Dr hab. inż. Jacek Kucharski, prof. nadzw.
Politechnika Łódzka
Instytut Informatyki Stosowanej

Łódź, 10 marca 2026 r.

RECENZJA
rozprawy doktorskiej mgr. Mohda Faizana ANSARIEGO
pt.: **„*Using a Camera to Determine Human Gaze Point*"**

Recenzja została przygotowana na zlecenie Dziekana Wydziału Automatyki, Elektroniki i Informatyki Politechniki Śląskiej, prof. dr. hab. inż. Dariusza Kani.

Recenzowana rozprawa została przygotowana pod kierunkiem prof. dr. hab. inż. Pawła Kasprowskiego. Monografia liczy 128 stron i składa się z 7 rozdziałów, bibliografii oraz spisów rysunków i tabel. Bibliografia obejmuje 183 pozycje, z których dwie są współautorstwa mgr. M.F. Ansariego.

1. Temat, zakres i cele pracy

Przedstawiona rozprawa dotyczy opracowania metod identyfikacji kierunku spojrzenia przy użyciu nisko-kosztowych kamer wykorzystywanych w sprzęcie powszechnego użytku i stanowi istotny wkład w upowszechnienie technologii śledzenia wzroku (eye-tracking). Głównym założeniem pracy jest umożliwienie wiarygodnej estymacji kierunku spojrzenia przy wykorzystaniu standardowych, niezmodyfikowanych kamer internetowych znajdujących się w laptopach i komputerach stacjonarnych, eliminując tym samym konieczność stosowania drogiego, specjalistycznego sprzętu lub urządzeń montowanych na głowie użytkownika. Przyjęty cel pracy jest zgodny z aktualnymi trendami w obszarze interakcji człowiek–komputer (HCI) oraz widzenia komputerowego obserwowanymi w ostatniej dekadzie, w których dążenie do powszechnego i nieinwazyjnego przetwarzania danych doprowadziło do przejścia od sprzętu laboratoryjnego do implementacji na urządzeniach brzegowych.

Mgr M.F. Ansari, definiując w sekcji 1.1 cel pracy, stwierdza, że głównym celem jego badań jest „*zbadanie, czy osoba posiadająca standardową kamerę internetową, zazwyczaj w laptopie lub komputerze stacjonarnym, może wykonywać estymację kierunku spojrzenia na swoim urządzeniu bez potrzeby używania dodatkowego oprogramowania lub sprzętu*". Ze względu na niską jakość obrazów zwykle uzyskiwanych z tego typu kamer (szczególnie, gdy obszar oka zajmuje jedynie niewielką część kadru), a także zmienne i niekontrolowane warunki oświetleniowe, konieczne jest zastosowanie specjalnej metodyki przetwarzania danych. Autor skierował swoje wysiłki w stronę technik głębokiego uczenia (Deep Learning – DL), a w szczególności konwolucyjnych sieci neuronowych (CNN).

Moim zdaniem cele przedstawione w rozprawie są ambitne i bez wątpienia stanowią wyzwanie adekwatne do zakresu rozprawy doktorskiej, a uzyskane rozwiązanie stanowi oryginalne osiągnięcie Doktoranta.

Zgodnie z tak określonym celem sformułowano trzy hipotezy badawcze:

Hipoteza 1 (H1): *możliwe jest opracowanie modelu opartego na CNN, który klasyfikuje kierunek spojrzenia użytkownika do regionów ekranu przy użyciu obrazów oczu i twarzy o niskiej rozdzielczości i jakości uzyskanych ze standardowej kamery internetowej.*

Hipoteza 2 (H2): *model trenowany na obrazach oczu i twarzy o niskiej jakości pochodzących ze standardowej kamery internetowej, dopasowany do konkretnej osoby, może osiągnąć dokładność estymacji spojrzenia porównywalną z nowoczesnymi systemami eye-tracking.*

Hipoteza 3 (H3): *estymacja spojrzenia przy użyciu obrazów niskiej jakości z kamer internetowych, wykorzystująca modele wstępnie trenowane na danych wielu użytkowników, a następnie dostrajane dla konkretnej osoby z użyciem uczenia transferowego, wymaga mniejszej ilości danych i szybciej osiąga zbieżność podczas dostrajania niż modele trenowane od podstaw dla danej osoby.*

Sposób sformułowania hipotez wskazuje, że badania zostały zaplanowane w sposób progresywny. Hipoteza H1 bada podstawową możliwość estymacji kierunku spojrzenia na podstawie zaszumionych danych poprzez uproszczenie problemu do zadania klasyfikacji (przewidywanie regionu). Hipoteza H2 dotyczy precyzyjnej regresji punktu widzenia, zakładając, że ograniczenie zadania do konkretnej osoby (modelowanie spersonalizowane) może zrekompensować ograniczenia kamery i umożliwić osiągnięcie dokładności porównywalnej z komercyjnymi systemami śledzenia wzroku. Hipoteza H3 sprawdza skalowalność modeli spersonalizowanych i proponuje wykorzystanie uczenia transferowego jako mechanizmu umożliwiającego osiągnięcie wysokiej jakości wyników przy minimalnej liczbie danych treningowych.

Pozytywnie oceniam kompletność rozprawy, w której Doktorant przedstawił w szczególności:

- podstawy teoretyczne dotyczące najważniejszych obszarów badań, tj. opis ludzkiego oka w kontekście rozumienia procesu widzenia i określania kierunku spojrzenia, jak również przedstawienie wybranych zagadnień sztucznych sieci neuronowych – od podstaw po CNN i uczenie transferowe;
- kompleksową analizę aktualnego stanu badań w zakresie estymacji kierunku spojrzenia, obejmującą różne podejścia do tego problemu (metody oparte na modelach, cechach oraz wyglądzie), uzupełnioną o przegląd zastosowań głębokiego uczenia i porównanie różnych zbiorów danych dotyczących tych zagadnień;
- propozycję oryginalnych struktur sieci CNN oraz procedur pozyskiwania i wstępnego przetwarzania danych;
- eksperymentalną weryfikację zaproponowanych rozwiązań oraz analizę uzyskanych wyników.

Moim zdaniem rozprawa stanowi więc kompleksowe opracowanie z jasno określonym celem i jego konsekwentną realizacją.

2. Ogólna ocena rozprawy

Główna część pracy poprzedzona jest zwięzłym, lecz prawidłowo skomponowanym rozdziałem wprowadzającym (rozdz. 2), w którym Doktorant przedstawia podstawowe informacje związane z tematyką badań. Zaprezentowana tu została anatomia oka jako narządu człowieka, wraz ze specyficzną notacją oraz wyjaśnieniem mechanizmu widzenia i ruchów gałki ocznej. Następnie przedstawiono przegląd sztucznych sieci neuronowych (ANN) obejmujący podstawowe pojęcia, struktury sieci oraz ich formalny opis matematyczny, a także bardziej zaawansowane rozwiązania i techniki, takie jak konwolucyjne sieci neuronowe (CNN) i uczenie transferowe. Ogólnie materiał teoretyczny przedstawiono poprawnie, w sposób wyważony i zapewniający niezbędny wgląd matematyczny w zastosowane metody. Świadczy to również o dobrym poziomie kompetencji mgr. M.F. Ansariego w obszarach objętych rozprawą.

Część wprowadzającą uzupełnia wnikliwy przegląd najnowszych metod estymacji spojrzenia (rozdział 3). Doktorant przedstawił obszerny przegląd badań oparty na licznych publikacjach z ostatniej dekady, co dowodzi aktualności podjętej problematyki. Uwzględniono również rys historyczny badań sięgający początków XX wieku, kiedy pojawiły się pierwsze próby estymacji kierunku spojrzenia. Zwrócono uwagę na zmianę paradygmatu badań około roku 2015 w kierunku metod głębokiego uczenia oraz omówiono rozwiązania oparte na CNN i architekturach typu transformer. Przedstawiono także opis kilku zbiorów danych powstałych w ostatniej dekadzie dzięki intensywnym badaniom nad estymacją spojrzenia, wraz z ich charakterystycznymi cechami i metodyką zbierania danych. Rozdział ten stanowi dobre uzasadnienie badań prowadzonych przez mgr. M.F. Ansariego i wskazuje kilka nierozwiązanych problemów w obszarze estymacji spojrzenia, które stały się celem pracy.

Kolejne trzy rozdziały (r. 4, 5 i 6) odnoszą się bezpośrednio do hipotez sformułowanych na początku rozprawy (odpowiednio H1, H2 i H3) i stanowią główną oraz najważniejszą część pracy. Struktura wszystkich tych rozdziałów jest podobna, ponieważ opierają się one na artykułach już opublikowanych (rozdz. 4 i 5) lub przygotowanych do publikacji (rozdz. 6). W naturalny sposób wzmacnia to wiarygodność przedstawionych rozwiązań i wyników (przynajmniej w przypadku rozdziałów 4 i 5), jednak ponieważ treść tych rozdziałów została w dużej mierze przejęta z cytowanych publikacji, zauważalne są pewne powtórzenia, co sprawia, że cała rozprawa jest nieco redundantna. Na przykład sekcje 4.2.1 i 4.2.2, poświęcone zbieraniu danych oraz ich wstępnemu przetwarzaniu, zawierają bardzo podobne opisy metodyki jak te przedstawione w sekcjach 5.2.1 i 5.2.2 oraz 6.2.1 i 6.2.2. Podobna sytuacja dotyczy również sekcji wprowadzających w tych rozdziałach (tj. sekcji 4.1, 5.1 i 6.1), gdzie kilkakrotnie powtarzane są zbliżone uwagi dotyczące znaczenia estymacji kierunku spojrzenia oraz typowych problemów związanych ze zbieraniem i przetwarzaniem danych.

Podstawowym założeniem przyjętym w rozdziale 4 jest zastąpienie regresyjnej estymacji kierunku patrzenia z dokładnością do pojedynczych pikseli, co może być zbyt ambitnym celem w przypadku zaszumionego sygnału z kamery internetowej, zadaniem klasyfikacji o mniejszej szczegółowości, czyli do dyskretnych regionów ekranu. W tym celu utworzono dwa zbiory danych, obejmujące obrazy odpowiednio jednego oraz czterech uczestników. We wszystkich przypadkach badane osoby były proszone o równomierne kierowanie uwagi na 20 dyskretnych punktów rozmieszczonych na ekranie. Procedura pozyskiwania danych została zorganizowana w taki sposób, aby uwzględnić różnorodne warunki ekspozycji, co sprawia, że zbiory danych są odpowiednie do prawidłowego trenowania sieci neuronowych. Zadanie uczenia maszynowego w tym punkcie zostało zdefiniowane jako problem wieloklasowej klasyfikacji

polegający na przewidywaniu, na którym z 20 obszarów ekranu użytkownik skupia wzrok w danym obrazie wejściowym. Zaproponowano trzy architektury konwolucyjnych sieci neuronowych (CNN), które jako dane wejściowe wykorzystują odpowiednio obraz jednego oka, obraz twarzy oraz obrazy obu oczu. Choć autor stwierdza na stronie 52, że „*zaproponowane architektury zostały wybrane spośród kilku możliwości, które początkowo analizowano*", ostateczny wybór ma charakter arbitralny i nie został szerzej omówiony w rozprawie, co należy uznać za słabość proponowanego rozwiązania. Wyniki przedstawione w sekcji 4.3.2 wskazują na możliwość uzyskania stosunkowo wysokiej dokładności śledzenia wzroku w zróżnicowanych warunkach środowiskowych (w większości przypadków dokładność przekraczała 80%, a najwyższa wartość sięgała ponad 88%), co stanowi potwierdzenie hipotezy H1. Jednocześnie tabela 4.3 pokazuje, że wyostrzanie obrazów znacząco poprawia całkowitą dokładność klasyfikacji, co potwierdza, że tego typu wstępne przetwarzanie może w pewnym stopniu kompensować „niską jakość" obrazów z kamer internetowych oraz wspomaga działanie sieci CNN. Nie jest również jasne, jaki jest cel przedstawienia szczegółowych danych w tabelach 4.3–4.5, a mianowicie wartości dokładności w trakcie procesu uczenia, ponieważ w tekście nie zamieszczono do nich żadnego komentarza.

Głównym przedmiotem rozdziału 5 jest estymacja kierunku patrzenia metodą regresją, a jego treść obejmuje trenowanie modeli na danych pochodzących od jednego użytkownika, co bezpośrednio odpowiada hipotezie H2. Przeanalizowano dwie różne architektury konwolucyjnych sieci neuronowych (CNN), wykorzystujące jako dane wejściowe jeden lub dwa obrazy. Podobnie jak w poprzednim przypadku, struktury sieci zostały wybrane w sposób arbitralny, jednak w tym przypadku kandydat przeanalizował zestawy wartości niektórych parametrów, takich jak współczynnik uczenia, dropout czy rozmiar jądra konwolucji. Wyniki przedstawione w sekcji 5.2.5 oraz omówione w sekcji 5.2.6, przy najlepszej uzyskanej dokładności w zakresie od 1,98 do 1,14 stopnia, niewątpliwie wskazują na wysoką skuteczność zaproponowanych modeli i stanowią potwierdzenie hipotezy H2. Należy jednak zauważyć, że rezultat ten osiągnięto poprzez trenowanie sieci na obrazach pochodzących od jednego użytkownika, co może prowadzić do silnego przeuczenia sieci, a tym samym znacząco ogranicza ogólność proponowanego rozwiązania. Ponadto zakładane w proponowanym rozwiązaniu zebranie około 12 000 etykietowanych próbek stanowi bardzo duże obciążenie dla użytkownika, co sprawia, że rozwiązanie to jest trudne do zastosowania w praktyce. Niektóre z wniosków sformułowanych w tym rozdziale wydają się dość oczywiste i niekoniecznie wnoszą istotną wartość poznawczą. Na przykład na stronie 65 Autor stwierdza, że „*po przeniesieniu modelu na komputer wyposażony w procesor graficzny (GPU) w celu trenowania nastąpiło wyraźne skrócenie czasu uczenia*"; na stronie 66 zauważa, że „*na podstawie tego eksperymentu staje się jasne, że współczynnik uczenia odgrywa kluczową rolę dla jakości sieci*", natomiast na stronie 73 wskazuje, że „*porównanie to pokazuje, iż trenowanie modelu na danych jednego użytkownika może prowadzić do wyższej dokładności niż w przypadku modeli trenowanych na danych wielu osób*". Ta część rozprawy wykazuje również pewne niedociągnięcia redakcyjne, które utrudniają jej lekturę. Tekst znajdujący się bezpośrednio pod tabelą 5.6 został dokładnie powtórzony pod tabelą 5.8, a inny fragment tekstu został powtórzony powyżej i poniżej tabeli 5.7.

Kluczowym elementem z punktu widzenia celu badań jest rozdział 6 rozprawy, poświęcony zastosowaniu uczenia transferowego w estymacji kierunku spojrzenia. Doktorant wykorzystuje dobrze znaną strategię polegającą na wykorzystaniu wiedzy pozyskanej na dużych zbiorach danych oraz ostatecznym dostrajaniu modelu do konkretnego zadania. W rozdziale tym analizowana jest skuteczność estymacji spojrzenia w warunkach ograniczonej dostępności danych. W tym celu zebrano dane od 19 uczestników, charakteryzujących się znacznym

zróżnicowaniem etnicznym; dane zostały zgromadzone podczas wielu sesji oraz w różnych warunkach oświetleniowych. Uczestnicy byli proszeni o klikanie w punkty z zestawu 54 punktów rozmieszczonych na całej powierzchni ekranu. Około 90 000 obrazów uzyskanych w ten sposób stanowi wiarygodny i dobrze przygotowany materiał do dalszych prac nad rozwojem metody estymacji kierunku spojrzenia. Ponadto powstały zbiór danych stanowi wymierny wkład rozprawy w badania w zakresie widzenia komputerowego i analizy obrazów. Procedura uczenia transferowego została zorganizowana w taki sposób, że model był wstępnie trenowany na danych pozyskanych od wszystkich uczestników z wyjątkiem jednego, a następnie dostrajany specjalnie dla pominiętego uczestnika. Analizie poddano zarówno skuteczność modeli, jak i ich zbieżność dla różnych wielkości zbiorów danych (od 500 do 100 obrazów, a w skrajnych przypadkach nawet do 10 obrazów). Wyniki uzyskane dla modeli z wagami wstępnie wytrenowanymi oraz bez nich potwierdzają główne założenia hipotezy H3, wykazując w większości przypadków wyższą skuteczność oraz lepszą zbieżność modeli wstępnie trenowanych. Najważniejszym wnioskiem z tej części badań jest wskazanie, że zastosowanie uczenia transferowego pozwala uzyskać zadowalającą dokładność estymacji kierunku spojrzenia przy ograniczonej liczbie dostępnych danych. Chociaż kandydat stwierdza na stronie 97, że „*istnieje możliwość dalszego udoskonalenia w celu osiągnięcia dokładności porównywalnej z zaawansowanymi systemami*", podejście to zwiększa praktyczną użyteczność proponowanego rozwiązania w rzeczywistych systemach. Metoda ta wydaje się szczególnie istotna w kontekście obliczeń brzegowych (edge computing), gdzie potrzebne są niewielkie modele dostrajane bezpośrednio na urządzeniu użytkownika. Rozdział 6 nie jest jednak wolny od zbędnych, oczywistych stwierdzeń. Na przykład na stronie 87 autor podaje informację, że „*zależność między calami a centymetrami wynosi 1 cal = 2,54 cm*", natomiast na stronie 91 stwierdza, że „*wyniki te prowadzą do wniosku, iż rezultaty mogą zostać dodatkowo poprawione poprzez zwiększenie liczby obrazów treningowych*".

W rozdziale 7, stanowiącym podsumowanie rozprawy, mgr M.F. Ansari w odpowiedni sposób określa jej wkład w obszar badań nad estymacją kierunku spojrzenia. Wskazuje, że opracował modele oparte na konwolucyjnych sieciach neuronowych (CNN) przeznaczone do dyskretnej klasyfikacji kierunku spojrzenia użytkownika, przeanalizował wpływ personalizacji rozwiązania, a także zbadał metody estymacji spojrzenia wykorzystujące uczenie transferowe. Wszystkie te osiągnięcia odnoszą się bezpośrednio do celu badań oraz sformułowanych w rozprawie hipotez badawczych.

3. Uwagi krytyczne i dyskusyjne

Recenzowana rozprawa nie jest pozbawiona fragmentów, które budzą pewne wątpliwości lub rodzą konkretne pytania.

1. Wyniki działania sieci CNN stosowanych do estymacji kierunku spojrzenia, szczególnie w przypadkach regresji, zawierają współrzędne punktu spojrzenia na ekranie w przestrzeni dwuwymiarowej (tj. współrzędne x i y, jak pokazano np. na rysunkach 5.3, 5.4 i 6.4). Jednak funkcja straty Mean Absolute Error (MAE) wykorzystana podczas trenowania sieci CNN, zdefiniowana równaniami 5.1 i 6.1, uwzględnia jedynie współrzędną y. Jaki jest powód pominięcia współrzędnej x w tych wzorach? Jaka jest różnica pomiędzy oznaczeniami $\hat{y}$ i $\overline{y}$ występującymi odpowiednio w równaniach 5.1 i 6.1?
2. Jaki wpływ miała niedokładność określenia odległości pomiędzy ekranem a oczami obserwatora (*Dist*) przy przeliczaniu błędów z pikseli (centymetrów) na stopnie,

zgodnie z równaniami 5.3 i 6.2? Przyjęto, że odległość ta wynosi 30–50 cm, jednak nie została ona bezpośrednio zmierzona.

3. Czy Doktorant mógłby skomentować możliwość wiarygodnego porównania opracowanej metody estymacji kierunku spojrzenia z innymi nowoczesnymi metodami przedstawionymi w tabeli 5.9, biorąc pod uwagę istotnie różne charakterystyki zbiorów danych wykorzystywanych w analizowanych przypadkach, w szczególności fakt, że zbiór danych użyty w rozdziale 5 został uzyskany jedynie dla jednej osoby?
4. Czy Doktorant dostrzega możliwość zastosowania innych metod wstępnego przetwarzania obrazów twarzy niż klasyfikator Viola–Jonesa i kaskada Haar do detekcji obiektów? Jak bardzo mogłoby to wpłynąć na wyniki estymacji kierunku spojrzenia w proponowanym rozwiązaniu?
5. Omawiając częściowe wyniki metody uczenia transferowego, Doktorant stwierdza w sekcji 6.3.2 na stronie 87, że „*dla prawego oka, w przypadku uczenia transferowego, wartość MAE wynosiła początkowo 164,51 dla 500 obrazów, a następnie nieznacznie wzrosła do 197,47 przy 100 obrazach, jak pokazano w tabeli 6.3. Stały wzrost wskazuje na zdolność modelu do utrzymania stabilnej wydajności pomimo zmniejszenia rozmiaru zbioru danych*". Nie jest jednak jasne, w jaki sposób należy interpretować to stwierdzenie.

4. Ocena poziomu edytorskiego rozprawy

Recenzowana rozprawa pana mgr. M.F. Ansariego przedstawia najważniejsze wyniki przeprowadzonych badań w sposób dość przejrzysty. Doktorant nie ustrzegł się jednak szeregu niedoskonałości stylistycznych w warstwie językowej oraz błędów redakcyjnych, z których niektóre zostały wymienione poniżej.

1. s. 17 jest "*each of its each serving*", powinno być *each serving*;
2. s. 18 jest odniesienie do równ. 2.4.1, powinno być 2.12;
3. s. 19-20 jest "*have been become*", powinno być *have been* lub *have become*;
4. s. 64 and s. 65 są odniesienia do Tabeli 4.2, powinny być do Tabeli 5.1;
5. s. 86 jest "*which we can be observed*", powinno być *which can be observed.*

5. Konkluzja końcowa

Recenzowana rozprawa doktorska mgr. M.F. Ansariego stanowi kompletne rozwiązanie problemu naukowego i – moim zdaniem – zawiera oryginalny wkład autora w rozwój dyscypliny informatyka techniczna i telekomunikacja. Doktorant wykazał się dobrym poziomem wiedzy w obszarach objętych rozprawą, w szczególności w zakresie uczenia maszynowego oraz przetwarzania obrazów. Sformułowane w recenzji uwagi krytyczne nie zmieniają mojej ogólnie pozytywnej oceny rozprawy.

W świetle ustawy Prawo o szkolnictwie wyższym i nauce z dnia 20 lipca 2018 r. stwierdzam, że rozprawa doktorska mgr. M.F. Ansariego spełnia wymagania ustawowe, wobec czego wnoszę o dopuszczenie jej do publicznej dyskusji i obrony.