Hybrydowa architektura VGG-Transformer w diagnostycznej wizji robotów Humanoidalnych

1. Strategiczny kontekst analizy behawioralnej w interakcji HRI

Ewolucja robotyki humanoidalnej w kierunku systemów afektywnych wymaga od nas, architektów, implementacji rozwiązań zdolnych do wyjścia poza prostą percepcję przestrzenną. W kontekście interakcji człowiek-robot (HRI), kluczowym wyzwaniem jest obiektywizacja stanów wewnętrznych pacjenta. Strategiczne połączenie analizy mimiki twarzy z kinetyką gestów i postawy ciała pozwala robotowi na diagnozowanie dobrostanu psychicznego w czasie rzeczywistym, co jest fundamentem płynności społecznej maszyn autonomicznych.

Uzasadnienie wdrożenia wizyjnych systemów diagnostycznych znajduje potwierdzenie w danych Światowej Organizacji Zdrowia (WHO). Szacuje się, że około 450 milionów ludzi na świecie zmaga się z zaburzeniami psychicznymi, a ponad 264 miliony cierpi na choroby umysłowe stanowiące główne obciążenie globalne. Statystyki są alarmujące – samobójstwa odpowiadają za blisko 700 000 zgonów rocznie, będąc drugą najczęstszą przyczyną śmierci w grupie wiekowej 15–29 lat. Roboty wyposażone w zaawansowane systemy wizyjne stanowią niezbędną odpowiedź na deficyty kadrowe w opiece medycznej. Przejście od subiektywnych, manualnych metod obserwacji do zautomatyzowanej, opartej na danych analizy pozwala na identyfikację subtelnych biomarkerów behawioralnych, które często umykają ludzkiemu oku.

2. Uzasadnienie architektury hybrydowej: synergia CNN i mechanizmów attention

Jako główny architekt systemu, zdecydowałem o implementacji modelu hybrydowego, który łączy stabilność sieci konwolucyjnych (CNN) z elastycznością kontekstową Transformerów. Wybór ten jest podyktowany koniecznością rozwiązania problemu zależności dalekosiężnych w sekwencjach wideo, gdzie klatka po klatce następuje ewolucja afektu.

- **VGG-16 jako ekstraktor cech:** Wykorzystujemy model VGG16 ze względu na jego sprawdzoną głęboką strukturę i małe filtry konwolucyjne, które doskonale radzą sobie z wychwytywaniem lokalnych wzorców przestrzennych (np. mikroekspresji wokół ust czy oczu). Należy podkreślić, że bazowy model VGG16 osiąga top-5 accuracy na poziomie 92,7% na zbiorze ImageNet, co czyni go potężnym fundamentem dla transfer learningu.
- **Visual Transformers (ViT):** Blok Transformera został zintegrowany w celu uchwycenia globalnego kontekstu i informacji czasowych. Z punktu widzenia inżynierii systemowej, musimy jednak pamiętać o bottlenecku obliczeniowym – kwadratowa złożoność mechanizmu self-attention wymaga precyzyjnej optymalizacji, aby system mógł operować na jednostkach krawędziowych (Edge AI).

- **Parametry treningowe:** Model wykorzystuje optymalizator Adam z niskim współczynnikiem uczenia (learning rate) wynoszącym 0,0001, co zapewnia stabilną zbieżność przy fuzji tak różnych architektur.

Poniższa tabela przedstawia porównanie skuteczności (Accuracy) proponowanego rozwiązania z modelami referencyjnymi:

Model / Architektura	Zbiór Danych	Skuteczność (Accuracy)
Proponowany VGG-Transformer	**RAF-DB**	**95,76% (Walidacja)**
MediaPipe (Real-time movement)	Dane wideo	97,67%
DeepFace (VGG-based)	LFW	98,61%
VGG16 (ImageNet baseline)	ImageNet	92,70% (Top-5)
SE-Resnet	RAF-DB	83,37%
YOLOv8 + Transformers	Dane ogólne	81,00%

3. Specyfikacja techniczna faz sieci neuronowej

System przetwarza sygnał w pięciu fazach, optymalizując przepływ danych pod kątem latencji krytycznej dla robotów humanoidalnych:

- **Faza 1 i 2 (wejście i ekstrakcja):** Wejście stanowi obraz RGB o wymiarach 100x100x3. Bloki konwolucyjne VGG16 (z zamrożonymi wagami) ekstrahują mapy cech, redukując wymiarowość przestrzenną do postaci gotowej do spłaszczenia.
- **Faza 3 (spłaszczanie i densyfikacja):** Wyjście konwolucyjne jest spłaszczane i przepuszczane przez warstwę Dense o 512 jednostkach z aktywacją ReLU, tworząc gęsty wektor reprezentacji.
- **Faza 4 (blok transformera):** Implementujemy TransformerBlock z 4 głowicami uwagi (Multi-Head Attention) i dropoutem 0,1. Pozwala to na "zważenie" istotności cech – system potrafi nadać wyższy priorytet kącikom ust przy rozpoznawaniu radości, ignorując nieistotne szumy tła.
- **Faza 5 (klasyfikacja):** To najbardziej rozbudowany etap. Cechy są podawane na dwie w pełni połączone warstwy (Fully Connected) po 4096 jednostek każda, z agresywnym dropoutem 0,5 w celu zapobiegania overfittingowi. Warstwa końcowa Softmax mapuje dane na 7 klas: Surprise, Fear, Disgust, Happy, Sad, Angry, Neutral.

4. Integracja DeepFace i MediaPipe: komplementarność w analizie behawioralnej

W rurociągu (pipeline) przetwarzania robota wykorzystujemy dwie wyspecjalizowane biblioteki, które pełnią rolę cyfrowych receptorów:

- **DeepFace:** Dzięki wykorzystaniu wag modelu VGG, biblioteka ta zapewnia precyzję rzędu 98,61% (LFW) w rozpoznawaniu dominant emocjonalnych. Jest kluczowa dla utrzymania stabilności identyfikacji w zmiennych warunkach oświetlenia.
- **MediaPipe:** Wybrana ze względu na skrajnie niską latencję, idealną dla procesorów Jetson Nano/Raspberry Pi. Śledzi 21 punktów dłoni, 33 punkty ciała (Pose) oraz do 478 punktów twarzy. W diagnostyce ASD (autyzmu) MediaPipe umożliwia automatyczną analizę ruchów stereotypowych (arm flapping, kołysanie się) poprzez oprogramowanie SMDA. System wykorzystuje MediaPipe Visualizer oparty na plikach konfiguracyjnych protobuf (.pbtxt) do definicji grafów przetwarzania, co pozwala na szybką rekalibrację pola widzenia robota.

5. Wskaźniki behawioralne i biomarkery w systemie wizyjnym robota

Obiektywizacja stanu pacjenta opiera się na wyznaczaniu matematycznych wskaźników (biomarkerów):

- **EAR (Eye Aspect Ratio):** Wykrywanie lęku (>20 mrugnięć/min) oraz depresji lub skrajnego zmęczenia (powolne mrugnięcia, EAR < 0,2 przez > 0,5s).
- **MAR (Mouth Aspect Ratio):** Detekcja ziewania oraz hiperwygilancji (nadmierne otwarcie ust w PTSD).
- **Gaze Direction & Iris Tracking:** Landmark 468 pozwala na śledzenie tęczówek. Dzięki stałej fizycznej średnicy tęczówki (11,7 mm), system estymuje odległość pacjenta od robota z dokładnością do ok. 2 cm.
- **Joint Angles:** Analiza 33 punktów Pose pozwala na wykrywanie sztywności chodu (choroba Parkinsona) lub stresu chronicznego poprzez analizę kątów stawowych.
- **rPPG (Zdalna fotopletyzmografia):** System analizuje mikrozmiany koloru skóry za pomocą tzw. "bit planes", co pozwala na nieinwazyjną estymację tętna (HR) i zmienności tętna (HRV) bez kontaktu fizycznego.

6. Etyka, odpowiedzialność i ryzyko w predykcji behawioralnej

Jako inżynierowie musimy balansować między skutecznością a etyką, zwłaszcza w opiece nad pacjentem wrażliwym.

- **Prywatność:** Wszystkie dane biometryczne podlegają natychmiastowej anonimizacji w potoku przetwarzania.
- **Ryzyko klasyfikacji:** Fałszywe wyniki dodatnie (False Positives) niosą ryzyko stygmatyzacji i błędnego leczenia. Z kolei fałszywe wyniki ujemne (False Negatives),

wynikające np. z umiarkowanego wskaźnika recall (ok. 68,07% w modelach detekcji agresji), mogą zagrażać bezpieczeństwu personelu.

- **Bias & Fairness:** Algorytmy mogą powielać uprzedzenia etniczne obecne w danych treningowych, co wymaga ciągłego monitorowania sprawiedliwości algorytmicznej.
- **Human-in-the-loop:** System AI jest wyłącznie narzędziem wsparcia decyzji klinicznej. Ostateczny werdykt diagnostyczny musi należeć do lekarza.

7. Wnioski końcowe i rekomendacje implementacyjne

Model hybrydowy VGG-Transformer wykazuje imponującą gotowość wdrożeniową (95,76% walidacji na RAF-DB). Integracja z MediaPipe i DeepFace tworzy komplementarny ekosystem sensoryczny dla robotyki społecznej.

Rekomendacje inżynieryjne:

1. **Optymalizacja pod NPU/GPU:** Blok Transformera należy akcelerować sprzętowo, aby uniknąć opóźnień w interakcji "twarzą w twarz".
2. **Monitorowanie F1-Score:** Ze względu na imbalans klas widoczny w logach treningowych (gdzie np. klasa 3 – Disgust osiąga znacznie niższą precyzję i recall niż klasa 7 – Neutral), zalecam optymalizację pod kątem F1-Score oraz stosowanie targetowanej augmentacji danych dla niedoreprezentowanych emocji.
3. **Kalibracja dynamiczna:** Wykorzystanie MediaPipe Visualizer i plików .pbtxt do bieżącego dostosowywania progów EAR/MAR w zależności od indywidualnej bazowej charakterystyki pacjenta.

Połączenie zaawansowanej wizji komputerowej z opieką nad zdrowiem psychicznym ma potencjał transformacyjny, redefiniując rolę robotów w medycynie prewencyjnej i diagnostyce afektywnej.