efekty agregacji próbek zapytań (w przypadku nagrań z monitoringu, prawdopodobne jest posiadanie wielu ujęć nieznanej osoby, więc taki zabieg jest jak najbardziej zasadny). Tym razem, zgodnie z intuicją uzyskuje znaczącą poprawę wskaźników rozpoznawania, przy czym przeprowadzone testy dotyczyły jedynie jednego scenariusza, w którym klasa była reprezentowana przez reprezentację odpowiadającą zdjęciu frontalnemu, a zapytanie było uśrednionym wektorem dla różnych ujęć danej twarzy ze zbioru testowego. Wyniki eksperymentów wykorzystujących wiele ujęć osób ze zbioru testowego były bardzo obiecujące, szkoda, że nie przeprowadzono sprawdzeń dla innych scenariuszy – być może okazałoby się, że dokonywanie rozważanych przez Doktorantkę sposobów agregacji nie przynosi poprawy w porównaniu z klasyfikacją w schemacie ‘najbliższych sąsiadów’, gdzie najbliższej pary próbek poszukiwano by w zbiorze galeryjnym i zbiorze zapytań.

Ponieważ przedstawione przez Doktorantkę wątki porównań wektorów (bez lub z agregacją różnymi sposobami) to użycie prostych i znanych idei klasyfikacji minimalnoodległościowej, naturalnym rozszerzeniem listy scenariuszy klasyfikacji może być klasyfikacja w schemacie k-NN lub klasyfikacja metodą najbliższej średniej, ale z uwzględnieniem odległości Mahalanobisa, czyli oceną rozrzutów w obrębie klasy galeryjnej (aby zachować prostotę obliczeniową, z użyciem np. diagonalnej macierzy kowariancji). Doktorantka nie podejmuje jednak tego tropu – być może z uwagi na intuicyjnie bardziej obiecujące możliwości oferowane przez metody agregacji bazującej na treningu, oferowanej przez modele agregujące.

Modele agregujące (ich istotą jest integracja uczenia modułu agregacji z procesem treningu klasyfikatora) rozważone w pracy Doktorantki – MVCNN i RotationNet, zostały zidentyfikowane przez Nią jako najbardziej obiecująca ścieżka klasyfikacji obrazów wieloujęciowych. Znowu, jest to zgodne z intuicją – zastąpienie ‘ręcznie’ dobieranych reguł wyznaczania reprezentacji danych przez metody uczenia stoi u podstaw sukcesów uczenia głębokiego. Weryfikacja poprawności rozpoznawania obrazów wieloujęciowych, która wymagała od Doktorantki przygotowania odpowiedniego materiału eksperymentalnego, została przeprowadzona, podobnie jak wszystkie wcześniejsze eksperymenty, w sposób poprawny metodycznie i podsumowana sformułowaniem uprawnionych wniosków.

3.2. Prace Doktorantki w zakresie adaptacji architektur MVCNN i RotationNet

Osiągnięcia wskazane przez Doktorantkę jako oryginalne, własne koncepcje w obszarze biometrycznej analizy twarzy to głębokie modele agregujące, stanowiące modyfikacje oryginalnych koncepcji przetwarzania obrazów wieloujęciowych. Dwa pierwsze z nich, to architektury

obliczeniowe stanowiące modyfikacje sieci MVCNN i RotationNet, nazwane przez Doktorantkę odpowiednio: MVCNN-Sygnalityka i RotationNet-Sygnalityka. Pierwszym elementem nowości, który Doktorantka wprowadza do bazowych architektur, jest ich przystosowanie do radzenia sobie w warunkach klasyfikacji na zbiorach ‘otwartych’, czyli zawierających klasy, które nie były używane w treningu klasyfikatora. Taki scenariusz klasyfikacji warunkuje praktyczne znaczenie metody, dlatego też przydatność zaproponowanego rozwiązania jest oczywista, chociaż zastosowana przez Doktorantkę strategia osiągnięcia tego celu: oparcie klasyfikacji na podobieństwie wektorów referencyjnego i badanego, generowanego przez wybrane warstwy gęstej sieci, nie jest nowa [1]. Druga modyfikacja, wprowadzona do metody MVCNN, dotyczy zastąpienia agregacji deskryptorów różnych widoków, dokonywanej w metodzie oryginalnej przez wybór maksymalnych wartości odpowiadających sobie komponentów wektorów składowych (max-pooling), przez uśrednienie tych elementów (average pooling). Chociaż zaproponowana, alternatywna strategia daje lepsze efekty, brak jakiejkolwiek analizy uzasadniającej to podejście (z pełną świadomością, że takiej analizy być może nie da się przeprowadzić), obniża rangę zaproponowanego pomysłu. Wreszcie, trzecim elementem autorskiej modyfikacji algorytmu treningu jest wykorzystanie modułu ekstraktora cech obrazu z parametrami pretrenowanymi na bazie twarzy (Doktorantka pokazuje, że daje to lepsze wyniki niż inicjalizacja losowa lub inicjalizacja w wyniku treningu na bazie zawierających inne niż twarze kategorie obrazów). Wprowadzenie rozważanego ulepszenia wydaje się być wskazaniem pewnej dobrej praktyki niż istotnym wkładem do budowy algorytmów klasyfikacji. W odniesieniu do drugiego z modeli: RotationNet-Sygnalityka, Doktorantka proponuje dopuszczenie losowej kolejności prezentacji sieci widoków, co jest nowością, ale wiąże się ze zwiększeniem złożoności obliczeniowej procedury.

3.2. Wykorzystanie transformera w analizie obrazów wieloujęciowych

Najciekawszym i najbardziej wartościowym, z punktu widzenia oceny dorobku Doktorantki, fragmentem rozprawy jest jej ostatnia część, która prezentuje koncepcję wykorzystania modułu kodującego transformera jako narzędzia realizacji zadania klasyfikacji obrazów wieloujęciowych.

Sednem pomysłu Doktorantki jest dostrzeżenie analogii występujących między problem analizy obrazów wieloujęciowych, których istotą jest reprezentacja trójwymiarowego obiektu poprzez agregację informacji cząstkowych, zawartych w dwuwymiarowych rzutach, z problemem analizy języka naturalnego, gdzie treść zdania wynika z agregacji jego komponentów. W efekcie, dla rozwiązania zadania rozpoznawania twarzy decyduje się użyć wspomnianej wcześniej koncepcji transformera, którą adaptuje do specyfiki rozwiązywanego zadania. Adaptacja obejmuje określenie organizacji procesu przetwarzania, niezbędnego dla realizacji postawionego celu, jak również nowe propozycje rozwiązań szczegółowych. Jako architekturę zapewniającą generację dyskryminatywnej, zagregowanej informacji o wyglądzie twarzy Doktorantka wskazuje część kodującą algorytmu transformera. Jako efekt transformacji danych wejściowych przez proponowany algorytm, Doktorantka przyjmuje dwie alternatywne reprezentacje. Pierwsza (nazywana przez Doktorankę SygnaT-token), to używany również w oryginalnej koncepcji wynik transformacji ‘tokenu’ pomocniczego (określanego typowo jako ‘CLS’), uzyskiwany w wyniku liniowej agregacji informacji o przetwarzanych obrazach, poddanej następnie nieliniowemu przekształceniu w wielowarstwowej sieci gęstej. Druga, to autorski pomysł wykorzystania reprezentacji generowanych w wyniku transformacji tokenów odpowiadających elementom sekwencji wejściowej, które po uśrednieniu są również argumentem nieliniowej transformacji w sieci gęstej (podejście, nazywane przez Doktorantkę SygnaT-avg). Aby nauczyć zaproponowaną

architekturę poprawnej reprezentacji informacji o wyglądzie osoby, Doktorantka dokonuje wstępnego treningu inspirowanego algorytmem BERT, sterowanego kryteriami poprawności klasyfikacji i poprawności predykcji reprezentacji brakujących komponentów zbioru wieloujęciowego. W zaprezentowanych scenariuszach uczenia (wstępnego i zasadniczego), Doktorantka proponuje potraktowanie sekwencji testowej jako następnika sekwencji galeryjnej (oddzielonego standardowym tokenem wejściowym ‘SEP’).

Zaproponowana metoda rozpoznawania wieloujęciowych zdjęć twarzy z użyciem transformera stanowi najskuteczniejsze (ze wszystkich rozważanych przez Doktorantkę sposobów) rozwiązanie problemu, a osiągane efekty są zdecydowanie konkurencyjne względem istniejących standardów dziedziny. Na szczególne podkreślenie zasługuje skuteczność autorskiej propozycji reprezentacji wyniku (‘SygnaT-avg’), przewyższająca efekty zastosowania podejścia wykorzystującego mechanizm oryginalnej koncepcji.

3.3. Podsumowanie oceny merytorycznej prac Doktorantki

Niewątpliwą zaletą przedstawionych w rozprawie prac jest gruntowna weryfikacja stosowanych obecnie podejść do problemu klasyfikacji zdjęć wieloujęciowych. Implementacja metod, obfitość scenariuszy eksperymentalnych i obszerność podsumowania wyników stanowią bardzo wartościowy efekt prac Doktorantki. Kluczowym elementem w ocenie wartości prac z perspektywy wymagań przewodu jest jednak ocena nowatorskiego wkładu Doktorantki do dziedziny. O ile w odniesieniu do zaproponowanych przez Nią ulepszeń algorytmów MVCNN i MV-RotationNet analizy obrazów wieloujęciowych, jej wkład w mojej opinii nie jest znaczący, o tyle zdecydowanie wartościowym, oryginalnym efektem jej badań jest propozycja metody analizy bazująca na koncepcji transformera, pozwalająca na uzyskanie poprawności analizy przewyższającej istniejące podejścia, co stanowi osiągnięcie o istotnym znaczeniu naukowym i praktycznym.

Na zakończenie opinii, chciałbym podjąć wątek alternatywnej i nie podjętej przez Doktorantkę strategii uczenia (określanej przez Nią jako ‘uczenie metryczne’, choć w literaturze znanej raczej jako tzw. meta-learning). Schemat ‘meta-learningu’ wydaje się lepiej pasować do podjętego problemu (zadanie to dopasować próbkę galeryjną do próbki testowej), niż zastosowany przez Doktorantkę schemat uczenia modeli konkretnych kategorii i oczekiwanie, że wyuczona reprezentacja będzie dyskryminatywna dla próbek wcześniej nieznanych. W pracy zabrakło mi próby konfrontacji obydwu podejść: proponowanego przez Doktorantkę oraz meta-learningu, dedykowanego dla przypadku testowania na zbiorach otwartych. Echa problemu, który w mojej opinii jest konsekwencją wybranej przez Doktorantkę drogi, odzywają się w wynikach eksperymentu podsumowanego na rys. 4.2d (wraz z postępem uczenia maleje poprawność klasyfikacji na zbiorze otwartym). Nie jest to niezrozumiałe – sieć podczas treningu zaczyna powoli specjalizować się w rozpoznawaniu klas pochodzących z zamkniętego zbioru treningowego, co pogarsza poprawność rozpoznawania dla klas nieznanych. Mimo, że jako przyczynę problemu, Doktorantka wskazuje zbyt dużą liczbę modyfikowanych parametrów w stosunku do posiadanego zbioru przykładów i proponuje użycie typowej dla transferu wiedzy metody „zamrażania” części wag, co przynosi częściową poprawę, wydaje mi się, że głównym źródłem problemu jest wybrana strategia uczenia. Wydaje się, (co wymagałoby oczywiście sprawdzenia), że zaobserwowane zjawisko mogłoby nie mieć miejsca gdyby użyto schematu ‘meta-learningu’, koncentrującego się na wyszukiwaniu różnic i podobieństw między niewielkimi zbiorami galeryjnymi i zapytaniami.

4. Uwagi szczegółowe

Praca jest napisana starannie, ale znajduje się w niej pewna liczba fragmentów, które są niejasne i wymagają doprecyzowania, które są polemiczne, wreszcie, które są w mojej opinii niepoprawne. Poniżej prezentuję ich listę, precyzując w każdym przypadku rodzaj formułowanego zastrzeżenia.

str. 35: „Przez głębokie sieci neuronowe najczęściej rozumiemy perceptrony wielowarstwowe, …: nie rozumiem

str. 36: „Zestawienie ze sobą warstw splotowych i skalujących przyczynia się do niewrażliwości modelu na niewielkie przesunięcia obiektów w obrazie.” - wyjaśnienie wprowadza w błąd: warstwa konwolucyjna zapewnia inwariantność względem translacji. To o co chodziło Autorce to zapewne selekcja najlepszego dopasowania filtru do treści obrazu w regionie określonym przez rozmiar okna decymacji.

str. 36: „dobrze sprawdzają się przy przetwarzaniu danych o strukturze siatki ...” to jest żargon, nie ma czegoś takiego, są dane określone na dyskretnej przestrzeni 2D

str. 37: Po co zwrot ‘pewnego rodzaju’ w zdaniu „Sam algorytm uczenia opiera się na pewnego rodzaju optymalizacji.”?

str. 40: Doktorantka definiuje odległość kosinusową pomijając wyjaśnienia używanej notacji. Jak rozumiem, symbole Ya i Yb oznaczają wektory reprezentacji. Jeśli tak, to przedstawiony iloczyn to chyba iloczyn skalarny, ale jeden z argumentów powinien być tu transponowany. Co więcej, Doktorantka twierdzi, że wartość odległości będzie zawarta w przedziale [0..1], czego nie rozumiem: kosinus zmienia się w zakresie od -1 do 1, więc skąd ten przedział?

str.49: W nagłówku Tabeli 2.2 pojawia się niezdefiniowany akronim TAR, który prawdopodobnie powinien mieć postać TPR.

str. 56: nie rozumiem sposobu ‘grafowego’ uporządkowania metod redukcji wrażliwości klasyfikacji na zmiany pozy, przedstawionych na rys. 3.1 – czy poprzez zastosowanie połączeń Autorka chciała uwypuklić bliskość koncepcyjną par metod? Chyba nie, bo podane podejścia można by ułożyć w dość dowolnej kolejności.

str. 91. Doktorantka mogłaby pomóc w wyjaśnieniu istoty metody ‘RotationNet-Sygnalityka’ poprzez odpowiednie wsparcie opisu tekstowego dobrze wyjaśnionym materiałem graficznym. Pewnie rys. 4.2 ma taki potencjał, ale niestety, trzeba się mocno domyślać znaczenia użytych tam oznaczeń i stylów. Co oznaczają litery ‘z’ i litery ‘i’ z indeksami – czy to są wyniki predykcji dla hipotezy danego widoku?. Czy dwa zacienione w różny sposób prostokąty, stanowią symboliczną reprezentację wektorów hipotez uzyskanych dla różnych widoków, gdzie węższy prostokąt, odpowiadający symbolowi ‘i’ oznacza ‘incorrect view’?. Co podlega agregacji? - prawdopodobnie, różne widoki.

str. 95: niezrozumiałe zdanie: ‘Wybór sekwencji jest wynika z z minimalizacji wartości neuronów kodujących ujęcie”

str. 103 „oraz zestaw par wektorów kluczy K i wektorów wartości W w wektor wyjściowy.” Zamiast symbolu W powinien być symbol V.

str. 106: „W eksperymentach używano modułu MHA złożonego z ośmiu równoległych modułów atencji tzw. głów, a znaczy to, że każdy z tokenów wejściowych jest cięty przed wejściem do modułu na osiem równych części.” - to stwierdzenie jest zaskakujące. Token to reprezentacja elementu sekwencji, zaś istota ‘wielogłowości’ to próba znalezienia szeregu alternatywnych kontekstów dla danej sekwencji. ‘Cięcie’ tokenów na kawałki to rezygnacja z wielowymiarowej reprezentacji danych wejściowych. W pracach dotyczących transformerów tokeny wejściowe są podawane równolegle na wejścia każdej z przetwarzających głów, a to co jest decymowane w stopniu proporcjonalnym do liczby głów, to rozmiary wektorów V, z których składane są wektory stanu lub wektory kontekstu, generowane przez warstwy gęste (bo wynik jest konkatenacją wyjściowych wektorów kontekstu).

str. 108: „Dla sieci SygnaT (avg) wyniki są nominalnie wyższe, ale mieściły się w zakresie zmienności.” - co Doktorantka chciała przez to powiedzieć?

5. Wniosek końcowy

W podsumowaniu niniejszej recenzji chciałbym stwierdzić, że przedstawiona praca zawiera oryginalne i wartościowe koncepcje, stanowiące zauważalny wkład do dziedziny biometrycznego rozpoznawania twarzy z użyciem tzw. obrazów wieloujęciowych. W konsekwencji uważam, że rozprawa doktorska Pani magister inżynier Weroniki Gutfeter pt. „Identyfikacja twarzy na podstawie obrazów wieloujęciowych z zastosowaniem głębokich sieci agregujących” **spełnia**

wymagania określone w odnośnych przepisach i tym samym **wnioskuję o dopuszczenie Doktorantki do publicznej obrony**.

Literatura

[1] Wang M., Deng W., „Deep Face Recognition: A Survey, CoRR, vol. abs/1804.06655, 2018, http://arxiv.org/abs/1804.06655