prof. dr hab. inż. Andrzej Pacut 30 września 2022

Instytut Automatyki i Informatyki Stosowanej
Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska

Naukowa i Akademicka Sieć Komputerowa, Państwowy Instytut Badawczy

Recenzja Rozprawy Doktorskiej

mgr inż. Artura Grudnia

**Biometryczne rozpoznawanie osób na podstawie analizy obrazów termowizyjnych twarzy**

Dla Rady Dyscypliny Automatyka Elektronika i Elektrotechnika

Wojskowej Akademii Technicznej im. Jarosława Dąbrowskiego

Niniejsza recenzja została napisana na wniosek Rady Dyscypliny Automatyka Elektronika i Elektrotechnika Wojskowej Akademii Technicznej im. Jarosława Dąbrowskiego i dotyczy pracy doktorskiej przedstawionej przez p. mgr inż. Artura Grudnia.

Praca składa się z wstępu, 9 rozdziałów i bibliografii i liczy 127 stron. Rozdziały 1, 2 i 3, są przeglądowymi rozdziałami wstępnymi. W r. 4 sformułowano podstawowe tezy pracy. Kolejne rozdziały stanowią opis merytoryczny proponowanego podejścia. W r. 5 omówiono stosowane bazy danych obrazów termowizyjnych. W r. 6 omawiane są metody detekcji twarzy w termowizji. W r. 7 określono i zbadano referencyjne metody rozpoznawania. Zasadniczą - i nowatorską - część pracy stanowi r. 8, w którym sformułowano i zbadano proponowaną metodę weryfikacji twarzy przez identyfikację. Rozdział 9 zawiera podsumowanie i wnioski końcowe. Pracę kończy Bibliografia.

## 1. Zagadnienie naukowe rozpatrywane w pracy (teza rozprawy)

Rozprawa ma charakter doświadczalny i dotyczy zagadnienia rozpoznawania tożsamości na postawie obrazów termowizyjnych twarzy (otrzymywanych w dalekiej podczerwieni). Zagadnienie to jest znacznie trudniejsze od rozpoznawania obrazów w świetle widzialnym, głównie ze względu na mniejszą zawartość informacji, znacznie mniejszą dostępność do danych doświadczalnych, a także trudniejsze technicznie - i droższe - tworzenie obrazów termowizyjnych. Z drugiej strony, rozpoznawanie na podstawie obrazów termowizyjnych jest istotne ze względu na zastosowania związane z odpornością na próby oszustwa i pasywny tryb pracy (bez oświetlenia zewnętrznego).

Rozpoznawanie obrazów twarzy w świetle widzialnym jest obecnie szeroko stosowaną dojrzałą techniką biometryczną, która może być stosowania "*in the wild*", tzn. dla zdjęć o różnym poziomie oświetlenia, różnych położeniach twarzy, zmiennej mimice, twarzy częściowo przesłoniętych, w przypadku wielu twarzy na zdjęciu etc. W wielu przypadkach wyniki są lepsze niż te otrzymywane przez ludzi. Podstawową metody detekcji i rozpoznawania są obecnie realizowane przy pomocy uczenia głębokiego

(głębokich sieci neuronowych). Metodologia dobrze rozwinięta dla obrazów tworzonych w świetle widzialnym może być "ostrożnie" przenoszona do problemów termowizyjnych.

Postawione przez doktoranta zagadnienie dotyczy detekcji twarzy na zdjęciu termowizyjnym i weryfikacji tożsamości obserwowanej twarzy. Pierwszym krokiem było zgromadzenie odpowiednich danych, w tym utworzenie własnej multimodalnej bazy zawierającej dane termowizyjne twarzy (rozdz. 5). Tworzenie baz danych jest często niedocenianym etapem prac, będącym jednak niesłychanie istotną i czasochłonną podstawą dalszych badań. Kolejnym krokiem było zbadanie możliwości detekcji obrazów termowizyjnych przy zastosowaniu gotowych metod sprawnie działających dla obrazów tworzonych w świetle widzialnym (r. 6). Analiza tych metod pozwoliła na wybór metod detekcji twarzy znanych dla światła widzialnego, które najlepiej działają dla analizowanych termowizyjnych baz danych. Podobne postępowanie dla weryfikacji tożsamości skończyło się niepowodzeniem. Doktorant zaproponował metodę alternatywną, w której rozpoznawanie twarzy zostało sprowadzone do rozróżniania par obrazów twarzy (“dubletów”) rozpoznawanej twarzy i osoby najbardziej do niej podobnej w bazie danych, a także par obrazów różnych osób. Nauczenie podobieństwa i niepodobieństwa takich par umożliwiło odróżnianie danej osoby od obrazu innych osób. Istota pomysłu polega zatem na utworzeniu klastrów par obrazów tej samej osoby i par obrazów osób różnych. W rezultacie, w miejsce bezpośredniej weryfikacji dokonuje się najpierw równoważnej klasyfikacji na klastry podobne i klastry niepodobne, kwalifikowane jako klasy obrazów podobnych lub niepodobnych.

Metoda ta pozwoliła na osiągnięcie poziomu TAR rzędu 83% dla FAR=1% w odróżnieniu od metod przeniesionych z obrazowania w świetle widzialnym (“metod referencyjnych”), gdzie największy osiągnięty współczynnik TAR był równy ok. 60%. Rozwiązanie takie jest nowatorskie w ramach rozpoznawania twarzy, choć stosowane było dla wykrywania fałszerstw podpisów odręcznych. Przebadano różne metody weryfikacji. Obliczenia wykonano przy użyciu skryptów wykorzystujących gotowe segmenty oprogramowania MATLAB.

Należy podkreślić, że zagadnienie budowy modelu systemu rozpoznającego zostało rozwiązane w ciekawy i oryginalny sposób, a jego prawidłowe działanie uzasadnione eksperymentalnie. Tym samym **główny cel pracy został osiągnięty.**

## 2. Oryginalność i przydatność rozprawy, jej pozycja w stosunku do stanu wiedzy

Metodologia opracowana przez autora jest oryginalna i stanowi alternatywne rozwiązanie problemu weryfikacji i - jako taka - wymaga wszechstronnego zbadania. W przypadku typowej weryfikacji dane są porównywane z danymi (własnymi i obcymi) w bazie danych. W przypadku proponowanej metody, dane porównywane są z reprezentantem klastra danych własnych lub obcych. Powodować to może zmniejszenie błędu, ale tylko wtedy, gdy reprezentant klastra został prawidłowo wybrany (etap identyfikacji).

Warto zbadać ten mechanizm dla znanych problemów weryfikacji w świetle widzialnym. Podkreślam jednak, że przeprowadzone przez Doktoranta eksperymenty dla badanych baz danych prowadzą do wyników akceptowalnych z praktycznego punktu widzenia.

## 3. Słabe strony rozprawy i jej główne wady

Metodologia detekcji i weryfikacji zaproponowana w pracy została jednak opisana niejasno i dość chaotycznie. Całkowicie brak uzasadnienia i interpretacji proponowanej metody. Brak również

uzasadnienia jej działania poprzez porównanie z metodą "bezpośrednią". Opisano szereg drobiazgowych operacji na obrazie, natomiast brak jest uzasadnienia całej metody. Nie zanalizowano wad metody, do której mogą należeć np. znacznie większa złożoność obliczeniowa a także trudność w rozszerzaniu osób weryfikowanych (konieczność powtórnej identyfikacji).

Do istotnych problemów zaliczam następujące:

1. Brak uzasadnienia działania proponowanej metody i porównania do mechanizmów metody "bezpośredniej". Jakie są różnice obu mechanizmów, dlaczego i kiedy proponowana metoda może być lepsza (czy gorsza)? Jej wady i zalety nie zostały omówione. W szczególności brak dyskusji złożoności obliczeniowej proponowanej metody w porównaniu do metody bezpośredniej weryfikacji, zajętości pomięci, problemu rozszerzania liczby osób weryfikowanych, etc. Badania takie są potrzebne przed szerszym wprowadzeniem proponowanej metodologii.
2. Nie zastosowano augmentacji obrazów stosowanej od początków głębokich sieci neuronowych w celu zwiększenia liczby danych w każdej z klas. Pozwala to na nawet kilkudziesięciokrotne zwiększenie rozmiaru bazy, co jest niesłychanie istotne ze względu na liczebność parametrów modeli. Doktorant wspomina o augmentacji polegającej na odwróceniu obrazu i stosuje augmentację polegającą na zamianie obrazów w dubletach obrazowych. Odwracanie obrazu jest tu oczywiście wykluczone ze względu na możliwe otrzymanie różnych tożsamości. Powszechnie stosuje się transformacje odpowiednio dużej ramki, w tym translacje, niewielkie obroty i transformacje liniowe ramki oraz modyfikacje kanału barwy (to ostatnie nie może być tu stosowane). Brak augmentacji znacząco obniżył możliwą liczbę danych co pogorszyło otrzymywane wyniki. Sugeruję konieczne augmentację i powtórzenie obliczeń w dalszych implementacjach metody. Co do dubletów: Zamiana obrazów w dubletach sprowadza się do odbicia w przestrzeni punktów dubletu. Nie widać powodu otrzymania różnych wyników, jeśli równocześnie zamienianie są oba obrazy (możliwe są jednak różnice losowe).
3. Ze względu na małą liczbę danych Doktorant stosował metody stosowane dla światła widzialnego bez "przeuczania" choćby ostatnich warstw sieci. Sytuacja może się zmienić po odpowiedniej augmentacji. Warto wypróbować proponowane podejście po zwiększeniu liczby danych.
4. Brak interpretacji wyników prawie wszystkich eksperymentów. Co np. oznacza, że dla bazy A otrzymano "lepszy" wynik niż dla bazy B? Jaki stąd wniosek? Co oznacza, że otrzymano lepsze wyniki dla metody X niż dla Y? Czy oznacza to, że metoda X jest lepsza dla wszystkich obrazów (światło widzialne, podczerwień, obroty twarzy, wielkość obrazu, etc.)? Tabele powinny być raczej wstępem do analizy metod, a nie jej zakończeniem.
5. W szczególności, nie zanalizowano przyczyn dramatycznie niskich dokładności i bardzo dużej wariancji wyników dla większości analizowanych metod i różnych popularnych w literaturze sieci neuronowych. Na przykład, wyniki uzyskane dla klasyfikatorów SVM są tak niskie, że podejrzewam błędy obliczeniowe lub błędne zastosowanie klasyfikacji wielokrotnej. Metoda SVM przez wiele lat była dominującą metodą klasyfikacji liniowej i nieliniowej. Jaki zastosowano rodzaj jądra i jak zostało ono dobrane?
6. Sieci neuronowe zostały w dysertacji zastosowane jedynie jako algorytmy wydobywania cech. Oczywiście, porównywanie tak wyliczanych cech jest możliwe przy użyciu innych metod (np. SVM), jednak czemu nie podjęto próby realizacji całego procesu weryfikacji przez sieć neuronową?
7. Brak informacji na temat sesji eksperymentalnych. Wiadomo, że pomiary tej samej charakterystyki biometrycznej dla danej osoby w ramach jednej sesji mogą być znacznie bardziej jednorodne niż

te otrzymane w czasie wielu sesji. Wspomniano o pomiarach w dwu różnych eksperymentach, jednak nie jest jasne, jak wykorzystano te wyniki. Eksperymenty przeprowadzone jednosesyjnie mogą prowadzić do sztucznie lepszej zgodności wyników.

8. Nie uwzględniono błędów FTE *(failure to enroll*, błąd rejestracji) i FTA ({failure to acquire}, błąd pozyskania próbki), które znacząco mogą zmienić wyniki weryfikacji. Nie wiadomo jednak, czy używane bazy danych umożliwiają taką analizę. Różnice pomiędzy wynikami dla baz danych mogą wynikać z - nieuwzględnionych - różnych poziomów tych błędów.
9. „Dla obrazów z zakresu LVIR obrazy zawsze zostają powielone trzykrotnie aby były zgodne z wymiarem warstwy wejściowej" (s.44). Czy warstwa wejściowa musi mieć 3 kanały ? - przecież to nieprawda. Może mieć dowolną liczbę kanałów! Postępowanie takie zwiększa złożoność obliczeniową.
10. W wielu miejscach pracy mowa o doborze eksperymentalnym parametrów, brak jednak odpowiedniej dokumentacji tych eksperymentów. Nie pozwala to na weryfikację decyzji strukturalnych (dotyczących hiperparametrów modeli).

## 4. Analiza źródeł i narzędzi pracy

Bibliografia pracy, zawiera 148 pozycji. Choć zagadnienia rozpoznawania twarzy mają niesłychanie bogatą literaturę, to dla obrazów termowizyjnych literatura jest znacznie uboższa. Przedstawiony przez Doktoranta wykaz literatury świadczy o odpowiednim zasobie narzędzi i zakresie wiedzy.

Nie przedstawiono osobnego spisu prac doktoranta. Z listy publikacji przedstawionej na moją prośbę przez Doktoranta wynika, że jest on współautorem 18 publikacji, w tym 8 publikacji w czasopismach i 8 w materiałach konferencyjnych w języku angielskim. Mimo braku publikacji samodzielnych **jest to dorobek znaczny**, a fakt publikacji zespołowych jest typowy dla prac eksperymentalnych.

Co do narzędzi:

Przegląd metod obrazowania termowizyjnego jest ciekawy i przydatny, jednak niektóre zaawansowane pojęcia nie są wytłumaczone (egzytancja energetyczna, bolometr).

Przegląd metod neuronowych w r.3 „Sieci neuronowe i deskryptory cech lokalnych" (s. 35-50) uważam za nietrafiony i w dużej mierze po prostu błędny. Rozdział ten jest zbyteczny dla osób o choćby niewielkiej wiedzy na ten temat i niewiele wyjaśniający dla osób bez odpowiedniej wiedzy: Wiadomości podawane są chaotycznie i często bez związku z dalszą częścią pracy, w której wykorzystywane są „gotowe" sieci neuronowe o znacznie większej złożoności. Niektóre błędy omówiono poniżej

- [s.35] nie uwzględniono metod częściowo nadzorowanych
- [s.35] regresja jako druga (po SVM) metoda klasyfikacji polega na „oszacowaniu danych testowych pod względem poznania ich dalszego trendu": tak jest tylko w szczególnych przypadkach
- [s.35] nie uwzględniono w opisie popularnych metod częściowo nadzorowanych
- [s.35] Metody SVM i Ada-Boost dotyczą różnych obiektów. SVM jest de facto uogólnieniem perceptronu Rosenblatta na modele optymalne i nieliniowe i stosowana jest dla wektorów. Ada-Boost z kolei dotyczy tworzenia cech dla obrazów (np. w $R^3$).
- [s.36] Metody regresyjne wywodzą się z modeli predykcyjnych, ale od dawna dotyczą znacznie szerszej klasy problemów modelowania

- [s.37] model McCullocha-Pittsa (nie ma algorytmu uczenia) jest pra-źródłem neuronu Rosenblatta. Rys 3.3 jest błędny - brak obciążenia (lub progu)
- [s.38, (3.2,3.3)] aktualizacja wag dla neuronu Rosenbatta powinna uwzględniać zmiany progu. Metoda dotyczy tylko historycznych modeli typu Rosenblatta. Wszystkie współczesne modele neuronowe są inne i wykorzystują różniczkowalne funkcje aktywacji a ich działanie minimalizuje funkcję kary za niedopasowanie do sygnału uczącego przez zastosowanie propagacji wstecznej gradientu.
- [s.38, (3.4)] „dodatkowo jest wyznaczana funkcja błędu". Tu nastąpił przeskok z opisu warstwy do sieci wielowarstwowych, dla których wyznaczenie funkcji błędu jest podstawą działania. Wzór dotyczy pojedynczej warstwy neuronu a nie sieci, jak sugeruje komentarz.
- [s.38] Dla typowego modelu perceptronu i dla wszystkich typowych modeli sieci neuronowych stosuje się metody gradientowej optymalizacji, a nie optymalizacji bezgradientowej! Inne metody wymieniane przez autora też są metodami gradientowymi!
- [s.38] Próby klasyfikacji metod nie wydają się są obecnie sensowne. Pod koniec 20 wieku powstały dziesiątki metod określania cech obiektów, które zostały jednak obecnie całkowicie zdominowane przez modele neuronowe, w szczególności przez głębokie sieci neuronowe.
- [(3.5)] Nie uwzględniono warunków brzegowych
- *W* (metoda Haara) nie jest macierzą.
- [s. 41] Element nieliniowy jest niezbędnym elementem neuronu
- Niezrozumiałe: „Maska filtrująca przed procesem uczenia ma różne wagi"
- [(3.8)] Funkcja aktywacji, a nie funkcja progowa
- [(3.9)] Funkcja sigmoidalna jest definiowana inaczej
- [s.39] Lustrzane odbicie obrazu twarzy nie jest właściwą augmentacją.
- [s.39] Nie zdefiniowano sieci jednokierunkowej. Podział względem kierunku uczenia: kierunek przedni i wsteczny?? „Najbardziej znane sieci z kierunkiem przednim to sieć perceptronowa i radialna". To nieprawda! Ponadto, sieć Hopfielda zwykle jest siecią bez uczenia, sieć Kohonena i Bayesowska wymaga bardziej szczegółowej charakteryzacji.
- [s.41] Nie uwzględniono w opisie typowego współdzielenia wag warstwy splotowej (choć uwzględniono tę metodę we wzorach). Nie uwzględniono problemów brzegowych.
- [s.41] „Maska filtrująca przed procesem uczenia ma różne wagi"??, ale „wagi są zmieniane tak, jak to ma miejsce w procesach uczenia sieci neuronowych". Przecież to właśnie są sieci neuronowe. „Jakie cechy są ekstrahowane z sieci neuronowej i każdej kolejnej warstwy zależy od danych uczących" - o co tu chodzi?
- [s.42] „Często występują funkcje aktywacji": bez nieliniowości otrzymujemy tylko modele liniowe, mogące jedynie służyć jako aproksymatory lokalne. Funkcja sigmoidalna (3.9) jest określona błędnie.

## 5. Sposób przekazania wyników, inne uwagi

- Rozdział dotyczący terminologii biometrycznej powinien raczej znaleźć się w Dodatku.
- [s.22, r.1.6] Metody rozpoznawania twarzy były przez lata klasyfikowane na dziesiątki sposobów (rys. 1.7 nie jest aktualny, sugeruje tylko dwie architektury rozpoznawania: syjamską i trypletową.)
- [s.24, r. 1.6] Przegląd literatury jest niejednorodny. Porównywane są różne wskaźniki jakościowe. Brak krytycznego ustosunkowania się do wyników literaturowych Podział metod rozpoznawania (rys. 1.7) chaotyczny i nie stosowany w (dominującym obecnie) kontekście sieci neuronowych.

- [s.47] dla sieci Yolo2 jako *backbone* używany jest *darknet* 19 (a nie 53)
- [s.49] LDP to lokalne wzorce kierunkowe a nie lokalne wzorce pochodnych.
- [s.65] wektor jako „rezultat metody Haara" ?
- [r. 6.2.5] – chaotyczny opis testowania.
- [r. 6.3] Nie wyjaśniono, co to są obwiednie kotwiczące.
- [s. 74] dlaczego „rozmiar partii liczby =1"?
- [s.81-89] Niewłaściwie opisana architektura syjamska. Czy nie stosowano uczenia? Istotą tej metody jest jeden proces uczenia prowadzący do identycznych parametrów gałęzi sieci.
- [s.82] Klasyfikacja na metody proste i metody złożone nie jest uzasadniona.
- W opisie metody SVM pominięto problem doboru funkcji jądra. Nie wiadomo w szczególności, jaki wielomian stosowano.
- Niektóre błędne czy niestandardowe definicje utrudniają czytanie. Obwiednia (s. 61) została zdefiniowana niezgodnie z jej znaczeniem; prawidłowy termin to w tym przypadku najczęściej „ramka", a przez obwiednie rozumie się brzeg badanego obiektu. „Poziom istotności" (s. 82) jest używany niezgodnie z jego definicją matematyczną w statystyce i analizie danych. Pojęcie „wydajności" powinno dotyczyć własności związanych z czasem, a na s.18 mamy „Im większa wartość FAR tym wydajniejszy jest analizowany system".

## 6. Podsumowanie

Rozprawa jest dobrym przykładem dysertacji, której wyniki mają bezpośrednie istotne przełożenie praktyczne. Metodologia proponowana przez Autora może stać się - w określonych przypadkach - alternatywnym podejściem do weryfikacji modeli. Podsumowując, stwierdzam, że

**przedstawiona przez pana mgr inż. Artura Grudnia rozprawa doktorska pt.**

**Biometryczne rozpoznawanie osób na podstawie analizy obrazów termowizyjnych twarzy**

**spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy i wnioskuję o jej przyjęcie i dopuszczenie do dalszych etapów przewodu doktorskiego**

Andrzej Sankt