RAPORT NAUKOWO-TECHNICZNY | 1 LISTOPADA 2026 R.

Mobilne śledzenie wzroku i multimodalna analiza afektywna w systemach Edge AI

Przegląd metod, architektur modeli, ograniczeń interpretacyjnych i procedur walidacji

Streszczenie

Przedstawiono krytyczną syntezę metod estymacji spojrzenia na urządzeniach mobilnych oraz automatycznej analizy sygnałów związanych z afektem. Zakres obejmuje obraz RGB, czujniki głębi i podczerwieni, dane inercyjne, ekspresję twarzy, prozodię mowy, tekst oraz sygnały EEG, EKG/PPG i aktywność elektrodermalną. Omówiono sieci CNN, modele rekurencyjne i transformery, mechanizmy fuzji między modalnościami oraz strategie wdrożenia na urządzeniach brzegowych. Wyniki reprezentatywnych badań śledzenia wzroku wskazują na znaczne korzyści personalizacji i wykorzystania dodatkowych sensorów, jednak nie uzasadniają bezpośredniego rankingu modeli testowanych na różnych zbiorach danych. W analizie afektywnej odróżniono mierzalne zachowania i sygnały fizjologiczne od nieobserwowalnych bezpośrednio stanów psychicznych. Zaproponowano architekturę przetwarzania z jawną oceną jakości danych, synchronizacją, obsługą brakujących modalności i niepewności predykcji. Sformułowano protokół walidacji obejmujący skuteczność, generalizację międzyosobniczą, opóźnienie całkowite, energię, ryzyko błędnej interpretacji i zgodność z ochroną danych. Opracowanie ma charakter przeglądu narracyjnego i projektu badawczego; nie przedstawia wyników własnego eksperymentu ani ilościowej metaanalizy standaryzowanych efektów.

Słowa kluczowe: eye tracking; informatyka afektywna; fuzja multimodalna; analiza ekspresji twarzy; estymacja spojrzenia; systemy Edge AI; walidacja; prywatność.

1. Wprowadzenie i cel badawczy

Współczesne systemy interakcji człowiek–komputer mogą łączyć lokalizację punktu spojrzenia na ekranie (point of gaze, PoG) z obserwacjami mimiki, mowy oraz aktywności autonomicznego i ośrodkowego układu nerwowego. Rozwój kamer frontowych i akceleratorów mobilnych umożliwia pozyskiwanie części tych informacji bez kosztownego stanowiska laboratoryjnego. Jednocześnie miniaturyzacja zwiększa znaczenie ograniczeń obliczeniowych, dynamiki naturalnego użytkowania i niejednoznaczności interpretacyjnej sygnałów.

Celem raportu jest zidentyfikowanie rozwiązań właściwych dla mobilnej, lokalnej analizy zachowania i sygnałów afektywnych oraz zaproponowanie sprawdzalnej architektury eksperymentalnej. Problem badawczy dotyczy kompromisu pomiędzy jakością estymacji, odpornością na zakłócenia, kosztami obliczeniowymi i ograniczaniem inwazyjności pomiaru. Główne pytanie ma postać: w jakich warunkach fuzja spojrzenia, zachowań ekspresyjnych i sygnałów fizjologicznych wnosi mierzalną informację ponad bazowe modele jednomodalne, przy akceptowalnym opóźnieniu na urządzeniu brzegowym?

2. Metodologia i granice wnioskowania

Zastosowano narracyjny przegląd metod oraz porównanie wybranych badań i klas modeli. Zestawiono definicje zadań, modalności pomiarowe, architektury, metryki i uwarunkowania wdrożenia. Porównania ilościowe ograniczono do wyników uzyskanych wewnątrz tego samego badania i na tym samym zbiorze danych. Nie przeprowadzono systematycznego przeszukiwania baz według protokołu PRISMA, oceny ryzyka błędu publikacyjnego ani agregacji wielkości efektów z niezależnych badań.

Przyjęto trzy poziomy wniosków: (a) wyniki empiryczne raportowane w konkretnych protokołach; (b) deklarowane właściwości technologii lub architektur; (c) propozycje projektowe wymagające walidacji. W szczególności nazwa klasyfikatora, wynik accuracy i liczba parametrów nie są wystarczającymi przesłankami jego przydatności w innym zadaniu. Dodatkowo nie należy utożsamiać przetwarzania sygnału korelującego z pobudzeniem z rozpoznaniem konkretnej emocji.

3. Podstawy pomiarowe i teoretyczne

3.1. Spojrzenie i ruchy gałek ocznych

W mobilnym eye trackingu rozróżnia się regresję punktu spojrzenia w płaszczyźnie ekranu PoG = (x, y) oraz estymację kierunku spojrzenia w przestrzeni trójwymiarowej. Miary w centymetrach i stopniach zależą od geometrii obserwacji i nie mogą być wymieniane bez określenia odległości oka od ekranu. W analizie zachowania okulomotorycznego istotne są fiksacje, sakkady, trajektorie skanowania, ruchy śledzące i mrugnięcia. Poprawna identyfikacja tych zdarzeń wymaga znajomości częstotliwości próbkowania i jakości sygnału. [1–4]

3.2. Ekspresja, afekt i inferencja

Modele kategoryczne opisują emocje poprzez klasy, natomiast modele wymiarowe przedstawiają afekt m.in. przez walencję i pobudzenie. Facial Action Coding System (FACS) operacjonalizuje widoczne jednostki działania mięśni twarzy (action units, AU), nie przypisując im z definicji jednoznacznego stanu psychicznego. Wnioski o emocjach formułowane na podstawie ekspresji muszą uwzględniać kontekst sytuacyjny, różnice indywidualne i kulturowe, stan zdrowia oraz możliwą celową regulację mimiki.

Sygnały elektrodermalne mogą informować o pobudzeniu współczulnym; EEG o określonych wzorcach aktywności elektrycznej mózgu, natomiast EKG/PPG o pracy układu krążenia i pośrednio o zmienności rytmu serca. Nie są to swoiste markery pojedynczych emocji. Zadanie rozpoznawania emocji należy więc definiować operacyjnie poprzez sposób pozyskania etykiet (samoopis, oceny obserwatorów, protokół zadaniowy), zamiast zakładać istnienie bezpośredniego fizjologicznego „odczytu” emocji.

3.3. Czasowość stanów i obserwacji

Zmiany ekspresji mogą trwać ułamki sekundy, prozodia i reakcje kontekstowe kilka sekund, a sygnały fizjologiczne wykazują również wolniejsze składowe. Dlatego wspólna analiza wymaga jawnego ustalenia rozdzielczości czasowej, długości okien, opóźnień fizjologicznych i sposobu synchronizacji. Stałe okna 0,5–2 s, 2–10 s i kilkudziesięciosekundowe mogą stanowić hipotezy projektowe, ale nie są uniwersalnymi granicami „emocji krótkiej” i „nastroju”.

4. Metody śledzenia wzroku na urządzeniach mobilnych

4.1. Sensory i rodziny algorytmów

Rodzina model-based wykorzystuje geometrię oka, położenie źrenicy i odbicia rogówkowe, zwykle z udziałem NIR/IR w konfiguracjach PCCR. Podejścia feature-based odwzorowują położenie cech obrazu na punkt spojrzenia. Metody appearance-based uczą regresji bezpośrednio z reprezentacji obrazu twarzy i oczu; stanowią podstawę licznych rozwiązań dla zwykłych kamer RGB. Architektury hybrydowe łączą geometrię, landmarki i sieci neuronowe. [1,2,5]

Modalność	Informacja	Potencjalna korzyść	Główne ograniczenie
RGB	twarz, oczy, układ źrenica–powieka	powszechność, niski koszt	oświetlenie, ruch i zasłonięcia
RGB + depth	obraz oraz geometria sceny	kompensacja pozycji/odległości	dostępność sensora, kalibracja
NIR/IR	kontrast oka, refleksy rogówkowe	stabilniejszy pomiar w wybranych warunkach	dedykowany hardware, odbicia
IMU	ruch, orientacja urządzenia	wykrywanie zmiany warunków i rekalkibracja	brak bezpośredniego pomiaru spojrzenia

4.2. Modele referencyjne i wyniki empiryczne

iTracker, trenowany z użyciem GazeCapture, wykazał błędy punktu spojrzenia wynoszące 1,71 cm na telefonach i 2,53 cm na tabletach bez kalibracji, a po kalibracji odpowiednio 1,34 i 2,12 cm. Autorzy raportowali około 10–15 klatek na sekundę na badanym wówczas urządzeniu. Wyniki te dokumentują wykonalność bez dodatkowego sensora, nie wyznaczają jednak współczesnego uniwersalnego progu jakości. [2]

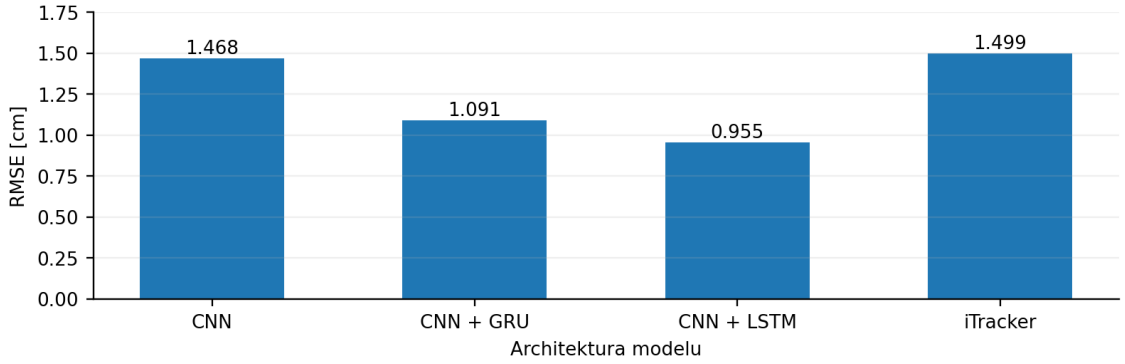
Valliappan i współautorzy odnotowali zmniejszenie średniego błędu z 1,92 ± 0,20 cm do 0,46 ± 0,03 cm po personalizacji obejmującej około 100 ramek kalibracyjnych, w szczególnie korzystnej konfiguracji obejmującej

frontalne ustawienie głowy i dystans 25–40 cm. Wartości opisują określone warunki i nie powinny być uogólniane na swobodny ruch użytkownika. [3]

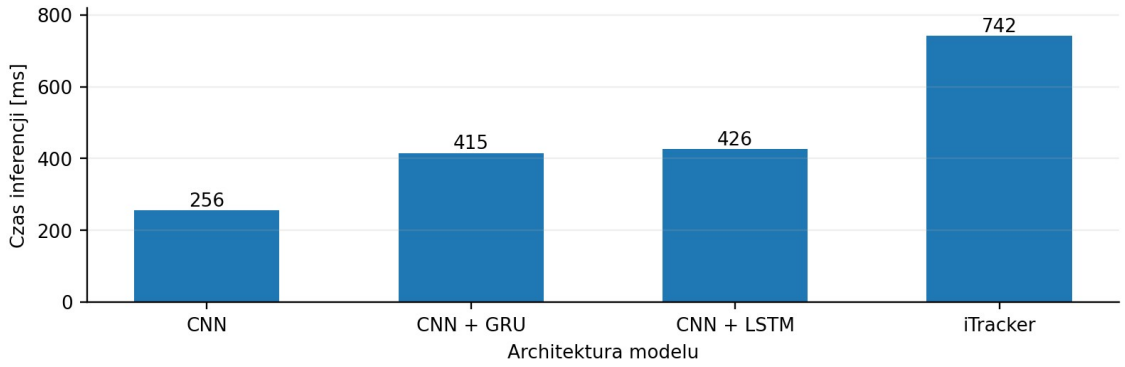
W RGBDGaze integracja obrazu RGB i głębi zmniejszyła średni błąd z 2,26 cm do 1,89 cm, tj. o 16,3%, w porównaniu przeprowadzonym w obrębie tego samego protokołu. Badanie obejmowało różne konteksty użytkowania telefonu. Odrębnym kierunkiem jest wykorzystanie IMU do uruchamiania adaptacyjnej kalibracji po zmianie ruchu lub postawy urządzenia. [4,6]

4.3. Modelowanie czasowe i koszt obliczeniowy

W badaniu obejmującym zbiór DynamicGaze porównano klasyczny CNN z wariantami CNN+GRU, CNN+LSTM oraz iTracker. W analizowanej konfiguracji najmniejszy RMSE uzyskano dla CNN+LSTM (0,955 cm wobec 1,468 cm dla CNN). Odpowiada to około 34,9% redukcji RMSE, ale kosztem dłuższej inferencji. Są to wyniki jednego protokołu, a nie przewaga gwarantowana przy innych zbiorach i urządzeniach. [7]



Rysunek 1. Błąd estymacji w obrębie benchmarku DynamicGaze. Mniejsze wartości oznaczają lepszą estymację. Dane z badania [7].



Rysunek 2. Czas samej inferencji modeli TFLite na Samsung S22 w badaniu [7]; bez detekcji twarzy, akwizycji i przetwarzania wstępnego.

Wyniki czasowe podkreślają, że wysoka jakość regresji i spełnienie wymogu czasu rzeczywistego stanowią odrębne kryteria. Dla aplikacji interaktywnej właściwym miernikiem jest opóźnienie całkowite od pozyskania klatki do udostępnienia wyniku, a nie wyłącznie czas przebiegu sieci. Kwantyzacja, pruning i destylacja wiedzy mogą obniżać wymagania sprzętowe, ale każdą optymalizację należy ocenić ponownie pod kątem błędu, niestabilności i zużycia energii. [7,8]

5. Detekcja zdarzeń okulomotorycznych

W badaniach interfejsów istotne jest przekształcenie szeregu próbek spojrzenia w zdarzenia. I-VT rozróżnia zdarzenia na podstawie prędkości, a I-DT identyfikuje skupienia próbek o małej dyspersji przestrzennej w oknie czasu. Modele probabilistyczne, w tym HMM, mogą wykorzystywać ciągłość stanów, lecz wymagają dodatkowej parametryzacji. Wybór algorytmu wpływa na liczbę, czas trwania i położenie wykrytych fiksacji, a więc także na końcowe wnioski badawcze. [9,10]

Przy ograniczonej częstotliwości próbkowania smartfonu i zmiennym ruchu głowy wiarygodne wykrywanie mikrosakkad jest szczególnie trudne. W badaniach wymagających wysokoczęstotliwościowej okulometrii należałoby zastosować referencyjny eye tracker i oddzielnie zwalidować detekcję drobnych ruchów oczu. Dla mobilnej interakcji

użytkowej większą wartość mogą mieć stabilność punktu spojrzenia i poprawność wykrywania fiksacji niż deklarowana możliwość analizy mikrosakkad.

6. Rozpoznawanie ekspresji i analiza afektywna

6.1. Jednomodalne reprezentacje wizualne

Przetwarzanie wizualne obejmuje detekcję twarzy, normalizację geometrii, ekstrakcję punktów charakterystycznych, analizę tekstury i dynamiki oraz klasyfikację bądź regresję. Wśród sieci konwolucyjnych spotyka się ResNet, EfficientNet i modele z mechanizmami uwagi, w tym DAF-MHSCA, MFER oraz DDAMFN. Wyniki takich modeli są silnie uzależnione od rodzaju etykiet, udziału ekspresji pozowanych, jakości anotacji i dystrybucji obrazów. Prawidłowa klasyfikacja etykiet ekspresji twarzy nie dowodzi bezpośrednio rozpoznania subiektywnie doświadczanej emocji.

6.2. Modele multimodalne i fizjologiczne

Architektury multimodalne łączą cechy obrazu, mowy i tekstu, a w wybranych zastosowaniach EEG, EKG/PPG oraz aktywność elektrodermalną. Przykładowo MemoCMT wykorzystuje transformację między modalnościami, JOYFUL integruje relacje grafowe i uczenie kontrastywne [16], a CNN-BiLSTM-Transformer łączy ekstrakcję lokalnych cech z modelowaniem sekwencji i uwagą. Modele mobilne i noszone, takie jak lekkie CNN-GRU, są projektowane pod ograniczony budżet pamięci i energii. Wyniki pomiędzy benchmarkami MELD, CMU-MOSEI, WESAD, DEAP, AMIGOS i bazami ekspresji twarzy nie są bezpośrednio porównywalne.

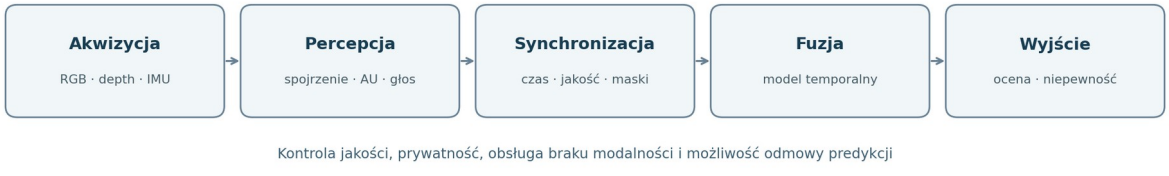
Rodzina / przykład	Modalności	Rodzaj zadania	Krytyczny wymiar oceny
CNN / DAF-MHSCA	obraz twarzy	klasyfikacja ekspresji	F1-macro, generalizacja międzyzbiorowa
CNN–RNN / CNN–GRU	wideo, mowa	dynamika ekspresji i prozodii	CCC / RMSE lub F1, opóźnienie
MemoCMT / JOYFUL	tekst, głos, obraz	afekt w dialogu	F1-macro, kontekst rozmowy
CNN–BiLSTM–Transformer	EEG, EKG/PPG, EDA	klasyfikacja etykiet afektu	LOSO, kalibracja międzyosobnicza
Lekki CNN–GRU / Edge-MER	twarz, głos, czujniki noszone	wnioskowanie na urządzeniu	moc, RAM, P95 opóźnienia
MLLM / AffectGPT	wideo, audio, tekst	opis afektu i uzasadnienie	trafność, kalibracja, halucynacje

W badaniu modelu DAF-MHSCA podano 92,49% accuracy na RAF-DB [14], natomiast dla architektury Attention-Based CNN-BiLSTM-Transformer raportowano 88,2% accuracy na DEAP i 89,5% na AMIGOS [15]. Są to wyniki uzyskane w odrębnych protokołach, a nie niezależnie potwierdzony ranking modeli. W szczególności nie wolno zestawiać w jednej kolumnie accuracy, F1, CCC i RMSE jako równoważnych miar.

6.3. Systemy generatywne i granice deskrypcji

Multimodalne modele językowe mogą generować opis zachowania i jego możliwych interpretacji, lecz opis ma charakter hipotezy zależnej od danych wejściowych i kontekstu. Nie należy traktować swobodnie generowanych zdań typu „użytkownik jest zestresowany” jako miary klinicznej bez zewnętrznego kryterium prawdy, kalibracji i oceny niepewności. Dla zastosowań badawczych właściwszy jest zapis cech obserwowalnych, wyników modelu i poziomu ufności oraz rozdzielenie komentarza wygenerowanego od zmierzonych wielkości.

7. Integracja modalności i proponowana architektura



Rysunek 3. Proponowana architektura wielomodalna z rozdzieleniem percepcji, synchronizacji, fuzji i oceny jakości.

System należy zorganizować w warstwy: akwizycja i synchronizacja sygnałów, kontrola jakości oraz detekcja artefaktów, ekstrakcja cech specyficznych dla modalności, fuzja temporalna, estymacja wraz z niepewnością oraz wyjście aplikacyjne. Warstwa synchronizacji musi uwzględniać opóźnienia pozyskania obrazu, buforów audio i pomiarów fizjologicznych. Braki próbek i utrata jakości jednej modalności nie powinny być niejawnie zastępowane wartościami zero.

W fuzji wczesnej łączy się cechy wejściowe, w fuzji późnej wyniki modeli jednomodalnych, a w fuzji pośredniej reprezentacje ukryte (np. cross-attention). Każda strategia wymaga kontrolowanego eksperymentu ablacyjnego. Podstawową konfiguracją odniesienia powinna być pojedyncza modalność o najwyższej jakości w danym zadaniu; efekt dołączenia kolejnych sensorów należy raportować wraz z kosztami pamięci, energii i ryzykiem braku danych.

W szczególności danych o spojrzeniu nie należy traktować jako bezpośredniej etykiety emocji. Mogą one wnosić informację o alokacji uwagi wzrokowej, przeszukiwaniu sceny lub reakcji na bodziec, lecz interpretacja afektywna wymaga kontekstu eksperymentalnego i niezależnego pomiaru kryterialnego. To rozróżnienie powinno być zachowane zarówno w strukturze modelu, jak i przy prezentacji wyników.

8. Wdrażanie na urządzeniach brzegowych

8.1. Budżet opóźnienia i zasobów

Opóźnienie końcowe można zapisać jako T_total = T_capture + T_preprocess + T_inference + T_fusion + T_output + T_queue. Składnik kolejkowania bywa pomijany, choć wzrost liczby buforowanych klatek może silnie zwiększyć czas odpowiedzi. W systemach interaktywnych wskazane są ograniczone kolejki i polityka wyboru najnowszej próbki, z monitorowaniem liczby odrzuconych danych. Oprócz średniej należy raportować medianę, P95 i maksima w reprezentatywnych warunkach.

Wydajność należy badać na docelowym CPU/GPU/NPU, dla określonych rozdzielczości, formatów numerycznych i trybu zasilania. Redukcja liczby parametrów nie gwarantuje skrócenia latencji: krytyczne mogą być transfery pamięci, wstępna detekcja twarzy i dostępność zoptymalizowanych operatorów. Platformy takie jak smartfony z TFLite/Core ML, układy Rockchip RK3588 i NVIDIA Jetson reprezentują odrębne środowiska programowe; benchmark jednego nie może zastąpić pomiaru na drugim.

8.2. Optymalizacja

Kwantyzację INT8/FP16, pruning, destylację wiedzy oraz redukcję częstości próbkowania należy weryfikować metodą porównań sparowanych, zachowując ten sam podział testowy. Oprócz średniego błędu istotne są degradacja w słabym świetle, błędy przy okularach, wzrost obciążenia termicznego, spadki FPS po długiej pracy i niestabilność opóźnienia. Odciążanie obliczeń do chmury wymaga odrębnego oszacowania zysków i ryzyka transmisji danych wrażliwych.

9. Protokół walidacji eksperymentalnej

Walidację należy zaprojektować jako zestaw z góry określonych zadań: regresji spojrzenia, detekcji zdarzeń okulomotorycznych, klasyfikacji ekspresji, regresji walencji/pobudzenia oraz wspólnej inferencji na danych multimodalnych. Etykiety powinny pochodzić z jawnie opisanej procedury i zostać niezależnie ocenione pod kątem rzetelności. Należy uwzględnić przynajmniej scenariusze ze stabilnym telefonem, swobodnym trzymaniem urządzenia, ruchem głowy i zmianą oświetlenia.

Obszar	Miara podstawowa	Wymóg sprawozdawczy
Regresja spojrzenia	MAE/RMSE [cm], błąd kątowy [°]	geometria ekranu, odległość, kalibracja
Fiksacje/sakkady	precision, recall, F1 zdarzeń	częstotliwość, progi, tolerancje czasu
Ekspresja / klasy	F1-macro, macierz pomyłek	definicja i źródło etykiet
Walencja / pobudzenie	MAE/RMSE, CCC	skala, opóźnienia etykiet
Fuzja modalności	Δ jakości względem baseline	ablacje, brakujące sensory
Generalizacja	wyniki LOSO, test zewnętrzny	brak przecieku osób i sesji
Zasoby	latencja end-to-end P50/P95, RAM, energia	model telefonu, tryb zasilania, czas testu
Bezpieczeństwo i jakość	ECE/Brier, odsetek abstencji	kalibracja i scenariusze awaryjne

W podziale danych należy oddzielić uczestników i sesje między zbiorem treningowym, walidacyjnym i testowym. Ocena leave-one-subject-out (LOSO) jest szczególnie istotna dla fizjologii i ekspresji, gdzie model może nauczyć się cech osoby zamiast właściwego związku z etykietą. Personalizację należy raportować oddzielnie od jakości modelu zero-calibration. Dla powtarzanych pomiarów należy stosować przedziały ufności uwzględniające zależności wewnątrzosobnicze.

Proponowana procedura obejmuje: etap pilotażowy stabilności rejestracji, zablokowanie protokołu i metryk przed analizą końcową, zbieranie danych w losowej kolejności warunków, ablacje modalności oraz test na niezależnej grupie. W analizie należy odróżnić efekt statystyczny od znaczenia praktycznego, uwzględniając koszty energetyczne i czasowe. Liczebność próby trzeba wyznaczyć z analizy mocy dla wcześniej zdefiniowanego efektu, a nie ustalać wyłącznie na podstawie dogodności rekrutacji.

10. Ryzyka naukowe, etyczne i regulacyjne

10.1. Niepewność interpretacji i uprzedzenia modelu

Najważniejszym zagrożeniem jest nadinterpretacja: pojedyncza ekspresja twarzy, fiksacja lub skok aktywności elektrodermalnej nie identyfikują jednoznacznie emocji. Na jakość predykcji mogą wpływać wiek, jakość sensora, oświetlenie, okulary, warunki kliniczne, różnice międzykulturowe i rodzaj zadania. Z tego względu wyniki należy prezentować jako estymaty ograniczone do danego protokołu, z możliwością powstrzymania się od decyzji przy niskiej jakości sygnału.

10.2. Prywatność i minimalizacja danych

Rejestracje twarzy, oka, głosu i danych fizjologicznych mogą umożliwiać identyfikację osoby lub ujawniać informacje szczególnie wrażliwe. Zalecane są przetwarzanie lokalne tam, gdzie jest wykonalne, jawne zgody lub inna właściwa podstawa prawna, minimalizacja zakresu rejestrowanych danych, rozdzielenie identyfikatorów od sygnałów oraz określenie terminów retencji. Dla eksperymentów należy opisać możliwość wycofania udziału, kontrolę dostępu do danych i procedurę zgłaszania incydentów.

10.3. Zastosowania rozpoznawania emocji w UE

Rozporządzenie (UE) 2024/1689, art. 5 ust. 1 lit. f, zakazuje określonych zastosowań systemów AI do wnioskowania o emocjach osób fizycznych w środowisku pracy i instytucjach edukacyjnych, z wyjątkiem zastosowań z przyczyn medycznych lub bezpieczeństwa. Zakaz ten ma odrębny zakres od dopuszczalności samego prowadzenia badań nad sygnałami psychofizjologicznymi. Każde konkretne wdrożenie wymaga kwalifikacji celu, funkcji systemu i podstaw przetwarzania danych zgodnie z aktualnym stanem prawa. [13]

System nie powinien automatycznie przypisywać osobie diagnozy psychologicznej, rekomendacji dotyczącej zatrudnienia ani oceny zdolności ucznia na podstawie samej estymacji ekspresji. Zastosowania w medycynie lub bezpieczeństwie wymagają niezależnej walidacji i oceny właściwego reżimu prawnego; sam fakt działania lokalnego nie zwalnia z tych wymagań.

11. Proponowany program badawczy i wdrożeniowy

1. Etap A – punkt odniesienia: regresja PoG na kamerze RGB, prosta metoda detekcji fiksacji oraz odrębny klasyfikator ekspresji; pomiar jakości i pełnej latencji.
2. Etap B – odporność: zmiana postawy i oświetlenia; osobno ocena personalizacji, głębi i wsparcia IMU przy wykrywaniu dryfu.
3. Etap C – fuzja: dołączenie prozodii i sygnałów fizjologicznych w kontrolowanych zadaniach; ablacją wykazać dodatkową informację każdej modalności.
4. Etap D – Edge AI: porównania INT8/FP16, destylacji i ograniczania rozdzielczości; pomiary P95 latencji, energii, pamięci i temperatury.
5. Etap E – replikowalność: ocena LOSO, niezależny test, dokumentacja protokołu, wersjonowanie danych i modeli, publikacja kodu bez danych osobowych.

W pierwszej fazie rekomendowane jest ograniczenie liczby celów predykcyjnych: odróżnić zadanie okulomotoryczne (mierzalny PoG) od zadania afektywnego (etykieta zależna od definicji), a następnie dopiero badać ich łączną wartość. Takie rozdzielenie zmniejsza ryzyko pozornej poprawy wynikającej z przecieku informacji o uczestniku lub scenariuszu.

12. Dyskusja i wnioski

Mobilne śledzenie wzroku osiągnęło poziom umożliwiający eksperymenty i wybrane zastosowania interakcyjne na powszechnym sprzęcie, zwłaszcza gdy dopuszcza się kalibrację. RGB pozostaje punktem wyjścia o najniższym progu wdrożenia, natomiast depth, IMU oraz modele temporalne stanowią uzasadnione kierunki zwiększania odporności. Poprawa dokładności może jednak zwiększać opóźnienie i zapotrzebowanie na energię, dlatego rekomendowanie „najlepszego” modelu bez określenia zadania i sprzętu jest nieuzasadnione.

W analizie afektywnej najważniejsze jest zachowanie granicy między obserwowalnym sygnałem a wnioskowaniem o stanie wewnętrznym. Fuzja multimodalna może poprawiać predykcję określonych etykiet, lecz jej korzyści trzeba wykazywać w obrębie wspólnych protokołów z prawidłową walidacją międzyosobniczą. Nie należy utożsamiać wzrostu accuracy na benchmarku z trafnością psychologiczną w warunkach naturalnych.

Najbardziej uzasadnioną strategią jest modułowa architektura Edge AI: lokalna akwizycja, kontrola jakości, zsynchronizowana fuzja wybranych modalności, jawna niepewność i testy generalizacji. Przyszłe badania powinny koncentrować się na odporności w realnych warunkach, replikowalnych metrykach czasu rzeczywistego, rzetelnych etykietach afektywnych oraz ocenie wpływu systemu na prywatność i autonomię użytkownika.

Bibliografia

[1] Lei Y., He S., Khamis M., Ye J. An End-to-End Review of Gaze Estimation and its Interactive Applications on Handheld Mobile Devices. ACM Computing Surveys, 2023. DOI: 10.1145/3606947.

[2] Krafka K. i in. Eye Tracking for Everyone. Proceedings of CVPR, 2016, s. 2176–2184. https://openaccess.thecvf.com/content_cvpr_2016/html/Krafka_Eye_Tracking_for_CVPR_2016_paper.html

[3] Valliappan N. i in. Accelerating eye movement research via accurate and affordable smartphone eye tracking. Nature Communications, 2020, 11, 4553. DOI: 10.1038/s41467-020-18360-5.

[4] Arakawa R., Goel M., Harrison C., Ahuja K. RGBDGaze: Gaze Tracking on Smartphones with RGB and Depth Data. Proceedings of ICMI, 2022. DOI: 10.1145/3536221.3556568.

[5] Cheng Y., Wang H., Bao Y., Lu F. Appearance-based Gaze Estimation With Deep Learning: A Review and Benchmark. arXiv:2104.12668, 2021.

[6] Lei Y. i in. MAC-Gaze: Motion-Aware Continual Calibration for Mobile Gaze Tracking. arXiv:2505.22769, 2025.

[7] Gunawardena N., Lui G. Y., Ginige J. A., Javadi B. Smartphone-based Eye Tracking System using Edge Intelligence and Model Optimisation. arXiv:2408.12463, 2024.

[8] Gou J., Yu B., Maybank S. J., Tao D. Knowledge Distillation: A Survey. International Journal of Computer Vision, 2021. DOI: 10.1007/s11263-021-01453-z.

[9] Salvucci D. D., Goldberg J. H. Identifying Fixations and Saccades in Eye-Tracking Protocols. Proceedings of ETRA, 2000.

[10] Andersson R., Larsson L., Holmqvist K., Stridh M., Nyström M. One algorithm to rule them all? An evaluation and discussion of ten eye movement event detection algorithms. Behavior Research Methods, 2017. DOI: 10.3758/s13428-016-0738-9.

[11] Khamis M., Alt F., Bulling A. The Past, Present, and Future of Gaze-enabled Handheld Mobile Devices. Proceedings of MobileHCI, 2018. DOI: 10.1145/3229434.3229452.

[12] Huang J., White R., Buscher G. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. Proceedings of IJCAI, 2016.

[13] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z 13 czerwca 2024 r. (AI Act), art. 5 ust. 1 lit. f. https://eur-lex.europa.eu/eli/reg/2024/1689

[14] A Directional Attention Fusion and Multi-Head Spatial-Channel Attention Network for Facial Expression Recognition. 2025, preprint. https://www.researchgate.net/publication/399038874

[15] Emotion Recognition Algorithm Based on Multi-Modal Physiological Signal Feature Fusion Using Artificial Intelligence and Deep Learning. International Journal of Advanced Computer Science and Applications, 2025, 16(6). https://thesai.org/Publications/ViewPaper?Code=IJACSA&Issue=6&SerialNo=13&Volume=16

[16] Joyful: Joint Modality Fusion and Graph Contrastive Learning for Multimodal Emotion Recognition. Proceedings of EMNLP, 2023. https://aclanthology.org/2023.emnlp-main.996/