

dr hab. inż. Dominik Belter, prof. PP
Politechnika Poznańska
Wydział Automatyki, Robotyki i Elektrotechniki

Poznań, 12 listopada 2025 r.

RECENZJA

rozprawy doktorskiej magister Ekemeyong Awong Lisiane Esther
pt.: „Predicting Human Actions for Utilisation in Assistive Robots”

1 Podstawa wykonania recenzji

Niniejsza recenzja dotyczy rozprawy doktorskiej mgr Ekemeyong Awong Lisiane Esther za-tytułowanej: „Predicting Human Actions for Utilisation in Assistive Robots” w dyscyplinie naukowej automatyka, elektronika i elektrotechnika i technologie kosmiczne. Promotorem opiniowanej rozprawy doktorskiej jest prof. dr hab. inż. Teresa Zielińska a promotorem pomocniczym dr Vibekananda Dutta. Recenzja została przygotowana na podstawie uchwały Rady Naukowej Dyscypliny Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne Politechniki Warszawskiej nr 155/III/2025 z dnia 30.09.2025 r. Zgodnie z informacją dodatkową załączoną do powyższego pisma, postępowanie o nadanie stopnia naukowego doktora mgr Ekemeyong Awong Lisiane Esther procedowane jest według wymagań określonych w art. 187 ustawy z dnia 20 lipca 2018 r. - Prawo o szkolnictwie wyższym i nauce.

2 Rozprawa doktorska

2.1 Struktura rozprawy

Rozprawa doktorska została zredagowana w formie 10 rozdziałów z jednym dodatkiem i obejmuje 208 stron. Zestawienie literatury zawiera 327 pozycji, w tym 186 prac z ostatnich pięciu lat. Trzy pozycje literaturowe stanowią współautorskie prace Doktorantki.

Rozdział pierwszy stanowi wprowadzenie do tematyki rozprawy i przedstawia motywację podjęcia badań nad rozpoznawaniem i przewidywaniem ludzkich działań przez roboty (ang. Human Action Prediction). Omawia rozwój tej dziedziny – od eksperymentów laboratoryjnych po zastosowania w rzeczywistych warunkach, takich jak opieka zdrowotna, nadzór nad osobami starszymi, sport, monitoring czy rozrywka. Podkreśla znaczenie postury ciała i sygnałów niewerbalnych w rozumieniu intencji człowieka, a także rolę jakości i różnorodności danych w opracowaniu metod działających w czasie rzeczywistym. Rozdział kończy się przedstawieniem struktury pracy.

Rozdział drugi poświęcony został przeglądowi literatury. Omawia techniki uczenia maszynowego, typy sensorów i danych wykorzystywanych w HAP oraz etapy przetwarzania informacji, a także dostępne zbiory danych służące do porównywania algorytmów HAP. Przedstawia również analizę porównawczą metod pod kątem dokładności, wydajności i szybkości reakcji, ukazując wyzwania oraz postępy w tej dziedzinie.

W rozdziale trzecim przedstawiono motywację badań, problem badawczy oraz nową metodę przewidywania ludzkich działań (HAP) opartą na rozpoznawaniu postur człowieka z danych RGB-D. Opisano założenia i cele pracy, koncentrując się na sekwencjach postur jako kluczowych elementach służących do rozpoznawania całych działań. Zaproponowano podejście łączące klasyfikację i klasteryzację postur w oparciu o bazę danych, umożliwiające dokładne i szybkie przewidywanie działań w czasie rzeczywistym. Omówiono również etapy metodyki: przetwarzanie danych, rozpoznawanie postur i predykcję działań. W tym rozdziale przedstawiono także tezę główną rozprawy oraz określono cele i wpływ naukowy pracy.

Rozdział czwarty opisuje proces pozyskiwania i przygotowania danych do systemu HAP. Przedstawiono sposób nagrywania danych z użyciem kamer RGB-D (ZED 2, Azure Kinect) oraz tworzenie własnych zbiorów we współpracy z Uniwersytetem Waseda. Omówiono etapy przetwarzania danych, takie jak oznaczanie wideo, segmentacja postur, ekstrakcja i normalizacja cech szkieletowych, co zapewnia wysoką jakość i skuteczność uczenia maszynowego w systemie przewidywania działań.

Rozdział piąty opisuje proces grupowania ludzkich postur z wykorzystaniem uczenia nienadzorowanego opartego na mapach samoorganizujących się (SOM, ang. Self-Organizing Maps). Celem jest identyfikacja i grupowanie postur 3D na podstawie ich podobieństw strukturalnych. Wykorzystano własne zbiory danych (WUT-10 i Waseda), zawierające naturalne sekwencje ruchów. Wprowadzono nowy wskaźnik oceny jakości klastrów nazwany Discriminant Score (DS), uzupełniający tradycyjne miary, takie jak Silhouette Coefficient (SC), Dunn Index (DI), Calinski–Harabasz Index (CH) i Davies–Bouldin Index (DBI). Wyniki potwierdzają skuteczność i skalowalność metody w analizie postur i przewidywaniu ludzkich działań.

Rozdział szósty opisuje klasyfikację ludzkich postur z wykorzystaniem metod uczenia nadzorowanego opartych na sieciach typu transformer. Przedstawiono trzy autorskie architektury: Pose Sequence Transformer (PoSeqTNet), Central Posture Transformer (CPT) i Isolated-Attention Transformer (IAT), zaprojektowane do analizy sekwencji postur człowieka. Modele zostały przetestowane na zbiorach WUT-10, Waseda i PKU-MMD, co pozwoliło ocenić ich skalowalność i zdolność uogólniania. Wyniki pokazują, że podejście oparte na

transformerach oraz uwzględnienie sekwencji postur znacząco poprawia dokładność, odporność i użyteczność systemów rozpoznawania postur w rzeczywistych warunkach.

Rozdział siódmy przedstawia metodę pozyskiwania cech niezależnych od orientacji dla grupowania postur opartej na mapach samoorganizujących się. Zaproponowano podejście eliminujące zależność od punktu widzenia kamery, wykorzystujące względne odległości między segmentami ciała i kąty między nimi zamiast absolutnych współrzędnych 3D. Dzięki temu system lepiej radzi sobie z różnymi kątami obserwacji i budową ciała analizowanej osoby. Eksperymenty na zbiorach WUT-10 i Waseda potwierdziły, że metoda znacząco poprawia jakość i spójność grupowania postur.

Rozdział ósmy opisuje rozwój systemu HAP opartego na uczeniu nadzorowanym i nienadzorowanym. Przedstawiono dwa podejścia: jedno wykorzystujące sieci typu transformer, a drugie – mapy samoorganizujące się (SOM). Oba systemy korzystają z relacyjnej bazy danych, która mapuje sekwencje postur na odpowiadające im działania człowieka. System na bieżąco analizuje dane szkieletowe, rozpoznaje postury w czasie rzeczywistym i umożliwia wczesne przewidywanie działań, zanim zostaną one zakończone. Opisano również opracowany graficzny interfejs użytkownika do wizualizacji wyników. Wyniki podkreślają komplementarność między mechanizmami wnioskowania w sieciach typu transformer a SOM.

Rozdział dziewiąty przedstawia hybrydowy system HAP łączący SOM, sieć typu transformer i relacyjną bazę danych do przewidywania działań. Każda sieć niezależnie klasyfikuje postury: SOM grupuje przestrzennie punkty charakterystyczne modelu człowieka, a transformer uwzględnia wzorce przestrzenne i sekwencyjne. Wyniki obu modeli są łączone z użyciem adaptacyjnego ważenia, a rozpoznane postury porównywane z bazą działań, co umożliwia wczesne przewidywanie złożonych działań.

Ostatni rozdział zawiera wnioski z pracy oraz propozycje dalszych badań. Podsumowano w nim wszystkie etapy pracy – od sformułowania problemu i pozyskania danych, po projektowanie modeli i ich ewaluację. Wskazano wkład rozprawy w rozwój systemów przewidywania działań człowieka w robotyce asystującej. Omówiono główne osiągnięcia, ograniczenia badań oraz zaprezentowano kierunki przyszłych prac nad bardziej niezawodnymi, przejrzystymi i przyjaznymi użytkownikowi systemami HAP.

2.2 Ogólna charakterystyka podjętej tematyki

Przedmiotem rozprawy jest opracowanie systemu przewidywania działań człowieka (HAP), opartego na sekwencyjnym rozpoznawaniu postur człowieka wyodrębnionych z danych RGB-D. System integruje podejścia uczenia maszynowego nadzorowanego i nienadzorowanego w celu efektywnego przetwarzania informacji o posturach człowieka. Zaproponowano nową architekturę opartą na sieciach typu transformer, umożliwiającą wysoką skuteczność w sekwencyjnej klasyfikacji postur, z wykorzystaniem mechanizmów *isolated* i *central attention*, zaprojektowanych do uchwycenia zależności czasowo-przestrzennych w przejściach pomiędzy posturami ciała. Dodatkowo opracowano ramy klasteryzacji nienadzorowanej jako podejście uzupełniające, umożliwiające identyfikację kluczowych postur bez konieczności etykietowania danych. Wsparciem dla tego procesu jest nowo wprowadzony wskaźnik

Discriminant Score (DS), który pomaga w doborze optymalnej liczby klastrów, zwiększając interpretowalność i rozróżnialność wyodrębnionych reprezentacji postur. Rozpoznane lub zgrupowane postury są następnie interpretowane w skalowalnym systemie opartym na bazie danych, mapującym sekwencje postur na konkretne działania człowieka.

2.3 Problem naukowy i teza rozprawy

Problem naukowy i tezy pracy zostały jednoznacznie zdefiniowane. Pierwszym problemem naukowym rozpatrywanym w rozprawie jest nienadzorowana klasyfikacja postur człowieka. Drugim problemem analizowanym w pracy jest klasyfikacja sekwencji postur człowieka oraz przewidywanie działań (HAP). Rozwiązanie tak postawionych problemów naukowych odbywa się przy następujących założeniach:

- złożoność systemu: unikanie nadmiernie skomplikowanych architektur, które zużywają zbyt wiele zasobów obliczeniowych i energii,
- efektywność danych: kluczowe jest opracowanie metod wnioskowania i przewidywania, które nie wymagają dużych zbiorów danych, a jednocześnie zapewniają odporność i zdolność adaptacji,
- długoterminowe przewidywanie: większość istniejących zbiorów danych i metod koncentruje się na działaniach krótkoterminowych, co ogranicza ich zastosowanie w szerszych obszarach, takich jak opieka zdrowotna czy robotyka asystująca.

Celem badań jest opracowanie wydajnego, elastycznego i opartego na danych systemu HAP, który wykorzystuje rozpoznawanie postur człowieka – zarówno metodami nadzorowanymi, jak i nienadzorowanymi – do przewidywania bieżących działań człowieka w czasie rzeczywistym. Dążeniem jest przezwyciężenie ograniczeń obecnych systemów poprzez poprawę zdolności uogólniania, zmniejszenie zależności od dużych, oznakowanych zbiorów danych oraz umożliwienie długoterminowego przewidywania działań.

W pracy została sformułowana następująca teza:

Dzięki wykorzystaniu rozpoznawania postur człowieka do przewidywania bieżących działań w systemie opartym na bazie danych możliwe jest osiągnięcie satysfakcjonujących wyników, które odpowiadają na kluczowe wyzwania w HAP, umożliwiając tworzenie bardziej elastycznych i praktycznych systemów asystujących człowiekowi. Przez „satysfakcjonujące wyniki” rozumiane jest osiągnięcie działania w czasie rzeczywistym w warunkach rzeczywistych przy akceptowalnej dokładności i zdolności do uogólniania.

2.4 Rozwiązanie postawionego problemu

W celu udowodnienia tezy rozprawy zaprojektowano i zaimplementowano system, który integruje podejścia uczenia maszynowego nadzorowanego i nienadzorowanego w celu efektywnego przetwarzania informacji o posturach człowieka. Zaproponowano nową architek-

ture opartą na sieciach typu transformer, umożliwiającą wysoką “świadomość” sekwencyjną w klasyfikacji postur, dopasowaną do uchwycenia zależności czasowo-przestrzennych w przejściach postur ciała człowieka. Dodatkowo opracowano ramy nienadzorowanej klasteryzacji jako podejście uzupełniające, pozwalające na identyfikację kluczowych postur bez oznaczonych danych. Wsparciem dla tego procesu jest nowo wprowadzony Discriminant Score, który pomaga w doborze optymalnej liczby klastrów, zwiększając interpretowalność i rozróżnialność wyodrębnionych reprezentacji postur.

Rozpoznane lub zgrupowane postury są następnie interpretowane w skalowalnym systemie opartym na bazie danych, który mapuje sekwencje postur na akcje wyższego poziomu. Mechanizm wnioskowania czerpie inspirację z metodologii drzew decyzyjnych. Proponowany system HAP został gruntownie oceniony przy użyciu benchmarkowych zbiorów danych RGB-D, obejmujących złożone aktywności oraz nagrania z wielu perspektyw. Wyniki eksperymentów wykazały, że system jest zdolny do przewidywania działań człowieka przed ich zakończeniem, z wysoką dokładnością predykcji i niskim narzutem obliczeniowym, co czyni go odpowiednim do zastosowań w czasie rzeczywistym.

2.5 Oryginalny dorobek autora i jego znaczenie poznawcze i aplikacyjne

Oryginalny dorobek Doktorantki związany jest ze zbiorem nowych narzędzi i metod do rozpoznawania postur człowieka oraz całych działań, i można go zdefiniować następująco:

- zaprojektowanie zmodyfikowanych architektur typu transformer do nadzorowanej klasyfikacji postur człowieka, z wprowadzeniem koncepcji takich jak Pose Sequence Transformer (PoSeqTNet), Central Posture Transformer (CPT) oraz Isolated-Attention Transformer (IAT);
- przeprowadzenie ilościowej oceny zaproponowanych architektur na różnych zbiorach danych przy użyciu standardowych miar wydajności (dokładność, precyzja, czułość, F1-score);
- opracowanie algorytmu nienadzorowanej klasteryzacji postur oraz porównanie jego wyników z rezultatami modeli nadzorowanych;
- wprowadzenie nowej miary „Discriminant Score”, wspierającej optymalny wybór liczby klastrów oraz ocenę jakości klasteryzacji postur w nieoznaczonych zbiorach danych;
- zaproponowanie metody ekstrakcji cech niezależnych od punktu widzenia, zwiększającej odporność systemu rozpoznawania postur przy różnych kątach kamery i zróżnicowanej morfologii ciała człowieka;
- opracowanie mechanizmu wnioskowania opartego na skalowalnej strukturze przypominającej drzewo decyzyjne, reprezentującej sekwencje postur i działań jako elementy bazy danych, umożliwiającą ich łatwą aktualizację i rozbudowę.

3 Uwagi merytoryczne

3.1 Silne i słabe strony pracy, uwagi dyskusyjne

Rozprawa obejmuje liczne interesujące koncepcje i wartościowe analizy, co w sposób naturalny sprawia, że część zagadnień może wymagać dodatkowego uszczegółowienia. Poniżej przedstawiono wybrane uwagi:

- W jaki dokładnie sposób zbiór danych został podzielony na część treningową i testową? Czy zestaw testowy zawiera nowych uczestników, którzy nie brali udziału w przygotowaniu zestawu treningowego?
- Proponowana metoda nie uwzględnia obiektów, z którymi człowiek wchodzi w interakcję, a kwestia ta została pozostawiona na przyszłe prace. Jak mógłby wyglądać taki system?
- Przegląd literatury nie uwzględnia wykorzystania precyzyjnych czujników do estymacji pozycji 3D, takich jak OptiTrack. Dlaczego? Tego typu sensor mógłby być użyteczny chociażby do określenia referencyjnych ruchów człowieka (groundtruth).
- Brakuje podsumowania literatury, które podkreślałoby, które metody są najbardziej zbliżone do tych zaprezentowanych w pracy i mogą później posłużyć do porównania. Tab. 2.2 przedstawia różne metody stosowane w HAP, ale zostały one przetestowane na różnych zbiorach danych, więc trudno porównać uzyskane wyniki. Być może warto byłoby porównać je na tym samym zbiorze danych. Później, gdy przedstawiane są wyniki, to proponowane metody są porównywane z innymi metodami z literatury, ale przegląd literatury to dobra okazja, aby wspomnieć o nich wcześniej i uzasadnić wybór metod do porównania.
- Rys. 4.2 pokazuje dane zarejestrowane przez kamery ZED 2 i Azure Kinect, ale nie jest jasne, które dane pochodzą z której kamery.
- Obie reprezentacje szkieletu pokazane na Rys. 4.4 mają 32 punkty. Zgodnie z tekstem liczba punktów wynosi odpowiednio 32 i 34 dla kamer Kinect Azure i ZED 2. Jakie punkty zostały pominięte?
- Rys. 4.5 powinien zawierać cztery diagramy pokazujące, jak reprezentacje wielopunktowe dla obu kamer są redukowane do 19 punktów.
- Co się dzieje, gdy stawy są zasłonięte? Czy oba sensory zawsze generują tę samą liczbę punktów?
- Rysunki 5.5–5.13: czy obrazy są ponumerowane numerem klatki?
- Jaka jest różnica między czynnościami „placing” i „picking” na Rys. 5.6? Rozróżnienie tych dwóch działań na podstawie pojedynczego obrazu wydaje się niemożliwe.

- Czy ta klasteryzacja zakłada, że „picking” odbywa się z podłogi, a „placing” na stole? Nawet w takim przypadku te postury są bardzo podobne, co może tłumaczyć, dlaczego znajdują się w tym samym klastrze uzyskanym przy użyciu odległości euklidesowej.
- Brakuje wizualizacji centroidów (przykładowych postur) w wielu sekcjach.
 - Rys. 5.16: czy ten rysunek przedstawia przykładowe postury reprezentowane przez centroidy?
 - Dlaczego podczas segmentacji nie rozważono klasteryzacji hierarchicznej (aglomeracyjnej)? Jej główną zaletą jest brak konieczności wyboru liczby klastrów.
 - Wszystkie punkty ciała są obliczane względem początku układu współrzędnych. Czy punkt odniesienia ustawiony jest na klatkę 0, czy dla każdej klatki osobno? Zakładam, że jest obliczany względem pozycji dolnego odcinka kręgosłupa w każdej klatce. To dobre podejście, ale w takim przypadku trudno rozróżnić, czy człowiek się porusza, skacze czy kuca. Czy dobrym pomysłem byłoby wzbogacenie wektora wejściowego o informacje o prędkości ciała?
 - „Traditional approaches in HAR and HAP often rely on classical neural networks and extensive datasets, which demand substantial computational resources and time.” – to stwierdzenie powinno być poparte cytowaniami.
 - Czy wyniki w Tab. 6.1 i 6.2 sugerują, że liczba próbek ma kluczowe znaczenie przy trenowaniu transformera?
 - W jaki sposób inne metody przedstawione w Tab. 6.3 kodują problem? Czy dane wejściowe są dokładnie takie same jak dla PoSeqTNet?
 - Ile parametrów mają architektury przedstawione w Tab. 6.6? Wydajność klasyfikacji zależy nie tylko od architektury, ale także od zbioru danych i liczby trenowanych parametrów.
 - Struktura pracy ma charakter przyrostowy. Dlaczego cechy odporne na orientację nie zostały zastosowane od początku, a inne podejścia nie zostały użyte w badaniu ablacyjnym, aby podkreślić wybory projektowe?
 - Początek sekcji 7.5: ilustracja różnych kątów ułożenia ciała opisana w pierwszym akapicie byłaby bardzo przydatna.
 - Dlaczego zbiór danych PKU-MMD nie został uwzględniony w sekcji 2.2?
 - Dlaczego w tabeli 6.4 porównano proponowaną architekturę do trzech innych: Factorised self-attention [304], SPRNet [305], DaViT-Tiny [306], natomiast wyniki klasyfikacji do innych metod - Ts-CNN [308] i IndRNN [309] (Tab. 6.6).

Praca jest praktycznie wolna od błędów językowych i pozbawiona usterek technicznych. Kilka, o których można wspomnieć, znajduje się poniżej:

- "Unlike traditional Transformers, which equally integrates" - integrate
- Koniec pierwszego akapitu w sekcji 6.5: tekst mówi, że liczba punktów została zredukowana z 19 do 18. Z kolejnych sekcji wynika jednak, że ostatecznie jest ich 8.
- "validation.In" - "validation. In"

4 Wnioski końcowe

Powyższe uwagi nie umniejszają wartości merytorycznej pracy. Przyjęte w rozprawie założenia są prawidłowe, a Doktorantka rozwiązała postawione problemy naukowe oraz zastosowała właściwe metody. Rozprawa świadczy o wiedzy na poziomie zaawansowanym o charakterze podstawowym dla dziedziny nauk inżynieryjno-technicznych oraz o charakterze szczegółowym, dotyczącym wykorzystania metod uczenia maszynowego i sztucznych sieci neuronowych do klasyfikacji postur człowieka oraz przewidywania czynności podejmowanych przez człowieka. Rozprawa obejmuje najnowsze osiągnięcia nauki i świadczy o znajomości współczesnej literatury oraz mieści się w dyscyplinie naukowej Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne. W rozprawie przeprowadzono właściwą analizę źródeł, w tym literatury światowej dotyczącej klasyfikacji postur i przewidywania czynności podejmowanych przez człowieka. Mgr Ekemeyong Awong Lisiane Esther wykazała się umiejętnością poprawnego i przekonującego przedstawiania uzyskanych wyników. Wyniki przedstawione są w sposób zwięzły i przejrzysty. Na uwagę zasługuje zaproponowanie metody uczenia nienadzorowanego oraz trzech własnych architektur opartych na sieci typu transformer (CPT, IAT oraz PoSeqTNet), z których dwie (CPT i IAT) przewyższają inne metody z literatury w zadaniu klasyfikacji oraz są znacznie szybsze.

Przedstawiona do recenzji rozprawa doktorska stanowi oryginalne rozwiązanie aktualnego i ważnego problemu naukowego w dyscyplinie naukowej Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne oraz potwierdza wiedzę teoretyczną i umiejętność samodzielnego rozwiązywania problemów naukowych przez Doktorantkę. Rozprawa w pełni spełnia ustawowe wymagania dotyczące rozpraw doktorskich określone w art. 187 ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce.

Wnioskuje do Rady Dyscypliny Automatyka, Elektronika i Elektrotechnika Politechniki Warszawskiej o dopuszczenie rozprawy mgr Ekemeyong Awong Lisiane Esther do publicznej obrony.

Dominik Beltr