Akademia Górniczo-Hutnicza im. St. Staszica w Krakowie
Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki

Rozprawa doktorska

# Automatyczne rozpoznawanie elementów mimiki w obrazie twarzy i analiza ich przydatności do sterowania

**Jaromir Przybyło**

Promotor: dr hab. inż. Piotr Augustyniak

Kraków, 2008

**Rodzicom**

*za to że wspierali mnie cały czas*

**Mirkowi J.**

*za motywowanie mnie zarówno do pracy jak i do odpoczynku*

**Izie G.**

*za wyrozumiałość, troskę oraz motywowanie do pogłębiania wiedzy w innych dziedzinach*

oraz
**Filipowi W.**

*życząc mu w przyszłości kariery naukowej (o ile będzie to jego marzeniem)*

# Spis treści

**Spis treści** **i**

**I O rozprawie** **3**

**1 Wstęp** **5**
1.1 Kontekst pracy, uzasadnienie podjęcia tematu . . . . . 5
1.2 Zakres, teza oraz cele rozprawy . . . . . 6
1.3 Streszczenie rozprawy . . . . . 8

**II Elementy mimiki** **11**

**2 Sposoby pomiaru oraz opisu ekspresji mimicznych twarzy** **13**
2.1 Mimika twarzy . . . . . 13
2.2 Facial Action Coding System . . . . . 15
2.3 MPEG-4 FAP . . . . . 18

**3 Określenie i wybór rozpoznawanych elementów mimiki** **21**
3.1 Systematyka typowych zadań wykonywanych przez użytkowników . 21
3.2 Własności i systematyka urządzeń wejściowych . . . . . 23
3.3 Analiza możliwości jakie oferuje mimika . . . . . 24
3.4 Wybór elementów mimiki potencjalnie przydatnych do sterowania 32

**4 Atrybuty elementów mimiki** **37**
4.1 Wstęp . . . . . 37
4.2 Atrybuty stałe — kształt . . . . . 38
4.3 Atrybuty zmienne — lokalne zmiany wyglądu . . . . . 39
4.4 Atrybuty dynamiczne — ruch i jego trajektoria . . . . . 40

**III Automatyczne rozpoznawanie elementów mimiki 41**

**5 Elementy systemu automatycznego rozpoznawania mimiki 43**
5.1 Wstęp . . . 43
5.2 Selektywne przetwarzanie informacji, segmentacja . . . 46
5.2.1 Segmentacja na podstawie barwy skóry . . . 48
5.2.2 Dobór przestrzeni kolorów oraz metody segmentacji . . . 50
5.3 Detekcja i lokalizacja twarzy . . . 55
5.3.1 Algorytm detekcji i lokalizacji twarzy . . . 57
5.3.2 Rezultaty detekcji i lokalizacji twarzy . . . 59

**6 Wyodrębnianie z obrazu twarzy elementów mimiki 63**
6.1 Wstęp . . . 63
6.2 Statystyczne modele kształtu . . . 64
6.3 Histogramy orientacji . . . 67
6.4 Detekcja ruchu . . . 69

**7 Rozpoznawanie wybranych elementów mimiki 75**
7.1 Wstęp . . . 75
7.2 Opis wybranych metod klasyfikacji . . . 76
7.3 Rozpoznawanie elementów mimiki — rezultaty i wnioski . . . 78
7.3.1 Statystyczne modele kształtu . . . 80
7.3.2 Histogramy orientacji . . . 81

**8 Adaptacja systemu rozpoznawania mimiki 87**
8.1 Wstęp . . . 87
8.2 Estymacja położenia głowy człowieka względem kamery . . . 89
8.2.1 Algorytm estymacji położenia głowy . . . 91
8.2.2 Rezultaty estymacji położenia głowy . . . 96
8.3 Wpływ parametrów kamery oraz zmian oświetlenia sceny . . . 99
8.3.1 Automatyzacja tworzenia modelu barwy skóry . . . 100
8.4 Metoda kalibracji systemu . . . 103
8.4.1 Detekcja mrugnięć . . . 107
8.4.2 Lokalizacja nozdrzy . . . 109

**9 Podsumowanie 113**
9.1 Główne rezultaty i wnioski . . . 113
9.2 Kierunki dalszych badań . . . 117

**Bibliografia 121**

**A Opis metod i algorytmów 131**
A.1 Przestrzenie barw . . . 131

A.2 Metody segmentacji twarzy . . . 132
A.3 Estymacja i usuwanie szumów z sekwencji video . . . 135
A.4 Metoda PCA . . . 137
A.5 Metoda przestrzeni skali (ang. scale-space) . . . 140
A.6 Filtry Gabora . . . 141
A.7 Wyznaczanie mapy prawdopodobieństwa oczu oraz ust . . . 142

**B Opis jednostek czynnościowych mimiki 145**

**C Rezultaty rozpoznawania elementów mimiki — wykresy i tabele 161**
C.1 Informacje ogólne . . . 161
C.2 Histogramy orientacji — konfiguracja nr 1 . . . 162
C.3 Histogramy orientacji — konfiguracja nr 2 . . . 164
C.4 Histogramy orientacji — konfiguracja nr 3 . . . 165
C.5 Histogramy orientacji — konfiguracja nr 4 . . . 166
C.6 Histogramy orientacji — badanie wpływu oświetlenia . . . 167
C.7 Histogramy orientacji — badanie wpływu skali . . . 168
C.8 Histogramy orientacji — badanie wpływu skali i rotacji . . . 169
C.9 Statystyczne modele kształtu — test 1 . . . 170
C.10 Statystyczne modele kształtu — test 2 . . . 171
C.11 Statystyczne modele kształtu — test 3 . . . 172
C.12 Statystyczne modele kształtu — test 4 . . . 173

**D Rezultaty segmentacji twarzy 175**

**E Rezultaty detekcji i lokalizacji twarzy 181**

**F Rezultaty estymacji położenia głowy — wykresy i tabele 183**
F.1 Informacje ogólne . . . 183
F.2 Sekwencja nr 1 — rezultaty . . . 184
F.3 Sekwencja nr 2 — rezultaty . . . 185
F.4 Sekwencja nr 3 — rezultaty . . . 186

**G Rezultaty detekcji mrugnięć 189**
G.1 Informacje ogólne . . . 189
G.2 Skuteczność detekcji . . . 191
G.3 Średni błąd lokalizacji oczu . . . 193

**H Rezultaty innych algorytmów 195**
H.1 Lokalizacja nozdrzy . . . 195

# Podziękowania

Autor chciałby serdecznie podziękować promotorowi dr hab inż. Piotrowi Augustyniakowi, za wiele inspirujących uwag które pomogły ukształtować niniejszą rozprawę doktorską, jego czas poświęcony w okresie bezpośredniej pracy nad rozprawą, jak również za okazywane zrozumienie i cierpliwość.

Podziękowania należą się również całemu zespołowi Laboratorium Biocybernetyki AGH, pod kierownictwem profesora Ryszarda Tadeusiewicza, a w szczególności:

dr inż. Zbigniewowi Mikrutowi za dobre rady i okazywane w każdej sytuacji poczucie humoru

dr n. techn. lek. med. Pawłowi Woloszynowi za inspirujące rady

dr hab inż. Markowi Gorgoniowi za pomoc na różnych etapach pracy

oraz mgr. inż. Mirosławowi Jabłońskiemu za motywowanie mnie do pracy oraz za wiele przydatnych pomysłów.

# Część I

# O rozprawie

# Rozdział 1

# Wstęp

## 1.1 Kontekst pracy, uzasadnienie podjęcia tematu

Głównym sposobem interakcji człowieka z komputerem jest — i zapewne długo jeszcze pozostanie — interfejs graficzny, oparty na technice okien, manipulacji myszką i wprowadzania danych z pomocą klawiatury. Możliwości wprowadzania danych, rozszerzane są poprzez szereg urządzeń, bardziej specjalizowanych dla określonych zastosowań. W przypadku rozrywki (gry) klasycznym przykładem są różnego rodzaju manipulatory (joystick, trackball) bądź też gamepady do gier. W zastosowaniach bardziej profesjonalnych (np. aplikacje typu CAD) wykorzystywane są również ekrany dotykowe, tablety czy też manipulatory w postaci „rękawicy" pozwalającej na pracę z interfejsem trójwymiarowym (3D).

Obecnie obserwuje się zmianę podejścia w konstruowaniu interfejsów komputerowych w kierunku wykorzystania wielu równoległych sposobów komunikacji między użytkownikiem a maszyną. Podejście to często określane jest w literaturze jako interfejsy multimodalne (ang. multimodal interfaces) [63]. Przykładem takiego sposobu interakcji jest system „Put That There" opracowany na uniwersytecie MIT [11]. Interfejs ten bazował na rozpoznawaniu gestów wskazywania (urządzenie elektroniczne zapewniające sygnał o pozycji i orientacji) oraz mowy, jak również zapewniał generację mowy i wizualizację na dużym ekranie ściennym. Dzięki połączeniu dwóch metod komunikacji skuteczność i wygoda interakcji została zwiększona. Innym przykładem, ukierunkowanym na zapewnienie możliwości pracy z komputerem osobom niewidomym, jest aplikacja „Meditor: Multimode Text Editor" wykorzystująca zarówno standardową klawiaturę jak i specjalny terminal Braille'a [5]. Dodatkowo Meditor używa technologii syntezy (TTS ang. text to speech) oraz rozpoznawania mowy (ASR ang. automated speech recognition).

W katedrze Automatyki AGH prowadzono pionierskie prace, których celem

było zapewnienie komunikacji głosowej z pomocą sygnału mowy [81]. Prowadzone są również badania mające na celu wykorzystanie do sterowania czujników potencjałów bioelektrycznych, na przykład elektronystagmografii, elektromiografii lub elektroencefalografii [12]. W tym kontekście wymienić również należy prace, których celem jest opracowanie systemu adaptacyjnego dla potrzeb zdalnego nadzoru kardiologicznego pacjentów [1].

Na świecie istnieją także projekty badawcze poświęcone wykorzystaniu informacji wizyjnej (obraz) do sterowania. Przeznaczony dla osób niepełnosprawnych ruchowo, interfejs oparty o metody wyszukiwania wzorców na obrazie, może być również wykorzystany przez osoby w pełni sprawne [6]. W katedrze Automatyki AGH prowadzone są badania mające na celu określenie reguł postrzegania naturalnego przez człowieka w celu ich wykorzystania w inteligentnych systemach wizyjnych [2][61].

Obserwując komunikację człowieka z innymi ludźmi łatwo można stwierdzić, że bardzo ważna jest tu komunikacja niewerbalna, której istotnym elementem jest mimika twarzy. Ten kanał łączności między człowiekiem a komputerem i innymi systemami technicznymi (na przykład robotem medycznym albo wózkiem inwalidzkim) odgrywa szczególną rolę w przypadku niektórych osób dotkniętych szczególnie głęboką niepełnosprawnością. W przypadku gdy kalectwo albo choroba odbiorą człowiekowi zręczność rąk konieczną do operowania myszką czy klawiaturą, gdy te same przyczyny utrudnią artykułowanie wyrazistych, nadających się do automatycznej interpretacji wypowiedzi słownych — mimika pozostaje jednym z ostatnich kanałów łączności chorego ze światem, w tym także ze światem systemów technicznych. Przykładem niepełnosprawności, która praktycznie eliminuje użycie tradycyjnych metod komunikacji człowieka z komputerem (myszka, klawiatura), są osoby z porażeniem czterokończynowym. W tym przypadku podstawowym problemem są ograniczenia sprawności ruchowej takich osób. Jedyną możliwością ruchu jest mimika twarzy, prawidłowa ruchomość żuchwy i języka oraz minimalne ruchy głowy. Dodatkowym utrudnieniem jest niewyraźna mowa, wynikająca ze stosowanych w wielu przypadkach elektronicznych stymulatorów oddechu.

## 1.2 Zakres, teza oraz cele rozprawy

W sytuacjach ograniczenia sprawności ruchowej alternatywę w sposobie sterowania komputerem, mogą stanowić metody i algorytmy rozpoznawania oraz analizy obrazów. Podstawową zaletą tego typu interfejsów sterujących jest możliwość dostosowania interfejsu do człowieka oraz sytuacji a nie odwrotnie. Trend taki można zaobserwować w aktualnie prowadzonych badaniach — zmiana podejścia w konstruowaniu interfejsów komputerowych z architektur skoncentrowanych na maszynie (ang. machine-centered architecture) do architektur skupionych na

użytkowniku (ang. human-centered interaction architecture) [55].

Obserwacja optyczna nie wymaga przeprowadzania odpowiedniej procedury dokowania użytkownika do stanowiska pracy; wystarczające jest, aby użytkownik znalazł się w obszarze działania systemu. Podobnie zakończenie pracy nie wymaga odłączania stanowiska lub odłączania urządzeń. W przypadku osoby poruszającej się z pomocą wózka lub innego pojazdu rozpoczęcie i zakończenie pracy z komputerem może więc nie wymagać udziału opiekunów. Samo stanowisko pracy może być także łatwo przenoszone. Ogromną zaletą jest również możliwość dowolnego wyboru obserwowanej okolicy ciała bez zmiany konstrukcji systemu. Daje to możliwość zastosowania nawet przy skrajnym ograniczeniu sprawności ruchowej użytkownika i zawężeniu jej na przykład do mimiki twarzy lub jednego palca ręki. System taki stwarza też możliwość jednoczesnego obserwowania dwóch różnych, nawet odległych okolic ciała, jeśli ruchomość każdej z nich osobno nie wystarcza do wyrażenia wszystkich poleceń sterujących.

Zaletą tego typu interfejsów sterujących jest również bezkontaktowość interakcji. Brak fizycznego kontaktu elementów systemu z ciałem użytkownika zapobiega niedogodnościom higienicznym i estetycznym, usuwa też zagrożenia elektryczne wymagające stosowania zabezpieczeń przed porażeniem prądem. System wizyjny nie wymaga stosowania dodatkowych elementów pogarszających wygodę użytkowania, jak różnego rodzaju elektrody, czujniki, opaski, okulary, ustniki i tym podobne.

Należy również nadmienić, iż system oparty o rozpoznawanie obrazów może również służyć osobom w pełni sprawnym, rozszerzając w ten sposób możliwości wymiany informacji między człowiekiem a maszyną.

**Prowadzone w ramach niniejszej rozprawy badania, ograniczono do rozpoznawania elementów mimiki przy użyciu metod i algorytmów analizy obrazów**. Wybór elementów mimiki, jako sposobu przekazywania informacji, podyktowany został faktem — potwierdzanym przez wyniki wielu badań psychologicznych — iż nie licząc głosu, dużą część informacji człowiek przekazuje przy pomocy komunikacji niewerbalnej, w tym dzięki mimice twarzy. W kontekście systemu dla osób niepełnosprawnych wydaje się to szczególnie istotne ze względu na zasygnalizowane wcześniej ograniczenia sprawności ruchowej. Celem prowadzonych prac jest również określenie sposobów wykorzystania elementów mimiki do rozszerzenia możliwości komunikacji między użytkownikiem a komputerem. Uwzględniając powyższą analizę sformułowano następującą tezę rozprawy:

**Elementy mimiki automatycznie wyodrębnione z cyfrowego obrazu twarzy są przydatne w komunikacji człowiek-maszyna i mogą być wykorzystane do sterowania.**

Aby wykazać słuszność powyższej tezy sformułowano następujące główne **cele rozprawy**:

1. Określenie sposobu pomiaru oraz precyzyjnego opisu elementów mimiki.
2. Określenie możliwości jakie oferuje mimika w kontekście typowych scenariuszy interakcji człowieka z maszyną i własności istniejących urządzeń. Wybór elementów mimiki przydatnych w komunikacji człowiek-maszyna.
3. Zdefiniowanie atrybutów elementów mimiki oraz odpowiadających im cech charakterystycznych na obrazie, pozwalających na rozpoznanie gestów mimicznych.
4. Opracowanie metody automatycznego rozpoznawania wybranych elementów mimiki.

Realizacja celu sformułowanego w punkcie -4- wymagała określenia następujących celów składowych:

a) Zaproponowanie struktury i elementów składowych systemu automatycznego rozpoznawania gestów mimicznych.

b) Opracowanie metod wyodrębniania z obrazu twarzy cech charakterystycznych, odpowiadających atrybutom rozpoznawanych elementów mimiki.

c) Opracowanie metod rozpoznawania gestów mimicznych wykorzystujących wyodrębnione cechy. Badanie skuteczności algorytmów dla typowych sytuacji występujących podczas interakcji człowieka z maszyną.

d) Usystematyzowanie czynników wpływających na skuteczność rozpoznawania oraz opracowanie metody adaptacji systemu do człowieka oraz zmieniających sie warunków otoczenia.

## 1.3 Streszczenie rozprawy

Struktura rozprawy jest następująca:

- W części pierwszej (rozdziały 2 - 4) przedstawiono sposoby pomiaru oraz opisu ekspresji mimicznych twarzy, dokonano wyboru elementów mimiki oraz określono ich atrybuty i cechy.

- Część druga (rozdziały 5 - 8) poświęcona została zagadnieniom automatycznego wyodrębniania z obrazu i rozpoznawania elementów mimiki. Przedstawiono również kwestie adaptacji systemu.
- Zestawienie wyników pracy i najistotniejszych wniosków oraz krytyczną analizę zaproponowanych rozwiązań przedstawiono w podsumowaniu (rozdział 9).
- Rezultaty przeprowadzonych badań oraz opisy algorytmów zamieszczone zostały w dodatkach.

W poszczególnych rozdziałach zagadnieniami wiodącymi są:

- **Rozdział 2 — Sposoby pomiaru oraz opisu ekspresji mimicznych twarzy.** Wykorzystanie elementów mimiki do sterowania, wymaga uwzględnienia szeregu zagadnień, wśród których szczególnie istotne są: psychofizjologiczne uwarunkowania człowieka oraz sposób reprezentacji i opisu mimiki. W rozdziale 2 przedstawione zostały zagadnienia związane z komunikacją niewerbalną (istotne w kontekście wykorzystania mimiki do sterowania) oraz istniejące sposoby opisu ekspresji mimicznych twarzy. Spośród metod reprezentacji mimiki szczegółowo opisane zostały dwie z nich, wykorzystane w dalszej części pracy.
- **Rozdział 3 — Określenie i wybór rozpoznawanych elementów mimiki.** Skonstruowanie urządzenia wejściowego umożliwiającego sterowanie przy pomocy gestów mimicznych, wymaga uwzględnienia wymagań wynikających z typowych zadań jakie wykonuje użytkownik podczas pracy z komputerem. W rozdziale 3 przedstawiono usystematyzowanie typowych zadań wykonywanych przez człowieka podczas interakcji z interfejsem graficznym komputera. Poprzez analizę akcji użytkownika oraz porównanie własności typowych urządzeń wejściowych z możliwościami jakie oferuje mimika, dokonano wyboru elementów mimiki, które są potencjalnie przydatne w komunikacji człowiek-maszyna.
- **Rozdział 4 — Atrybuty elementów mimiki.** Rozpoznawanie różnych elementów mimiki w oparciu o informacje wyodrębnione z obrazu, wymaga określenia atrybutów oraz cech charakterystycznych opisujących poszczególne gesty mimiczne (np. kształt, zależności geometryczne, wygląd...). Temu zagadnieniu poświęcony jest rozdział 4. Dokonana analiza elementów mimiki pozwoliła na zdefiniowanie podstawowych atrybutów oraz odpowiadających im cech, które mogą zostać następnie wyodrębnione z obrazu.
- **Rozdział 5 — Elementy systemu automatycznego rozpoznawania mimiki.** Rozpoznawanie obiektów przez ludzi wydaje się łatwe i bezproblemowe. Na podstawie ogólnej wiedzy o danym obiekcie, np. sylwetka ludzka

— człowiek natychmiast jest w stanie rozpoznać nowe osoby, niezależnie od tego z jakiego punktu widzenia są one obserwowane czy też np. siedzą lub stoją. Z punktu widzenia komputerowej analizy obrazów, rozpoznawanie wymaga wyodrębnienia z obrazu użytecznej informacji na temat interesujących obiektów. W rozdziale 5 zaproponowano schemat struktury systemu automatycznego wyodrębniania z obrazu elementów mimiki, zawierający kilka bloków składowych realizujących wymagane zadania. Omówione zostały etapy przetwarzania i analizy obrazu, na które składają się: selektywne przetwarzanie informacji (segmentacja), detekcja i lokalizacja twarzy, wyodrębnianie cech, klasyfikacja oraz adaptacja systemu. Wśród istotnych rezultatów wymienić można opracowaną metodę doboru przestrzeni barw dla algorytmu segmentacji oraz algorytm detekcji twarzy.

- **Rozdział 6 — Wyodrębnianie z obrazu twarzy elementów mimiki.** Zdefiniowane w pierwszej części rozprawy atrybuty elementów mimiki oraz odpowiadające im cechy oczekiwane na obrazie twarzy, stanowią podstawę umożliwiającą rozpoznanie wybranych gestów mimicznych. W pierwszej kolejności odpowiednie cechy (kształt, wygląd...) muszą zostać wyodrębnione z obrazu. W rozdziale 6 przedstawiono wybrane metody i algorytmy pozwalające na ekstrakcję z obrazu twarzy wymaganych na etapie rozpoznawania wektorów cech, reprezentujących elementy mimiki. Wybrane metody to: statystyczne modele kształtu, histogramy orientacji oraz detekcja ruchu.

- **Rozdział 7 — Rozpoznawanie wybranych elementów mimiki.** W zadaniu rozpoznawania celem jest określenie przynależności różnego typu obiektów (w tym przypadku elementów mimiki) do pewnych klas na podstawie znanych wcześniej przynależności do klas innych obiektów. Jest to zatem typowe zadanie klasyfikacji, w którym obiekty opisane są przy pomocy zestawu atrybutów/cech. W rozdziale 7 przedstawiono metodykę oraz rezultaty rozpoznawania wybranych elementów mimiki, wykorzystując dane będące wynikiem etapu wyodrębniania cech. Na podstawie analizy wyników wybrano najlepszą metodę rozpoznawania oraz zaproponowano dalsze kierunki prac.

- **Rozdział 8 — Adaptacja systemu rozpoznawania mimiki.** Duża zmienność wyglądu twarzy oraz elementów mimiki, wynikająca z wpływu różnych czynników stanowi istotne utrudnienie w automatycznym rozpoznawaniu gestów. W rozdziale 8 usystematyzowano czynniki wpływające na skuteczność rozpoznawania oraz zaproponowano metody adaptacji systemu rozpoznawania mimiki do człowieka oraz zmieniających się warunków otoczenia.

# Część II

# Elementy mimiki

# Rozdział 2

# Sposoby pomiaru oraz opisu ekspresji mimicznych twarzy

**Streszczenie**

Wykorzystanie elementów mimiki do sterowania, wymaga uwzględnienia szeregu zagadnień, wśród których szczególnie istotne są: psychofizjologiczne uwarunkowania człowieka oraz sposób reprezentacji i opisu mimiki. W niniejszym rozdziale przedstawione zostały zagadnienia związane z komunikacją niewerbalną (istotne w kontekście wykorzystania mimiki do sterowania) oraz istniejące sposoby opisu ekspresji mimicznych twarzy. Spośród metod reprezentacji mimiki szczegółowo opisane zostały dwie z nich, wykorzystane w dalszej części pracy.

## 2.1 Mimika twarzy

Twarz człowieka może być traktowana jako bardzo elastyczny system zdolny do generacji wielu sygnałów informacyjnych oraz komunikatów jednocześnie. Można wyróżnić cztery ogólne klasy sygnałów:

- sygnały statyczne, na które składają się stałe cechy takie jak: naturalna budowa twarzy, zmarszczki mimiczne i fałdy skórne,
- sygnały wolnozmienne reprezentujące zmiany wyglądu twarzy zachodzące stopniowo w czasie (np. starzenie się),
- sztuczne elementy umieszczane na twarzy takie jak okulary, makijaż,
- szybkie sygnały czyli gesty mimiczne twarzy wywołane ruchami odpowiednich mięśni.

Spośród powyższych klas sygnałów głównie gesty mimiczne twarzy niosą informacje użyteczne z punktu widzenia ich wykorzystania do sterowania. Dlatego konieczne jest zdefiniowanie takiego sposobu opisu, który pozwoli na usystematyzowanie gestów mimicznych oraz nada im określone znaczenie. Pod uwagę należy również wziąć psychofizjologiczne uwarunkowania człowieka — podczas absorbującej pracy z aplikacją systemu komputerowego użytkownik może w odruchowy sposób zmieniać mimikę twarzy w odpowiedzi na bodźce somatyczne lub emocjonalne (np. mruganie oczami, uśmiech, ziewanie, marszczenie brwi). Z punktu widzenia niesionych informacji istnieje następujący podział funkcjonalny [22][23] związany z komunikacją niewerbalną (mimika):

- Wskaźniki uczuć — przekazują odbiorcy aktualny stan emocji nadawcy (radość, smutek, zaskoczenie, gniew, wstręt, strach), wyrażane głównie twarzą, parajęzykiem, tonem głosu i pozycją ciała.
- Emblematy (inaczej znaki autonomiczne) — zastępują określone słowa i frazy, wyrażają konkretne emocje i postawy (przykładem emblematu jest wzruszenie ramion jako odpowiednik wypowiedzi „nie mam pojęcia"). Przyjmuje się, że emblemat to taki sygnał, który jest identycznie interpretowany przez co najmniej 70 % populacji, w której się go używa.
- Ilustratory — służą do uzupełnienia treści wypowiadanych słów. Najczęściej wykonuje się je rękami, np. w celu zasygnalizowania wielkości, odległości i kierunku, ale używa się tu również mimiki i kanału wokalnego, aby np. podkreślić wagę pewnego słowa.
- Regulatory — organizują całość sytuacji komunikacyjnej, a więc takie sygnały, jak: tempo mówienia, przerwy, intonacja pytająca lub kończąca frazę, wzrokowe sygnały zamiaru przejęcia inicjatywy w dialogu lub odwrotnie — chęci przekazania głosu któremuś z pozostałych uczestników konwersacji i tym podobne.
- Manipulatory — nazywane też adaptatorami zapewniają większy komfort danej osobie, np. zmiana pozycji na krześle, oparcie się o ścianę, założenie nogi na nogę, poprawianie włosów. Należą tu także ruchy, takie jak: pocieranie jednej części ciała o drugą, drapanie, czyszczenie, przestawianie przedmiotów na stole i tym podobne.

Odróżnianie mimiki spontanicznej od celowych, znaczących ruchów jest zadaniem spoczywającym na oprogramowaniu interfejsu.

Spośród systemów opisu mimiki można wyróżnić MAX (ang. Maximally Discriminative Affect Coding System) [40]. Jest to teoretyczna metoda pomiaru określająca mimikę w kategoriach zmian wyglądu twarzy wywołanych poszczególnymi emocjami. Istnieje również wiele metod pozwalających na precyzyjny

opis ruchów mimicznych twarzy bez nadawania im konkretnego znaczenia. Do najczęściej używanych zaliczyć można systemy:

- FACS (ang. Facial Action Coding System) — oparty na anatomii system pomiaru ruchów mimicznych twarzy, grupujący je w jednostki czynnościowe (ang. action units) ściśle związane z anatomią muskulatury twarzy,
- FAP (ang. Facial Animation Parameters) — zdefiniowana w ramach standardu MPEG-4 specyfikacja oraz metoda opisu i animacji twarzy i ciała ludzkiego.

Poniżej zostały omówione dwie z wymienionych metodyk opisu mimiki — FACS oraz FAP — użyteczne z punktu widzenia wyboru gestów mimicznych przydatnych do sterowania.

W tabelach B.1, B.2, B.3 zawarto krótki opis jednostek czynnościowych wykorzystywanych w dalszej części pracy. Natomiast tabele B.4 do B.13 zawierają bardziej szczegółowe informacje o elementach mimiki. Tabele jak również zdjęcia wybranych elementów mimiki zamieszczono w dodatku B.

## 2.2 Facial Action Coding System

Metodyka FACS (ang. Facial Action Coding System) jest jednym z częściej wykorzystywanych sposobów pomiaru oraz opisu ekspresji mimicznych twarzy. Została ona opracowana przez psychologów P. Ekmana oraz W.V. Friesen w latach 1970, w oparciu o analizę zmian wyglądu twarzy wywołanych przez ruchy poszczególnych mięśni. Naukowcy ci przebadali wiele nagranych materiałów filmowych obrazujących różne ekspresje mimiczne. Na podstawie wiedzy o budowie anatomicznej twarzy oraz badań palpacyjnych (czyli badania przez dotyk), analizowane były specyficzne zmiany wyglądu wywołane skurczem poszczególnych mięśni lub ich grup oraz próbowano określić sposoby odróżniania poszczególnych ruchów mimicznych od siebie. Celem prowadzonych badań było opracowanie wiarygodnego sposobu rozróżniania i kategoryzowania ekspresji mimicznych twarzy przez wykwalifikowanych obserwatorów. Wynikiem prowadzonych prac jest metodyka FACS, która została opisana w podręczniku [24] zawierającym również bogaty materiał filmowy.

W metodyce tej wykorzystuje się pojęcie „jednostki czynnościowej" (ang. Action Unit, AU). Jednostka czynnościowa rozumiana jest jako izolowany i niepodzielny ruch mimiczny możliwy do świadomego wykonania przez człowieka oraz wywołujący obserwowalne zmiany w wyglądzie twarzy. Zmiany geometrii oraz wyglądu szczegółów anatomicznych twarzy wywołane są głównie przemieszczaniem obszarów skóry pociąganych przez mięśnie mimiczne. Wykonanie ruchu mimicznego może prowadzić na przykład do powstania bruzdy, wygładzenia zmarszczek

bądź uwypuklenia fragmentu skóry. Klasyfikacja FACS opisuje 44 jednostki AU. Większość z nich (trzydzieści) związana jest z ruchami mięśni, pozostałe natomiast — z ruchem głowy lub oczu. Dwanaście AU dotyczy górnej części twarzy, osiemnaście dolnej. Podział mimiki na jednostki czynnościowe nie odpowiada dokładnie anatomii muskulatury twarzy, ale jest z nią ściśle związany. Wynika to z dwóch powodów — po pierwsze, w niektórych przypadkach skurcze różnych mięśni wywołują podobne wizualnie zmiany w wyglądzie twarzy. Po drugie, skurcz poszczególnych części jednego mięśnia często wywołuje różne zmiany, które mogą być zaklasyfikowane jako odmienne jednostki czynnościowe. Ilość podstawowych jednostek czynnościowych jest relatywnie niewielka, w rzeczywistości występują one jednak częściej w kombinacjach. Zaobserwowana całkowita liczba kombinacji wynosi ponad 7000 [73]. Kombinacje mogą być addytywne, gdy połączenie nie wpływa na wygląd poszczególnych jednostek, oraz nie nieaddytywne — wygląd jednostek składowych zmienia się.

Klasyfikacja FACS opisuje zmiany wyglądu twarzy wywołane jednostkami czynnościowymi zarówno w aspekcie dynamicznym, jak i statycznym. Pozwala to na analizę ruchów mimicznych nie tylko obserwowanych w czasie rzeczywistym, ale także zarejestrowanych na nieruchomym obrazie twarzy.

W opisie FACS można wyróżnić następujące elementy:

- określenie jednostek czynnościowych AU składających się na dany ruch mimiczny,
- pomiar stopnia intensywności oraz czasu wykonywanych ruchów w pięciostopniowej skali — od A (śladowy gest mimiczny) do E (maksymalny gest),
- podział na ruchy jednostronne oraz symetryczne,
- uwzględnienie zależności z innymi AU, m.in. ruchami głowy oraz oczu,
- instrukcje w jaki sposób wykonywać ruch.

FACS jest klasyfikacją ściśle opisową — nie zawiera informacji o znaczeniu danego gestu mimicznego. Przyporządkowanie jednostkom AU emocji zostało opracowane w ramach innego systemu nazwanego FACSAID (ang. Facial Action Coding System Affect Interpretation Dictionary), który został opracowany przez tą samą grupę badawczą.

Wśród pozycji literaturowych, poświęconych automatycznemu rozpoznawaniu jednostek czynnościowych przy użyciu metod analizy obrazów, wymienić można prace prowadzone na uniwersytecie CMU (Carnegie Mellon University) [86][87][88]. W ramach projektu badawczego powstał system rozpoznawania mimiki AFA (ang. Automated Facial Image Analysis). Pozwala on na automatyczne rozróżnianie subtelnych zmian mimiki w oparciu o klasyfikację FACS dla twarzy widocznych frontalnie. Algorytm rozpoznawania wykorzystuje metody śledzenia

wybranych cech oraz sieci neuronowe. Śledzone części twarzy mogą stanowić stałe cechy charakterystyczne (takie jak np. oczy, brwi, usta) oraz cechy pojawiające się przejściowo (np. zmarszczki mimiczne). Stałe cechy reprezentowane są przez szczegółowe parametryczne modele (nazywane: ang. multistate facial component models), uwzględniające różne stany w jakich może znaleźć się śledzony element (np. otwarte-zamknięte oczy). System rozwijany jest nie tylko w kierunku automatycznego rozpoznawania mimiki, ale również rozpoznawania i interpretacji zachowania ludzi. Posiada on (wg autorów) średnią skuteczność rozpoznawania 96.4% dla AU górnej części twarzy oraz 96.7% AU dolnej części twarzy. Testy przeprowadzane były na dwóch bazach danych: CMU Facial Expression Database [44] oraz Ekman-Hager Facial Action Exemplars [25].

Prace w kierunku rozpoznawania jednostek czynnościowych prowadzone są również w innych ośrodkach naukowych. Wśród nich wyróżnić należy system opracowany przez Valstara [92]. Zaproponowana metoda wykorzystuje czasowe szablony (ang. temporal templates) oraz klasyfikację opartą na algorytmie kNN (ang. k-Nearest-Neighbor) połączonym z algorytmem regułowym (ang. rule-based). System pozwala na rozpoznawanie 15 jednostek AU występujących pojedynczo lub w kombinacjach i charakteryzuje się średnią skutecznością rozpoznawania 76.2% (testy przeprowadzane na bazie CMU). Autorzy prowadzą również prace nad rozpoznawaniem emocji opartym o jednostki AU [64].

Interesujące podejście do problemu rozpoznawania mimiki zaprezentował A. Kapor w pracy [45]. Przedstawione rozwiązanie pozwala na automatyczne rozpoznawanie jednostek czynnościowych oraz ich kombinacji w obrębie górnej części twarzy. System wykorzystuje dodatkowe urządzenie — aktywny oświetlacz podczerwieni. Zastosowanie oświetlacza daje możliwość skutecznej detekcji oczu człowieka poprzez obserwację odbicia światła od siatkówki oka. Siatkówka oka ludzkiego działa jak lustro sferyczne odbijając od dna oka promieniowanie, które wpada poprzez obszar źrenicy. Zjawisko to jest powszechnie znane jako efekt „jasnych źrenic" (ang. bright pupil effect) i występuje gdy źródło światła i system optyczny z sensorem zostaną umieszczone współosiowo. W przypadku oświetlenia bocznego (niewspółosiowego), obszar źrenicy na obrazie uzyskanym z kamery jest znacznie ciemniejszy od otoczenia — efekt „ciemnych źrenic" (ang. dark pupil effect). Wykorzystanie oświetlacza w dużym stopniu zwiększa skuteczność oraz automatyzuje detekcję oczu. Informacja o położeniu oczu jest następnie wykorzystywana do pobrania odpowiednich obszarów obrazu zawierających elementy mimiki (oko, powieka, brwi). Dalsza analiza wykorzystująca statystyczny algorytm PCA oraz liniową kombinację wzorców, pozwala na otrzymanie parametrów kształtu. Kształt ten jest następnie rozpoznawany poprzez klasyfikator z wykorzystaniem metody SVM (ang. support vector machines). Przedstawione rozwiązanie charakteryzuje się skutecznością rozpoznawania około 69% (testy na własnej bazie autorów) oraz 81.22% — testy dla bazy CMU. Jak wynika z analizy

autorów, metoda ta nie radzi sobie z dużymi zmianami skali obserwowanej twarzy oraz obrazami nowych twarzy o odmiennym wyglądzie.

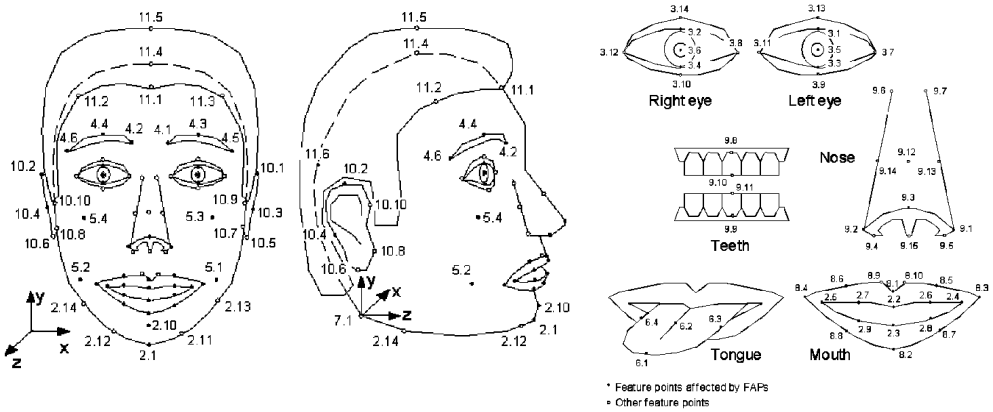
Istotnym problemem w automatycznym rozpoznawaniu jednostek czynnościowych jest fakt, iż gesty mimiczne wykonywane przez ludzi podczas interakcji (np. rozmowa) z innymi osobami bardzo często połączone są z dużymi ruchami głowy (np. potrząsanie, przechylanie głowy, itp). Ruchy te przekładają się na rotacje poza płaszczyzną obrazu (ang. out-of-plane rotations), będącego dwuwymiarową reprezentacją trójwymiarowej sceny. Skutkuje to istotnymi zniekształceniami kształtu twarzy oraz wyglądu jej elementów. Problem ten jest poruszany w pracy [4]. Autorzy uwzględnili zniekształcenia obrazu wynikające z ruchów głowy w następujący sposób. W pierwszej kolejności estymowane są parametry modelu kamery (ang. camera calibration parameters) oraz wyznaczane jest geometria twarzy na scenie trójwymiarowej (3D). Następnie informacje te wykorzystywane są do utworzenia modelu twarzy niezależnego od parametrów sceny (skala, rotacja). Na podstawie takiego modelu generowany jest obraz twarzy w pozycji frontalnej względem kamery (ang. image warping). Mając do dyspozycji frontalny obraz twarzy, wyodrębniane są jej cechy charakterystycznych przy użyciu filtrów Gabora (ang. Gabor wavelets). Zestaw takich cech jest następnie klasyfikowany przy pomocy SVM oraz ukrytych modeli Markowa (ang. Hidden Markov Models). Zaproponowane rozwiązanie testowany było na własnej bazie autorów — osiągnięta skuteczność wynosiła 95.9%.

## 2.3 MPEG-4 FAP

W ramach standardu MPEG-4 [62][53] zdefiniowana została specyfikacja metody opisu i animacji twarzy oraz ciała ludzkiego. Definiuje ona model geometryczny twarzy (siatka wieloboków lub model powierzchni NURBS), określa parametry animacji i sposób ich kodowania oraz zawiera reguły pozwalające na obliczanie deformacji modelu geometrycznego na podstawie parametrów animacji.

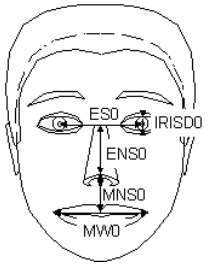
W odróżnieniu od metodyki FACS, ściśle związanej ze szczegółami anatomicznymi (mięśnie mimiczne, zmarszczki i fałdy skórne), standard ten określa 84 punkty charakterystyczne pozwalające na opisanie wyglądu twarzy. Zostały one podzielone na kilka grup w zależności od ich położenia oraz przynależności do określonych elementów twarzy (np. oczy, brwi, policzki, język, itp.) — rysunek 2.1.

Niektóre z punktów związane są z parametrami animacji FAP (ang. Facial Animation Parameters), które pozwalają na opisanie ruchu pojedynczych punktów, bądź też zapisanie całych gestów mimicznych (przesunięcie całej grupy). Ruch definiowany jest względem stanu neutralnego, tzw. pojęcia „neutralnej twarzy", określającego nie tylko wygląd elementów twarzy (np. usta zamknięte, wargi złączone), ale również położenie głowy względem kamery oraz układ odniesienia.




Rysunek 2.1: Punkty charakterystyczne twarzy definiowane w standardzie MPEG-4 FAP, [62]

Amplituda ruchu mierzona jest względem trzech osi współrzędnych w jednostkach określanych przez tablicę FAPU (ang. Facial Animation Parameter Unit). FAPU (rys. 2.2) pozwala na pomiar ruchu w jednostkach względnych — przykładowo jednostka ES0 związana jest z odległością między źrenicami. Oprócz gestów mimicznych FAP pozwalają również na opisanie ruchów głowy oraz języka.




Rysunek 2.2: Standard MPEG-4 — jednostki FAPU, [62]

Zdefiniowana specyfikacja pozwala na generację realistycznych animacji, poprzez odpowiednie deformacje modelu geometrycznego twarzy w zależności od zmian parametrów FAP. Animacje takie wykorzystywane są nie tylko do tworzenia realistycznych „avatarów" (wirtualna postać reprezentująca człowieka), ale również pozwalają na realizację wideokonferencji nie wymagającej dużych przepustowości łączy — przesyłane są tylko parametry FAP oraz informacje o teksturze. Istotnym elementem w tego typu zastosowaniach jest automatyczna lokalizacja oraz śledzenie punktów charakterystycznych twarzy na obrazach otrzymywanych

z kamery. Stosowane są rozwiązania oparte na umieszczaniu łatwo odróżnialnych markerów w odpowiednich miejscach twarzy, a następnie wykorzystanie metod przetwarzania i analizy obrazów do ich lokalizacji. Przykładem może być system animacji [51] umożliwiający rozpoznawanie wybranych gestów mimicznych, a następnie animowanie sztucznie generowanego avatara. System wykorzystuje sztuczne markery umieszczane w określonych punktach charakterystycznych twarzy. W artykule [30] autorzy zaprezentowali podejście nie wymagające stosowania markerów. Punkty charakterystyczne określane są ręcznie przez operatora podczas inicjalizacji systemu — pobierane są informacje o kolorze oraz krawędziach w otoczeniu punktu. Następnie elementy twarzy są śledzone na kolejnych obrazach sekwencji video. Do śledzenia używane są algorytmy wykorzystujące budowę morfologiczną poszczególnych elementów twarzy — np. dla oka są to: położenie źrenicy, powiek oraz brwi.

# Rozdział 3

# Określenie i wybór rozpoznawanych elementów mimiki

**Streszczenie**

Skonstruowanie urządzenia wejściowego umożliwiającego sterowanie przy pomocy gestów mimicznych, wymaga uwzględnienia wymagań wynikających z typowych zadań jakie wykonuje użytkownik podczas pracy z komputerem. Na potrzeby niniejszej rozprawy dokonano usystematyzowania typowych zadań wykonywanych przez człowieka podczas interakcji z interfejsem graficznym komputera. Poprzez analizę akcji użytkownika oraz porównanie własności typowych urządzeń wejściowych z możliwościami jakie oferuje mimika, dokonano wyboru elementów mimiki, które są potencjalnie przydatne w komunikacji człowiek-maszyna.

## 3.1 Systematyka typowych zadań wykonywanych przez użytkowników

Sterowanie komputerem przy pomocy gestów mimicznych, wymaga od projektanta wzięcia pod uwagę szeregu czynników. Oprócz specyfikacji technicznej (dobór kamery, sposobu transmisji danych...) i informatycznej (określenie w jaki sposób sygnał wizyjny będzie interpretowany przez komputer), konieczne jest uwzględnienie wymagań wynikających z typowych zadań jakie wykonuje użytkownik podczas pracy z interfejsem graficznym.

Na wczesnym etapie rozwoju informatyki, wymagania te zostały określone w postaci standardów: GSPC, 1977; GSPC, 1979; ISO, 1983. Wprowadzają one koncepcje „logicznych urządzeń wejściowych" (ang. logical devices), stanowiących

poziom pośredni między sterownikami sprzętu a aplikacjami. Programista aplikacji dysponuje standardowym API (ang. Application Programming Interface), udostępniającym funkcje pozwalające m.in. na: określenie pozycji (np. kursora we współrzędnych ekranu), wybór (np. elementu graficznego), wprowadzanie tekstu, liczb. Z powyższymi sposobami wprowadzania informacji do komputera powiązane są konkretne typy urządzeń (np. klawiatura, myszka).

W pracy [27] autorzy zaprezentowali podejście, które systematyzuje urządzenia wejściowe z punktu widzenia udostępnianej przez nie funkcjonalności. Wyszczególnili oni sześć elementarnych zadań pozwalających na realizację dowolnego celu sterowania odpowiadającego intencjom użytkownika:

- wybór obiektu lub elementu graficznego spośród kilku możliwości (ang. selection),
- określenie pozycji 1,2 lub 3 wymiarach (ang. position),
- określenie orientacji w 1,2 lub 3 wymiarach (ang. orientation),
- podanie ścieżki, rysowanie (ang. path or ink), co wymaga podania kolejnych pozycji lub orientacji w czasie,
- wprowadzanie tekstu (ang. text entry),
- ustalenie wartości numerycznej (ang. quantify).

Zadania te odpowiadają w przybliżeniu podejściu zawartemu w standardzie GSPC.

Powyższy podział nie jest do końca jednoznaczny, ponieważ zależy od rodzaju wykorzystanego urządzenia [35]. Jako przykład można przedstawić zadanie wyboru elementu graficznego. Wykorzystanie myszki wymaga w pierwszej kolejności określenia pozycji, a następnie dokonania wyboru (kliknięcie). Aby osiągnąć cel potrzebne są dwa elementarne zadania. Natomiast przy pomocy ekranu dotykowego lub tabletu graficznego wystarczające jest jedno zadanie — bezpośrednie wskazanie żądanego obiektu. Ominięcie problemu jednoznaczności można osiągnąć poprzez zdefiniowanie pojęcia „zadań złożonych" (ang. compound tasks), czyli akcji które z punktu widzenia użytkownika stanowią jedno zadanie (lub mogą być wykonane hierarchicznie). Przykładem może być wybór odnośnika na stronie WWW, który wymaga przewijania dokumentu oraz wskazania kursorem oraz wybrania odnośnika.

Wymienione sposoby opisu funkcjonalności urządzeń wejściowych prezentują dwa odmienne podejścia, które można zaobserwować w konstruowaniu interfejsów komputerowych [55]. Koncepcja „logicznych urządzeń wejściowych" jest przykładem architektury skoncentrowanej na maszynie (ang. machine-centered architecture). Z kolei w architekturach skupionych na użytkowniku (ang. human-centered

interaction architecture), urządzenia wejściowe rozpatruje się nie od strony interfejsu programowego (API), ale z punktu widzenia udostępnianej funkcjonalności. W takim podejściu istotny staje się odpowiedni dobór urządzenia wejściowego do wymaganego zadania.

## 3.2 Własności i systematyka urządzeń wejściowych

Użytecznymi kryteriami w doborze urządzenia do określonego zadania, mogą być własności danego urządzenia. Do najistotniejszych własności można zaliczyć [36]:

- **Mierzona fizyczna cecha.** Typowy sprzęt wykorzystuje pomiar pozycji (np. tablet), ruchu (myszka), siły (dżojstik izometryczny[1]) lub kąta (dżojstik izotoniczny[2]). Urządzenia mierzące pozycję pozwalają na jej określenie w sposób bezwzględny. Umożliwiają również emulację względnej pozycji. Z kolei urządzenia oparte na pomiarze ruchu są zdolne tylko do podawania względnego położenia poprzez sterowanie kursorem na ekranie. Zazwyczaj wymagana jest również odpowiednia funkcja przejścia (ang. transfer function) zapewniająca odpowiednie przekodowanie siły na położenie. Może one przyjąć postać funkcji akceleracji pozwalającej na realizację mniejszego przesunięcia na początku ruchu, a większego przy końcu. Definiuje się również pojęcie stosunku sterowania do wyświetlania (ang. control to display ratio), czyli współczynnika definiującego przesunięcie kursora na ekranie w stosunku do ruchu np. myszki.

- **Ilość stopni swobody.** Akcje wykonywane przez użytkownika podczas obsługi interfejsu graficznego wymagają określenia położenia w jednym (suwak), dwóch (położenie na ekranie), lub trzech wymiarach oraz określanie orientacji (zastosowania CAD, gry). Urządzenia które nie zapewniają żądanych informacji wymagają dodatkowych sposobów interakcji (np. podział zadania na kilka wykonywanych sekwencyjnie).

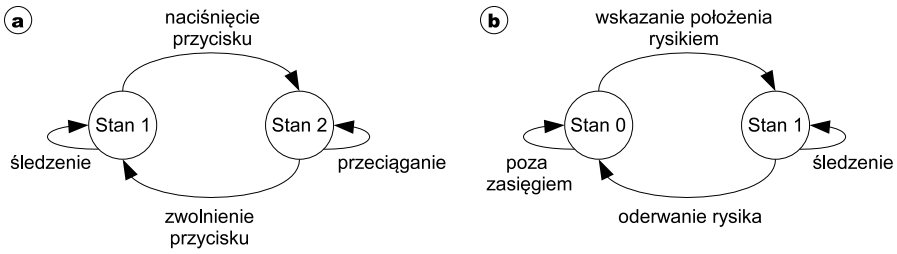
- **Szybkość i dokładność.** Pod tymi pojęciami rozumie się potocznie szereg następujących parametrów takich jak: rozdzielczość, częstotliwość próbkowania, opóźnienie reakcji, szum, nieliniowości.

- **Pośrednie lub bezpośrednie sterowanie.** Przykładem bezpośredniego sterowania jest ekran dotykowy, w którym powierzchnia wyświetlająca informacje służy równocześnie do wprowadzania informacji. Pozostałe urządzenia udostępniają sterowanie pośrednie, np. myszka.

[1] sygnał wyjściowy jest funkcją siły przyłożonej do drążka
[2] sygnał wyjściowy jest funkcją wychylenia drążka z pozycji neutralnej

- **Ciągły pomiar wielkości lub wykrywanie zdarzeń.** Wszelkiego rodzaju przyciski (ang. buttons) zapewniają sygnały w postaci zdarzeń (naciśnięty/zwolniony...), w odróżnieniu od np. dżojstika umożliwiającego ciągły pomiar siły.
- **Obsługiwane stany.** Zadania takie jak np. zaznaczenie ikony na pulpicie, operacje „przeciągnij i upuść", wymagają dodatkowej sygnalizacji stanu odpowiadającego intencji użytkownika. W pracy [13] określone zostały trzy podstawowe stany wymagane dla poprawnej obsługi interfejsu graficznego.
  - stan 0: poza zasięgiem (ang. out of range) — przykładowo oderwanie rysika od tabletu graficznego,
  - stan 1: śledzenie (ang. tracking) — ruch kursora,
  - stan 2: przeciąganie (ang. dragging) — pozwala na przesuwanie elementów interfejsu np. ikon.

Nie wszystkie urządzenia wspierają obsługę powyższych stanów. Jako przykład podać można pracę z myszką oraz z panelem dotykowym — 3.1. W przypadku panelu odróżnienie operacji przeciągania od śledzenia wymaga zastosowania dodatkowego przycisku. Z kolei myszka nie obsługuje stanu „poza zasięgiem". Ma to znaczenie w niektórych typach zadań.




Rysunek 3.1: Stany obsługiwane przez: (a) myszkę, (b) panel dotykowy.

## 3.3 Analiza możliwości jakie oferuje mimika

Przedstawiona systematyka własności urządzeń wejściowych oraz analiza typowych zadań stanowi dobry punkt wyjścia do wyboru potencjalnych gestów mimicznych, które zapewnią sterowanie komputerem. Na potrzeby niniejszej pracy dokonano usystematyzowania i wyboru typowych zadań wykonywanych przez człowieka podczas interakcji z interfejsem graficznym. Są to:

- wskazanie i wybór elementu graficznego (ang. point and select),

- operacja „przeciągnij i upuść” (ang. drag and drop),
- operacja „rozciągania gumki” (ang. rubber banding) służąca np. do rysowania ramki zaznaczania wielu obiektów,
- rysowanie (ang. ink),
- menu rozwijalne i wyskakujące (ang. pull-down menu, pop-up menu),
- wprowadzanie tekstu (ang. text entry),
- rozpoznawanie gestu (ang. gesture recognition) wykonywanego np. myszą,
- operacje złożone (ang. compound tasks), np. manipulacja obiektem 3D.

Analizując przedstawione zadania (tab. 3.1) można zauważyć, iż urządzenie wejściowe oraz jego interfejs programowy powinno:

- udostępniać sygnały sterujące takie jak: pozycja, orientacja, trajektoria, wybór,
- oraz obsługiwać następujące stany: śledzenie i przeciąganie (w niektórych przypadkach również stan „poza zasięgiem”).

Udostępnianie sygnałów sterujących jest realizowane w odmienny sposób przez różne urządzenia (myszka, tablet), które charakteryzują się odmiennymi własnościami.

Bezpośrednie wskazanie pozycji na ekranie (ang. direct locator) możliwe jest zazwyczaj w tabletach graficznych. Z kolei myszka komputerowa pozwala na pośrednie wskazanie pozycji poprzez przesunięcie kursora w zadane miejsce (ang. indirect locator). Często wykorzystuje się również wskazanie pozycji przy pomocy dedykowanych klawiszy (ang. location direction keys). W przypadku gestów mimicznych sterowanie odbywa się pośrednio, ponieważ powierzchnia wyświetlająca nie może być zintegrowana z urządzeniem. Nie wydaje się możliwe również bezwzględne mapowanie niewielkich zmian mimiki na położenie kursora, stąd taki sposób sterowania można zaklasyfikować jako względny. Wyjątkiem może być ograniczona możliwość bezwzględnego mapowania ruchów głowy na orientację.

Różne zadania wymagają od urządzenia obsługi innej ilości stopni swobody. Przykładowo do przesuwania suwaka wystarczający jest jeden stopień swobody. Dlatego w konstrukcji myszek komputerowych dodaje się rolkę sterującą. Z kolei wybór obiektu na ekranie wymaga zapewnienia w urządzeniu dwóch stopni swobody (wskazanie pozycji). W przypadku mimiki zapewnienie jednego stopnia swobody nie stanowi problemu — prawie każdy gest może zostać do tego celu wykorzystany — ruch brwi w górę (jednostka czynnościowa AU 1+2) oraz w dół (AU4). W przypadku dwóch stopni swobody konieczny jest wybór przynajmniej

Tabela 3.1: Zestawienie typowych zadań oraz wymaganych dla nich sygnałów i stanów

| **Zadanie** | **Wymagane sygnały sterujące** | **Wymagana obsługa stanów** |
|---|---|---|
| wskazanie elementu graficznego | pozycja w 1, 2 lub 3 wymiarach | 1 (opisy stanów na rysunku 3.1 ) |
| wskazanie i wybór elementu graficznego | pozycja w 1, 2 lub 3 wymiarach oraz sygnał wyboru | 1, 2 |
| operacja „przeciągnij i upuść” | pozycja w 1, 2 lub 3 wymiarach oraz sygnał wyboru | 1, 2 |
| operacja „rozciągania gumki” | pozycja w 1, 2 lub 3 wymiarach oraz sygnał wyboru | 1, 2 |
| rysowanie | trajektoria w 1, 2 lub 3 wymiarach oraz sygnał wyboru | 1, 2 |
| menu rozwijalne i wyskakujące | pozycja lub trajektoria w 1, 2 lub 3 wymiarach oraz sygnał wyboru | 1, 2 |
| wprowadzanie tekstu (klawiatura) | wymagana klawiatura lub w przypadku klawiatury programowej — pozycja w 2 wymiarach oraz sygnał wyboru (dla jednowymiarowej klawiatury programowej – sygnał pozycji 1D) | W przypadku klawiatury programowej — 1, 2 |
| wprowadzanie tekstu (rozpoznawanie pisma) | trajektoria w 1, 2 lub 3 wymiarach oraz sygnał wyboru | 0 (ze względu na konieczność sygnalizacji końca wpisywania tekstu), oraz 1 i 2 |
| rozpoznawanie gestu | trajektoria w 1, 2 lub 3 wymiarach | 0, 1 |
| operacje złożone — np. rotacja w programie CAD | orientacja w 1, 2 lub 3 wymiarach | 1, 2 |

czterech gestów. Możliwe jest również uzyskanie trzech stopni swobody, poprzez np. pomiar ruchów głowy w różnych płaszczyznach (góra/dół, przechylenie prawo/lewo, obrót prawo/lewo).

W przypadku mimiki możliwy jest pomiar następujących parametrów: wykonanie lub brak wykonania gestu, intensywność gestu, oraz szybkość wykonania gestu. Wszystkie zmiany odbywają się względem mimiki neutralnej, a ich am-

plituda jest raczej niewielka. Sugeruje to wykorzystanie mimiki do sterowania podobnie jak w przypadku joysticka. Istnieje możliwość wyboru pomiaru intensywności lub szybkości wykonania gestu (pomiar wielkości fizycznej ciągłej — dżojstik analogowy) lub faktu wykonania lub braku wystąpienia gestu (detekcja zdarzenia — dżojstik cyfrowy).

Zadanie wyboru obiektu lub elementu graficznego również może być realizowane na kilka sposobów. Bezpośredni wybór (ang. direct pick) udostępniają urządzenia takie jak tablety graficzne lub ekrany dotykowe. Pośrednie wybranie elementu graficznego realizowane jest zazwyczaj na kilka różnych sposobów:

- poprzez wskazanie pozycji oraz akcję wyboru (ang. indirect locator),
- pośrednie wybranie elementu graficznego poprzez wskazanie pozycji oraz oczekiwanie określonego czasu (ang. time scan),
- wybranie jednego ze zdefiniowanych wcześniej elementów poprzez naciśnięcie dedykowanego przycisku.

Możliwe jest również wybranie elementu poprzez rozpoznanie rysowanego gestu (przykład — sterowanie przeglądarką Opera przy pomocy gestów wykonywanych kursorem). Wymienione sposoby wymagają od urządzenia bądź udostępnienia bezpośredniego wyboru (tablet, dedykowane klawisze), lub obsługi stanów śledzenia, przeciągania (czasami również stanu „poza zasięgiem"). W kontekście sterowania mimiką możliwy jest wybór obiektu poprzez wcześniejsze wskazanie pozycji oraz akcję wyboru (analogicznie jak dla myszki). Wymaga to jednak zdefiniowania gestów zapewniających sygnały sterujące pozycją oraz dodatkowego gestu dla akcji wyboru. Możliwe jest również pominięcie dodatkowe gestu wyboru i emulacja poprzez oczekiwanie określonego czasu (ang. time scan). Potencjalną zaletą mimiki jest duża różnorodność gestów mimicznych (np. ilość kombinacji jednostek czynnościowych = 4000), która pozwala na wybranie jednego z wielu zdefiniowanych wcześniej elementów poprzez przypisanie gestom akcji naciśnięcia dedykowanych przycisków.

Uzyskanie sygnału trajektorii, w przypadku mimiki, jest również możliwe. Najprostszym sposobem może być zdefiniowanie sekwencji gestów (góra, dół, prawo, lewo) definiujących dane pozycji w czasie i składających się na trajektorię. Możliwe wydaje się również rozpoznawanie ruchów głowy lub ruchów gałek ocznych i kierunku patrzenia (ang. eye gaze direction).

Wszystkie wykorzystane gesty powinny być łatwe do wykonania dla człowieka oraz w miarę możliwości intuicyjne. Wymaga to uwzględnienia umiejętności oraz psychofizjologicznych uwarunkowań człowieka — nie każda osoba potrafi wykonać wszystkie gesty, występuje mimika mimowolna (emocje) oraz problem powtarzalności. Wiele zależy również od budowy i specyfikacji technicznej urządzenia (parametry toru przetwarzania i analizy — częstotliwość próbkowania obrazu

FPS, szumy, nieliniowości), która będzie miała wpływ na szybkość i dokładność pomiaru.

W tabelach 3.2 oraz 3.3 i 3.4, zawarte zostało podsumowanie własności typowych urządzeń wejściowych w kontekście systemu opartego o mimikę oraz podsumowano możliwości jakie oferuje mimika (sygnały sterujące, ilość potrzebnych gestów).

Tabela 3.2: Podsumowanie własności typowych urządzeń wejściowych w kontekście sterowania mimiką

| **Własność** | **Myszka / joystick** | **Tablet** | **Gesty mimiczne** |
|---|---|---|---|
| Mierzona cecha fizyczna | Myszka – ruch, dżojstik – siła lub kąt | Pozycja | Można mierzyć: wykonanie lub brak wykonania gestu, intensywność gestu, szybkość wykonania gestu. |
| Pomiar pozycji względny lub bezwzględny | Względny (użytkownik musi sterować kursorem) | Bezwzględny oraz możliwość emulacji względnego pomiaru pozycji | Ze względu na specyfikę wykonywanych gestów (zmiana w stosunku do mimiki neutralnej) możliwy jest pomiar względny (podobnie jak joystick). Wyjątkiem mogą być ruchy głowy, które w ograniczonym zakresie można wykorzystać do podania orientacji bezwzględnej. |
| Ilość stopni swobody | Dwa | Dwa | Jeden, dwa, lub trzy – w zależności od gestu (np. ruch głowy w pewnych granicach może być mapowane do pozycji 3D, natomiast ruch brwi – jako pozycja 1D). |
| Szybkość, rozdzielczość, dokładność | W zależności od urządzenia, z reguły duża | W zależności od urządzenia, z reguły duża | Raczej mała, ze względu na to iż amplituda większości gestów jest niewielka. Uwzględnić należy również psychofizjologiczne uwarunkowania człowieka — nie każda osoba potrafi wykonać wszystkie gesty, występuje mimika mimowolna (emocje) oraz problem powtarzalności. Szybkość i dokładność będzie również zależała od parametrów toru przetwarzania i analizy (np. szumy). |
| Ciągły pomiar wielkości lub wykrywanie zdarzeń | Ciągłe podawanie pozycji oraz zdarzenia – stany przycisków | Ciągłe podawanie pozycji oraz zdarzenia – stany przycisków. Dodatkowo pomiar siły i nacisku | Możliwy ciągły pomiar (intensywności lub szybkości wykonania gestu), lub detekcja zdarzeń (wykonanie lub brak gestu). |
| Obsługiwane stany | 1, 2 (rys. 3.1) | 0, 1 (2 z dodatkowym przyciskiem) | Stan 0 (poza zasięgiem) – możliwy<br>Stan 1 (śledzenie) – możliwy<br>Stan 2 (przeciąganie) – możliwy jeśli zostaną dobrane gesty, które mogą być wykonane niezależnie i jednocześnie (np. ruch ust i mruganie) |

Tabela 3.3: Podsumowanie możliwości oferowanych przez mimikę w kontekście zadań — cz.1

| **Zadanie** | **Sygnały sterujące** | **Uwagi** |
|---|---|---|
| Bezpośrednie wskazanie pozycji na ekranie | Uzyskanie sygnału z wymaganą rozdzielczością niemożliwe | — |
| Pośrednie wskazanie pozycji poprzez przesunięcie kursora | Wykonanie lub brak wykonania gestu lub intensywność / szybkość wykonania gestu | Wymagana ilość gestów: dla dla 1D – min. 1 gest, optymalnie 2 gesty, 2D co najmniej 4 gesty, 3D – 6 gestów |
| Bezpośrednie wybranie elementu graficznego | Bezpośrednie wskazanie pozycji oraz wyboru niemożliwe | — |
| Pośrednie wybranie elementu graficznego poprzez wskazanie pozycji oraz akcję wyboru | Wykonanie lub brak wykonania gestu, pomiar intensywności / szybkości wykonania gestu | Wymaga zdefiniowania gestów do określenia pozycji (np. 2D – 4 gesty) oraz dodatkowych gestów sygnalizujących wybór (kliknięcie, podwójne kliknięcie, prawy lewy klawisz) |
| Pośrednie wybranie elementu graficznego poprzez wskazanie pozycji oraz oczekiwanie określonego czasu | Wykonanie lub brak wykonania gestu lub intensywność / szybkość wykonania gestu | Podobnie jak dla pośredniego wskazania pozycji poprzez przesunięcie kursora. Dodatkowo obsługa emulacji czasowej akcji wyboru |

Tabela 3.4: Podsumowanie możliwości oferowanych przez mimikę w kontekście zadań — cz.2

| **Zadanie** | **Sygnały sterujące** | **Uwagi** |
|---|---|---|
| Wybranie jednego ze zdefiniowanych wcześniej elementów poprzez naciśnięcie dedykowanego przycisku | Wykonanie lub brak wykonania gestu | Ilość gestów odpowiadająca ilości dedykowanych przycisków. Gesty powinny być łatwo odróżnialne od siebie |
| Wybranie elementu poprzez rozpoznanie rysowanego gestu. | Wykonanie lub brak wykonania gestu lub intensywność / szybkość wykonania gestu | Wymaga zdefiniowania określonej sekwencji gestów mimicznych (kolejne dane pozycji) lub trajektorii ruchu głowy / oczu |
| Ustawienie bezpośrednie orientacji elementu (mapowanie orientacji rzeczywistego obiektu np. wirtualna rękawica) | Pomiar intensywność wykonania ruchu głową lub ruchu gałek ocznych | W ograniczonym zakresie mapowanie ruchów głowy / oczu na orientację |
| Ustawienie pośrednie orientacji elementu | Wykonanie lub brak wykonania gestu lub intensywność / szybkość wykonania gestu | Wymaga oprócz gestów do ustalenia pozycji, niezależnych gestów dla orientacji, lub ew. obsługi dodatkowego stanu sygnalizującego wyznaczanie pozycji lub orientacji |

## 3.4 Wybór elementów mimiki potencjalnie przydatnych do sterowania

Z przedstawionej analizy wynika, że podstawowym sygnałem wymaganym do realizacji większości zadań, jest informacja o pozycji oraz sygnał wyboru. Informacja o trajektorii lub orientacji może zostać podana w sposób pośredni, podobnie jak w typowych urządzeniach wejściowych (zapamiętanie zmian pozycji w czasie, zamiana pozycji na orientację). Operacje wymagające obsługi kilku stanów (śledzenie, przeciąganie) są również możliwe do obsłużenia, jednakże warunkiem jest zapewnienie dodatkowego sygnału informującego o aktualnym stanie lub wykorzystanie do tego celu sygnału wyboru (analogicznie jak przycisk w myszce). W niniejszym rozdziale dokonano wyboru gestów, które będą potencjalnie przydatne do sterowania. Przy selekcji wykorzystano systematykę FACS, która określa m.in. stopień trudności wykonania danego gestu mimicznego, oraz wzięto pod uwagę możliwość niezależnego wykonania wymaganej ilości gestów. Propozycje zostały podsumowane w tabelach 3.5, 3.6 oraz 3.7. Metodyka FAP również może zostać wykorzystana. Należy jednak mieć na uwadze fakt, że standard ten definiuje model geometryczny i jego animacje — w odróżnieniu od standardu FACS, który jest ściśle związany z ruchami mięśni.

Duża różnorodność obserwowanych i możliwych do wykonania przez człowieka gestów mimicznych pozwala na elastyczny dobór elementów mimiki i powiązanie ich z określonymi sygnałami sterującymi. Dlatego zaproponowany wybór nie jest jedynym możliwym — w zależności od umiejętności lub preferencji człowieka do tych samych celów mogą zostać użyte różne zestawy gestów. **Jest to szczególna zaleta w kontekście wykorzystania systemu przez osoby niepełnosprawne.** Kilka przykładów które obrazują możliwości sterowania przy pomocy mimiki — do emulacji myszki/dżojstika z dwoma przyciskami mogą zostać wykorzystane elementy mimiki nr 1 i 2 (tab. 3.5, tab. 3.6). Sterowanie pozycją zapewniane jest poprzez gesty mimiczne dolnej części twarzy (różne ruchy ust), natomiast udostępnianie sygnału wyboru dzięki gestom górnej części twarzy (ruch brwi). Jako alternatywne źródła sygnałów można zaproponować — zestawy nr 4 i 2, wykorzystujące ruchy głowy do sterowania kursorem oraz ruchy brwi do emulacji przycisków. Osoby które posiadają lepszą kontrolę mimiki górnej części twarzy zaproponować można zestaw nr 5, wykorzystujący gesty mrugania oraz ruchów brwiami.

Jak widać na powyższych przykładach wybór gestów do sterowania jest dość szeroki. Należy jednak pamiętać, iż mimika jest sprawą bardzo indywidualną i gest łatwy do wykonania dla jednego człowieka może być niewykonalny dla innego (panowanie nad muskulaturą twarzy, stopień niepełnosprawności). Do psychofizjologicznych uwarunkowań człowieka można również zaliczyć emocje, których skutkiem często jest odruchowa zmiana mimiki twarzy. **Dlatego interfejs**

**człowiek-komputer oparty o rozpoznawanie mimiki powinien raczej udostępniać użytkownikowi „alfabet" z którego może on skorzystać wybierając najbardziej dla siebie naturalne gesty, niż ustalone z góry sposoby sterowania.**

Użycie gestów mimicznych do emulacji funkcjonalności myszki lub dżojstika nie wykorzystuje pełnych możliwości jakie oferuje mimika. Duża ilość istniejących elementów mimiki pozwala na zaprojektowanie interfejsu bardziej elastycznego, który będzie udostępniał wiele równoczesnych sposobów komunikacji (interfejs multimodalny). Wykorzystanie redundancji gestów mimicznych może w znaczny sposób uprościć realizację bardziej złożonych zadań takich jak np. nawigacja po dokumentach. Wybrane zadania złożone oraz sposoby użycia elementów mimiki przedstawiono w tabeli 3.7.

Analizując możliwości jakie oferuje mimika do sterowania, wymienić należy również urządzenia przenośne — takie jak telefony komórkowe, palmtopy oraz komputery przenośne PDA. Zastosowania tych urządzeń są bardzo szerokie (kalendarz, terminarz, notatnik, przeglądanie multimediów, gry, współpraca z urządzeniami do nawigacji GPS, itp).

W przypadku palmtopów oraz PDA podstawowym sposobem interakcji wykorzystywanym przez użytkowników jest rysik pozwalający na wybór elementów, czy też rysowanie na ekranie dotykowym. Do wprowadzania informacji w telefonach komórkowych, najpopularniejsza jak dotąd, jest klawiatura udostępniająca wybrany zestaw cyfr i liter oraz klawisze funkcyjne.

Prowadzone są również badania i wdrożenia nad alternatywnymi możliwościami interakcji, takimi jak rozpoznawanie mowy. Innym przykładem jest urządzenie firmy Apple — iPhone. Udostępnia ono wielodotykowy ekran (ang. multitouch) pozwalający na bardziej naturalną dla człowieka pracę z urządzeniem. Ekran wielodotykowy, w odróżnieniu od typowego ekranu dotykowego, pozwala na jednoczesną detekcję wielu dotknięć ekranu.

Coraz więcej urządzeń przenośnych jest wyposażanych w kamerę. Otwiera to możliwości do analizy obrazu i rozpoznawania mimiki. Przykładem może być praca [32], w której autorzy zaproponowali sterowanie typowymi aplikacjami komputera przenośnego (gry, mapa) przy pomocy analizy obrazu z wbudowanej kamery.

Potencjalne możliwości sterowania jakie oferuje mimika twarzy mogą zostać również wykorzystane w urządzeniach przenośnych. Jako przykład można podać wspomaganie rozpoznawania mowy poprzez jednoczesną analizę ruchu ust. Rozpoznawanie jednostek fonetycznych mowy (fonemy) może być wspomagane przez równoczesną detekcję analogicznych jednostek w dziedzinie obrazu — tzw. „visemów" (ang. visemes) [15]. Nie zawsze mogą one być traktowane jako odpowiedniki fonemów — raczej niosą informacje uzupełniające, przydatne do wspomagania rozpoznawania mowy np. w przypadku dużego hałasu w pomieszczeniu.

Tabela 3.5: Wybrane elementy mimiki potencjalnie przydatne do realizacji elementarnych zadań — cz.1

| **Nr** | **Elementarne zadania** | **Wymagane sygnały sterujące** | **Odpowiadające elementy mimiki (wg FACS)** | **Uwagi** |
|---|---|---|---|---|
| 1 | Pośrednie wskazanie pozycji 2D | •prawo-lewo<br>•góra<br>•dół | •AU14 (lewy,prawy)<br>•AU12<br>•AU15 | Element AU14 może być wykonany niesymetrycznie (prawa/lewa strona twarzy). Możliwe dwa tryby pracy: emulacja joysticka cyfrowego (detekcja wykonania gestu) oraz emulacja joysticka analogowego (pomiar intensywności gestu). |
| 2 | Wybór obiektu | •wybór1<br>•wybór2 | •AU1+2<br>•AU4 | Emulacja lewego i prawego klawisza myszki. W odróżnieniu od mimowolnego mrugania, mrugnięcia powinny być dłuższe. |
| 3 | Ustawienie bezpośrednie orientacji elementu | •rotacja X<br>•rotacja Y<br>•rotacja Z | •AU51, AU52<br>•AU53, AU54<br>•AU55, AU56 | Orientacja tylko w niewielkim zakresie (±45°) i z małą rozdzielczością. Problem 1 – konieczność dodatkowego sygnału dla odróżnienia stanu zmiany rotacji od stanu powrotu głowy do położenia początkowego. Problem 2 – duży obrót głowy powoduje zniknięcie monitora z pola widzenia człowieka. W przypadku pośredniej zmiany orientacji (na zasadzie joysticka analogowego) problem 1 znika. |

Tabela 3.6: Wybrane elementy mimiki potencjalnie przydatne do realizacji elementarnych zadań — cz.2

| Nr | Elementarne zadania | Wymagane sygnały sterujące | Odpowiadające elementy mimiki (wg FACS) | Uwagi |
|---|---|---|---|---|
| 4 | Pośrednie wskazanie pozycji na ekranie (sposób 2) | •prawo<br>•lewo<br>•góra<br>•dół | •AU52<br>•AU51<br>•AU53<br>•AU54 | Możliwe dwa tryby pracy: emulacja joysticka cyfrowego (detekcja wykonania gestu) oraz emulacja joysticka analogowego (pomiar intensywności gestu). |
| 5 | Pośrednie wskazanie pozycji na ekranie i wybór | •prawo-lewo<br>•góra<br>•dół<br>•wybór | •AU46<br>•AU1+2<br>•AU4<br>•AU19 | Element AU46 może być wykonany niesymetrycznie (prawe/lewe oko). Mruganie może powodować ograniczenie widzialności zmian na ekranie. Potencjalna zaleta – sterowanie kursorem tylko poprzez mimikę górnej części twarzy. Gesty dolnej części twarzy mogą zostać użyte do emulacji przycisków wyboru zdefiniowanych akcji (np. C i D). |
| 6 | Wybranie jednego ze zdefiniowanych wcześniej elementów | •A<br>•B<br>•C<br>•D | •AU1+2<br>•AU4<br>•AU25 (AU26)<br>•AU14 | Emulacja naciśnięcia dedykowanego przycisku. Przyciski A, B – mimika górnej części twarzy. Przyciski C, D – mimika dolnej części twarzy. |

Tabela 3.7: Wybrane elementy mimiki potencjalnie przydatne do realizacji złożonych zadań

| **Nr** | **Złożone zadania** | **Wymagane sygnały sterujące** | **Odp. elementy mimiki (wg FACS)** | **Uwagi** |
|---|---|---|---|---|
| 1 | Przeglądanie otwartego dokumentu lub obrazu | •przesuwanie zawartości tekstu góra-dół<br>•przesuwanie strony „page up/down”<br>•zmiana skali „zoom in/out” | •AU1+2, AU4<br>•AU46<br>•AU25(AU26) oraz AU14+AU12 | Element AU46 może być wykonany niesymetrycznie (prawe/lewo oko). |
| 2 | Nawigacja na stronach WWW | •wskazanie pozycji kursora (prawo, lewo, góra, dół)<br>•przesuwanie zawartości tekstu góra-dół<br>•przejście do poprzedniej/następnej strony<br>•odświeżenie strony | •AU14(l/p), AU12, AU15<br>•AU1+2,AU4<br>•AU46(l/p)<br>•AU25(AU26) | Elementy AU14 i AU46 mogą być wykonane niesymetrycznie (prawa/lewa strona twarzy). |
| 3 | Wspomaganie rozpoznawania mowy | „visemy” | — | Konieczne powiązanie jednostek czynnościowych z „visemami” lub wykorzystanie wyglądu „visemów” |

# Rozdział 4

# Atrybuty elementów mimiki

**Streszczenie**

Rozpoznawanie różnych elementów mimiki w oparciu o informacje wyodrębnione z obrazu, wymaga określenia atrybutów oraz cech charakterystycznych opisujących poszczególne gesty mimiczne (np. kształt, zależności geometryczne, wygląd...). Temu zagadnieniu poświęcony jest niniejszy rozdział. Dokonana analiza elementów mimiki pozwoliła na zdefiniowanie podstawowych atrybutów oraz odpowiadających im cech, które mogą zostać następnie wyodrębnione z obrazu.

## 4.1 Wstęp

Analizując bardziej szczegółowo wybrane elementy mimiki (dodatek: B, tabele: B.4 do B.13), można zauważyć że gesty mimiczne powodują powstawanie różnych zmian w wyglądzie twarzy. Zmiany te mogą zostać wykorzystane do wyodrębnienia z obrazu danego gestu. Przykładowo — marszczenie brwi (rys. B.2) skutkuje m.in. opuszczeniem oraz przyciągnięciem brwi do siebie połączonym z wywołaniem różnego rodzaju zmarszczek (zmiana wyglądu). Z kolei podczas unoszenia brwi (rys. B.1) następuje wyraźna zmiana kształtu brwi (powstają również zmarszczki). Istotna dla gestów mimicznych jest również szybkość ich wykonania.

Opierając się na powyższych obserwacjach przyjęto, że elementy mimiki mogą składać się z następujących elementarnych atrybutów:

- Atrybuty stałe, związane z budową morfologiczną twarzy (oczy, usta, brwi, bruzdy). Są one widoczne cały czas i mogą podlegać deformacji w zależności od rodzaju gestu. Zaliczamy do nich: kształt elementów twarzy (np. brwi) oraz ich wzajemne położenie (np. wielkość szczeliny ust).

- Atrybuty zmienne, wśród których wymienić należy różnego rodzaju zmarszczki, bruzdy oraz fałdy skórne. Może wystąpić również deformacja już istniejących bruzd lub pogłębienie zmarszczek. Powstają one podczas ruchów mimicznych i wywołują najczęściej lokalne zmiany wyglądu.

- Atrybuty dynamiczne takie jak np. szybkość wykonania gestu lub trajektoria ruchu. Mimo iż człowiek jest w stanie rozpoznać np. emocje ze statycznego obrazu, dynamika zmian wyglądu twarzy niesie dużą część informacji. Dlatego też uwzględnienie tego aspektu może ułatwić proces rozpoznawania mimiki. Przykładem wykorzystania dynamiki może być odróżnienie gestów wykonywanych intencjonalnie od wykonywanych mimowolnie (mruganie od przymknięcia oczu).

Przedstawione powyżej atrybuty stanowią raczej abstrakcyjną reprezentację elementów mimiki. Aby były one użyteczne do rozpoznawania, konieczne jest zgromadzenie ewidencji o wystąpieniu danego atrybutu na obrazie (cechy charakterystyczne). Przykładowo, jeśli obiekt jest reprezentowany przez jego kształt, to spodziewać się można istnienia na obrazie krawędzi odpowiadających konturowi obiektu.

Wybór cech jest istotnym zagadnieniem, od którego zależy skuteczność rozpoznawania oraz złożoność obliczeniowa algorytmów. Zagadnienia te opisane zostały w bardzo interesujący sposób w [19][47], gdzie można znaleźć informacje na temat tego w jaki sposób zmieniało się podejście do rozpoznawania w trakcie rozwoju dziedziny przetwarzania i analizy obrazów.

## 4.2 Atrybuty stałe — kształt

Badania nad percepcją człowieka [7], dowiodły że do rozpoznania obiektu człowiek wykorzystuje głównie informację o kształcie. W przypadku kształtu, może być on reprezentowany na wiele sposobów, które można podzielić na trzy kategorie [18]:

- Kontury — przybliżenie kształtu przy pomocy parametryzowanych konturów [3][46], zestawu punktów charakterystycznych lub też aproksymacji krzywej [72].

- Regiony — dekompozycja kształtu na prostsze elementy składowe [34], aproksymacja predefiniowanym elementem geometrycznym, lub reprezentacja przy pomocy zestawu cech związanych z wnętrzem kształtu (np. szkielet) [76].

- Transformacje — kształt reprezentowany przez parametry liniowej lub nieliniowej transformacji przestrzennej z kształtu bazowego, np. transformata Houga [20].

Z powyższych reprezentacji możliwe jest uzyskanie zestawu cech charakterystycznych opisujących dany kształt. W najprostszym przypadku mogą to być, często wykorzystywane, współczynniki kształtu. Możliwe jest również opisanie kształtu przy pomocy sygnałów jednowymiarowych (np. orientacja wzdłuż konturu), oraz wykorzystanie algorytmów takich jak transformacja Fouriera lub falkowa. Bardziej zaawansowane algorytmy, oparte o analizę statystyczną, pozwalają na uwzględnienie zmienności kształtu obiektów tej samej klasy. Wymienić tu należy aktywne modele kształtu (ang. active shape models, point distributed models) [16].

## 4.3 Atrybuty zmienne — lokalne zmiany wyglądu

W przypadku, występujących podczas ruchów mimicznych deformacji istniejących oraz powstawania nowych zmarszczek i bruzd, kształt nie jest wiarygodnym atrybutem. Konieczne jest wyodrębnienie z obrazu lokalnych zmian wyglądu. Można do nich zaliczyć m.in.:

- Teksturę — określenie zmian tekstury skóry wywołanych mimiką realizowane jest przez metody dopasowania wzorców (ang. template matching). Do najprostszych można zaliczyć znormalizowaną korelację (ang. normalized cross-corelation) wykorzystaną w systemie rozpoznawania mimiki zaproponowanym przez Margrit Betke [6]. Do opisu tekstury wykorzystuje się również liniową transformację PCA (analiza składowych głównych) [91] oraz sieci neuronowe [82].

- Lokalną orientację krawędzi — uwypuklanie się zmarszczek i bruzd powoduje powstawanie i znikanie krawędzi na obrazie. Ilościowy i jakościowy pomiar stopnia zmian może być wykonany przy pomocy lokalnych histogramów orientacji [28][85].

- Lokalne cechy obrazu — uwypuklanie się zmarszczek i bruzd powoduje również powstawanie i znikanie innych cech. Do ich wykrycia bardzo dobrze nadają się falki Gabora, które umożliwiają detekcję cech w różnych skalach i orientacjach [89]. Dodatkową zaletą falek Gabora jest ich teoretyczna niewrażliwość na zmiany oświetlenia sceny. Działanie tych falek można porównać do operacji realizowanych przez pierwszorzędową korę wzrokową człowieka [67]. Do opisu zmian obrazu wywołanych mimiką, wykorzystany może być również algorytm SIFT (ang. Scale Inwariant Features Transform), który pozwala na detekcję cech charakteryzujących się odpornością na obrót, zmianę skali oraz pewną odpornością na zmiany oświetlenia [65].

## 4.4 Atrybuty dynamiczne — ruch i jego trajektoria

Istotnym elementem mimiki są jej dynamiczne atrybuty. Szybkość wykonania gestu mimicznego, trajektoria ruchu poszczególnych cech lub kolejność wykonywania jednostek czynnościowych, niosą również istotne informacje. Wśród metod pozwalających na detekcję zmian można wyróżnić m.in.:

- Detekcję ruchu — wykrycie ruchu w określonym miejscu na obrazie twarzy stanowi istotną przesłankę o wykonywanym geście mimicznym. Przykładowo wykrycie zmian w obszarze ust może świadczyć o wykonaniu gestu ustami (uśmiech, otwarcie ust, itp.). Wśród metod wykorzystywanych w rozpoznawaniu mimiki znaleźć można algorytm wykorzystujący szablony ruchu [92].
- Wykorzystanie modeli ruchu. Wykonanie gestu powoduje ruch niektórych elementów charakterystycznych twarzy (np. brwi). Śledzenie tych zmian w czasie pozwala na estymację parametrów ruchu, m.in. trajektorii oraz szybkości. W praktyce stosowany jest filtr Kalmana, lub w przypadku nie spełnienia założeń tego filtru (nieliniowości, wielomodalny rozkład szumu obserwacji), filtr cząsteczkowy (ang. particle filter, condensation algorithm) [8]. Oprócz śledzenia cech wykorzystywana jest również informacja uzyskana z analizy lokalnego przepływu optycznego (ang. local optic flow).

# Część III

# Automatyczne wyodrębnianie z obrazu oraz rozpoznawanie elementów mimiki

# Rozdział 5

# Elementy systemu automatycznego rozpoznawania mimiki

**Streszczenie**

Rozpoznawanie obiektów przez ludzi wydaje się łatwe i bezproblemowe. Na podstawie ogólnej wiedzy o danym obiekcie, np. sylwetka ludzka — człowiek natychmiast jest w stanie rozpoznać nowe osoby, niezależnie od tego z jakiego punktu widzenia są one obserwowane czy też np. siedzą lub stoją. Z punktu widzenia komputerowej analizy obrazów, rozpoznawanie wymaga wyodrębnienia z obrazu użytecznej informacji na temat interesujących obiektów.

W rozdziale zaproponowano schemat struktury systemu automatycznego wyodrębniania z obrazu elementów mimiki, zawierający kilka bloków składowych realizujących wymagane zadania. Omówione zostały etapy przetwarzania i analizy obrazu, na które składają się: selektywne przetwarzanie informacji (segmentacja), detekcja i lokalizacja twarzy, wyodrębnianie cech, klasyfikacja oraz adaptacja systemu. Wśród istotnych rezultatów wymienić można opracowaną metodę doboru przestrzeni barw dla algorytmu segmentacji oraz algorytm detekcji twarzy.

## 5.1 Wstęp

Jedną z podstawowych funkcji systemu wzrokowego człowieka jest rozpoznawanie obiektów, czyli zdolność do powiązania informacji docierającej do ludzkiego oka (fale elektromagnetyczne) z posiadaną wiedzą na temat otaczającego świata. Pozwala to człowiekowi na wyciąganie wniosków oraz podejmowanie działań.

Przykładem może być powitanie osoby znajomej dzięki rozpoznaniu jej twarzy lub określenie jej zamiarów poprzez obserwację mimiki.

Istnieje bogata literatura poświęcona ludzkiemu systemowi wzrokowemu, wśród wielu publikacji wymienić można pracę powstałą w Katedrze Automatyki AGH [58]. W niniejszym rozdziale przedstawiono natomiast spojrzenie na problem z punktu widzenia komputerowej analizy obrazów.

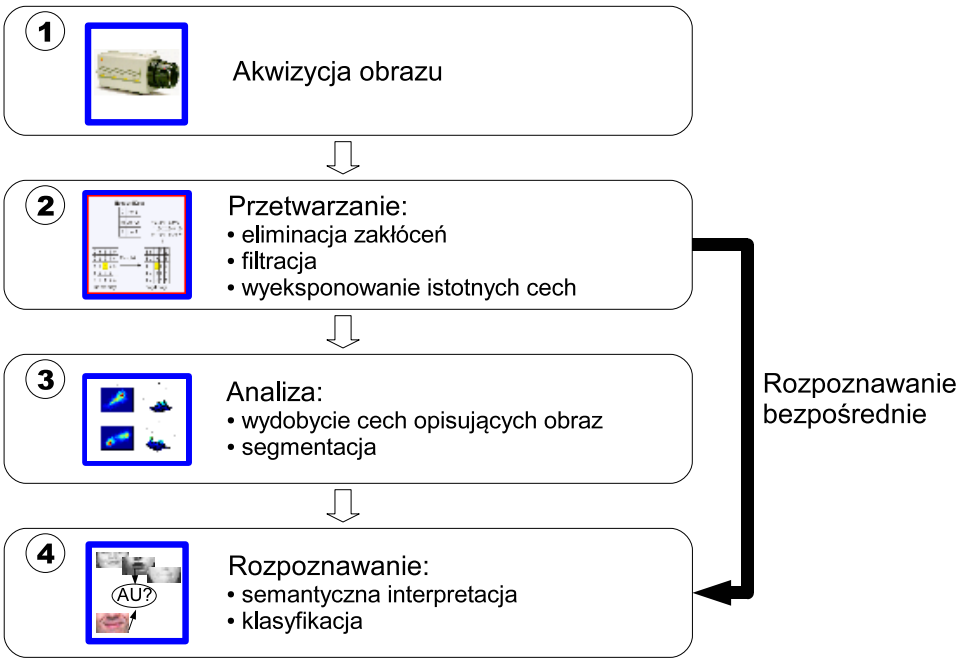
*„Celem sztucznego przetwarzania lub analizy obrazu jest takie automatyczne przetworzenie i przeanalizowanie obrazu wybranych obiektów lub całego otoczenia systemu zautomatyzowanego, aby uzyskać użyteczną informację na temat interesujących obiektów (np. będących przedmiotem manipulacji ze strony robota przemysłowego) lub na temat otoczenia, które może wpływać (i zwykle znacząco wpływa) na sterowany automatycznie proces”* [84].

Klasyczny schemat komputerowego przetwarzania obrazu przedstawiony został na rysunku 5.1. Składa się on z kilku etapów — począwszy od akwizycji obrazów, a skończywszy na rozpoznaniu obrazu i jego semantycznej interpretacji. Celem wstępnego przetwarzania jest eliminacja zakłóceń, często poprawa jakości obrazu (np. wyostrzanie), oraz wyeksponowanie istotnych cech (np. detekcja krawędzi). Za nim następuje etap analizy pozwalający na wyodrębnienie z obrazu interesujących obiektów lub ich części (segmentacja), oraz wydobycie cech charakteryzujących obiekty (np. współczynniki kształtu). W literaturze podejście takie określane jest często jako „metody oparte na ekstrakcji cech” (ang. feature-based approaches). Możliwe jest również bezpośrednie rozpoznanie obrazu z pominięciem etapu analizy (ang. image-based approaches). Przykładem mogą być metody dopasowania wzorców, w których zawartość obrazu porównywana jest bezpośrednio z zapamiętanym wzorcem wyglądu obiektu.

W kontekście automatycznego rozpoznawania elementów mimiki, klasyczny schemat przetwarzania obrazu przedstawia się zazwyczaj w odmienny sposób [26]. Na rysunku 5.2 przedstawiono schemat systemu rozpoznawania mimiki zaadoptowany i dostosowany na potrzeby niniejszej rozprawy.

Pierwszym etapem jest recepcja obrazu oraz selektywne przetwarzanie informacji, których celem jest wydzielenie z obrazu obszarów zawierających użyteczne dane. Proces ten porównać można do działania drogi wzrokowej człowieka, w szczególności funkcjonalności ciała kolankowatego bocznego, odpowiedzialnego za porządkowanie informacji o kolorze, ruchu i formie [58].

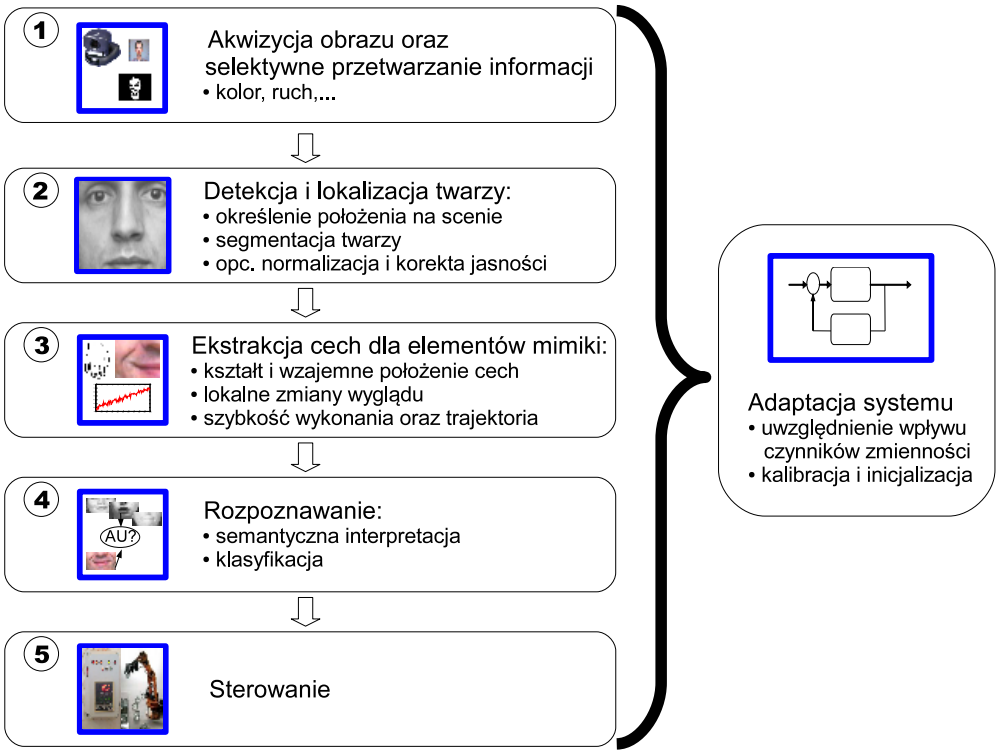
Kolejnym etapem jest detekcja i lokalizacja twarzy na obrazie, której celem jest odróżnienie twarzy od innych obiektów występujących na scenie oraz segmentacja obszarów zainteresowań. Etap ten często łączony jest z normalizacją twarzy (pod względem skali, orientacji połączonej z korektą jasności), aczkolwiek istnieją metody rozpoznawania, które tego nie wymagają. W literaturze funkcjonuje wiele rodzajów usystematyzowania i pogrupowania metod detekcji i lokalizacji twarzy [37],[99].




Rysunek 5.1: Operacje składające się na proces automatycznego widzenia.

Ponieważ elementy mimiki mogą mieć różny charakter, konieczne jest wyodrębnianie z obrazu cech opisujących poszczególne gesty mimiczne. Takimi cechami mogą być np. kształt, zależności geometryczne, kolor lub też bezpośredni wygląd obiektu. Z etapem tym ściśle łączy się kolejny krok algorytmu jakim jest rozpoznawanie elementów mimiki. Podczas rozpoznawania, dane będące wynikiem etapu wyodrębniania cech podlegają klasyfikacji do jednej z założonych wcześniej kategorii. Obszerną systematykę metod rozpoznawania elementów mimiki można znaleźć w pracy [26]. Zagadnienia te zostały opisane w rozdziałach 6 oraz 7.

Duża zmienność wyglądu twarzy oraz elementów mimiki, wynikająca z wpływu wielu czynników, powoduje iż praktycznie niemożliwe jest opracowanie jednej uniwersalnej metody rozpoznawania. Dlatego istotnym zagadnieniem przy projektowaniu systemu rozpoznawania mimiki, jest jego dostosowywanie się do człowieka oraz zmieniających się warunków otoczenia. Zagadnienia adaptacji zostały opisane w rozdziale 8.

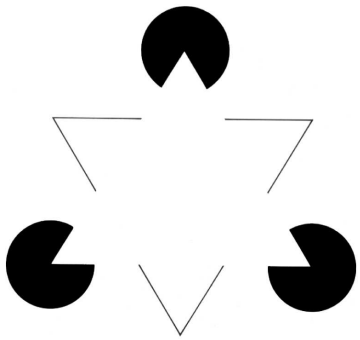



Rysunek 5.2: Ogólny schemat systemu rozpoznawania mimiki.

## 5.2 Selektywne przetwarzanie informacji oraz segmentacja obszarów zainteresowań

W procesie postrzegania sceny przez człowieka istotną rolę pełni selektywne przetwarzanie informacji istotnych dla aktualnie przyjętego celu działania. Potwierdzają to rezultaty badań funkcjonowania zmysłu wzroku oraz kory wzrokowej (ang. visual cortex) [31]. Przykładem skupiania uwagi na wybranych elementach postrzeganej sceny jest rozmowa z drugą osobą — człowiek często spogląda na twarz rozmówcy aby zorientować się w jego aktualnym stanie emocjonalnym. Skupiając uwagę na elementach twarzy (brwi, oczy, usta) oczekuje ich zmian w zależności od mimiki i jej znaczenia (radość, gniew, smutek, zaskoczenie). Dzięki temu człowiek może zareagować w odpowiedni sposób. Trzeba jednak zwrócić uwagę iż sam proces selektywnego przetwarzania informacji nie jest równoznaczny z prawidłowym rozpoznaniem rzeczywistego obiektu. Często zdarza się że „widzimy tylko to czego się spodziewamy". Przykładem mogą być różnego rodzaju iluzje takie jak na rysunku 5.3.

Z selektywnym przetwarzaniem informacji związany jest również proces grupowania percepcyjnego (ang. perceptual grouping). Tematyka ta omówiona zo-



Rysunek 5.3: Przykład optycznej iluzji — „trójkąt Kanizsa".

stała w teorii percepcji, nazywanej teorią Gestalt (ang. gestalt psychology), która została zaproponowana przez berlińską szkołę psychologii na początku XX wieku [48][94]. Opisywała ona pierwotnie organizowanie przez system nerwowy człowieka percepcji i spostrzeżeń odbieranych ze wzrokowego pola widzenia. Zauważono, iż ludzie mają tendencję do grupowania i konfigurowania bodźców pochodzących z wzrokowego pola widzenia wg zasad: bliskości, podobieństwa, domykania się w pewną całość oraz wyróżnionej cechy prostoty. Później twórcy teorii Gestalt uogólnili te zasady na opis działania innych zmysłów oraz przenieśli spostrzeżenia na pole psychologii, dostrzegając, że te same prawa dotyczą także zjawisk psychicznych.

W kontekście rozpoznawania obrazów przez maszynę, selektywne przetwarzanie informacji ma istotne znaczenie i realizowane jest najczęściej poprzez segmentację. W literaturze proces segmentacji określany jest również jako problem grupowania przestrzennego pikseli (ang. spatial grouping). W przypadku analizy obrazów dynamicznych (sekwencja video), konieczne jest również uwzględnienie kontekstu czasowego (ang. correspondence problem). Polega on na ustanowieniu asocjacji czasowych pomiędzy elementami obrazu należącymi do tej samej części fizycznego obiektu. Segmentacja która stanowi wstępny etap przetwarzania, zapewnia szereg wskazówek dla algorytmów rozpoznawania obrazu. Pozwala na zawężenie obszaru poszukiwań, poprzez wyeliminowanie pikseli które nie są istotne dla dalszej analizy. Zawężenie obszaru poszukiwań skutkuje zmniejszeniem czasu obliczeń, a w konsekwencji ułatwia pracę systemu w czasie rzeczywistym.

Do najczęściej wykorzystywanych w selektywnym przetwarzaniu informacji metod można zaliczyć:

- Kolor — odgrywa istotną rolę w rozumieniu obrazu przez człowieka. Jest silną wskazówką odporną na zmiany skali, orientacji oraz przysłonięcie części obiektu.

- Ruch — pozwala na stwierdzenie czy w obserwowanej scenie zaszły jakieś

zmiany np. w położeniu cech. Powoduje zmiany intensywności pikseli obrazu.

- Tekstura powierzchni — umożliwia odróżnienie od siebie części obiektów oraz wyznaczenie ich kształtu.

- Rozbieżność stereoskopowa (ang. stereo disparity) — pozwala na ocenę odległości.

- Charakterystyczne cechy (ang. salient features) — widoczne cechy obiektów odróżniające się w dużym stopniu od otoczenia (np. plamy, krawędzie, narożniki).

Z powyższych wskazówek szczególnie użyteczny w systemach detekcji twarzy jest kolor, który odgrywa istotną rolę w rozumieniu obrazu przez człowieka. Pozwala na odróżnienie od siebie obiektów — wśród nich wyróżnić można m.in. twarze innych ludzi. Również w komputerowej analizie obrazów wykorzystuje się informacje o kolorze do segmentacji obszarów zainteresowań (ang. ROI — region of interest).

### 5.2.1 Segmentacja na podstawie barwy skóry

Barwa skóry człowieka teoretycznie jest niezależna od rasy (z wyjątkiem albinosów) [37], dlatego też segmentacja w oparciu o kolor pozwala na wydzielenie miejsc obrazu w których istnieje duże prawdopodobieństwo wystąpienia twarzy. Wyniki mogą zostać wykorzystane do właściwej detekcji i lokalizacji twarzy. Właściwości niektórych przestrzeni kolorów (YCbCr) pozwalają również na wydzielenie z obrazu potencjalnych obszarów oczu oraz ust [38]. Jest to pomocne w weryfikacji wyników detekcji twarzy.

W pracy [14] autorzy skupili się na automatycznej segmentacji twarzy przy użyciu mapy kolorów skóry człowieka. Celem prowadzonych badań była inteligentna selekcja informacji z obrazu dla potrzeb aplikacji wideotelefonu. W artykule [74] przedstawiona została realizacja podobnego celu — wybór obszarów zainteresowań dla potrzeb systemu kodowania i kompresji sygnału wideo. Segmentacja i śledzenie twarzy stanowi w opisywanym systemie istotny element realizowany poprzez użycie danych o kolorze. W wielu przypadkach segmentacja oparta o barwę skóry jest częścią etapu detekcji i rozpoznawania twarzy [38]. Często skuteczność detekcji zwiększa się poprzez kompensację oświetlenia i wykorzystanie dodatkowej wiedzy (np. budowa morfologiczna twarzy). W kontekście wykorzystania koloru wymienić również należy prace prowadzone w polskich ośrodkach naukowych. Jedną z nich jest segmentacja twarzy dla potrzeb systemu rozpoznawania mimiki [41] — laboratorium Biocybernetyki AGH. Interesujące zastosowanie przedstawiła autorka pracy doktorskiej [57] — wykorzystanie in-

formacji o barwie skóry w algorytmie rozpoznawania znaków polskiego alfabetu palcowego.

Wykorzystanie informacji o kolorze wymaga określenia dwóch podstawowych założeń:

- wybór przestrzeni barw (ang. color space) — różne przestrzenie kolorów charakteryzują się różnymi własnościami,
- sposób modelowania rozkładu kolorów segmentowanego obiektu.

Jedną z podstawowych przestrzeni barw jest RGB, szeroko wykorzystywany w torach obróbki sygnału wizyjnego. Większość barw można wywołać przez zmieszanie w ustalonych proporcjach trzech wiązek światła o barwie czerwonej, zielonej i niebieskiej, czyli światła o odpowiedniej częstotliwości fali elektromagnetycznej. Przestrzeń kolorów RGB stanowi również punkt wyjścia do transformacji pikseli obrazu do innych przestrzeni barw. Wykorzystanie RGB do segmentacji skóry człowieka napotyka na szereg trudności, wynikających ze specyfiki tej przestrzeni. Przede wszystkim zwrócić należy uwagę na istnienie wysokiej korelacji pomiędzy składowymi oraz brak odseparowania informacji o chrominancji i luminancji. Istotny jest również brak percepcyjnej jednolitości (ang. perceptual uniformity) polegający na tym, że jednostkowe zmiany jednej ze składowej nie są tak samo postrzegane dla różnego zakresu tej składowej [1]. Dlatego w algorytmach segmentacji wykorzystuje się inne przestrzenie barw, otrzymywane poprzez odpowiednie transformacje:

- znormalizowana RGB [38],
- HSI, HSV, HSL (Hue Saturation Intensity – Value, Lightness), [60],
- YCrCb [14],
- inne.

Więcej informacji na temat wykorzystania różnych przestrzeni kolorów można znaleźć w pracy [77].

Celem modelowania rozkładu kolorów jest segmentacja obiektu (twarz, ręka) — inaczej mówiąc określenie reguły decyzyjnej pozwalającej na zaklasyfikowanie danego piksela jako części obiektu lub tła. Wśród stosowanych metod wyróżnić można:

- segmentację poprzez bezpośrednie określenie zestawu reguł [66],
- metody nieparametryczne nie wymagające modelu lecz bezpośrednio wykorzystujące rozkład statystyczny kolorów pikseli uzyskany z analizy przykładów [78],

- metody parametryczne, w których na podstawie danych uczących tworzony jest model rozkładu kolorów [95],

- metody dynamiczne wykorzystywane w przypadku jednoczesnej detekcji i śledzenia twarzy.

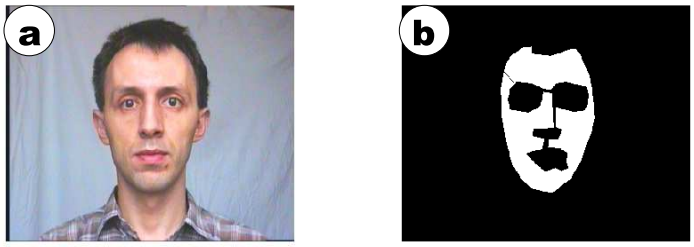
Zestawienie różnych metod modelowania rozkładu kolorów i segmentacji można znaleźć w pracy [93].

### 5.2.2 Dobór przestrzeni kolorów oraz metody segmentacji

Istnieje wiele metod segmentacji wykorzystujących różne przestrzenie kolorów. Porównanie skuteczności różnych metod segmentacji w oparciu o kolor jest trudne ze względu na testowanie dla różnych baz danych zawierających obrazy twarzy pobrane w różnych warunkach i różnymi kamerami.

Z punktu widzenia systemu rozpoznawania mimiki istotny jest odpowiedni wybór reprezentacji kolorów oraz sposobu modelowania pikseli skóry człowieka. W niniejszej pracy porównano następujące przestrzenie kolorów: znormalizowaną RGB, HSV oraz YCbCr. Spośród metod segmentacji wybrano do porównania następujące: bezpośrednie określenie zestawu reguł, znormalizowane tablice przekodowań (ang. LUT – look-up-table) oraz gaussowski model barwy skóry. Opis przestrzeni kolorów oraz metod segmentacji zamieszczony został w dodatku A (rozdziały: A.1 oraz A.2). W dalszej części rozprawy używane będzie pojęcie „modelu barwy skóry", które odnosi się do sposobu reprezentacji barwy skóry człowieka, dla potrzeb algorytmów segmentacji.

W celu obiektywnego porównania skuteczności segmentacji przy pomocy różnych metod i przestrzeni kolorów, na wybranych obrazach sekwencji video (uznanych za referencyjne) wyznaczono manualnie obszary zawierające piksele reprezentujące skórę twarzy. W ten sposób utworzona została binarna maska twarzy (rys. 5.4).




Rysunek 5.4: Przykładowy obraz sekwencji zawierający twarz (a) oraz przyjęta maska twarzy (b).

Następnie zdefiniowano kryterium porównawcze, w postaci funkcji celu (5.1):

$$f_s(\psi) = \frac{nb_{nonface}}{n_{nonface}} - \frac{nb_{face}}{n_{face}} \tag{5.1}$$

gdzie:

$f_s(\psi)$ – funkcja celu

$nb_{face}$ – ilość pikseli poprawnie zaklasyfikowanych do skóry w obszarze maski twarzy

$n_{face}$ – całkowita ilość pikseli obszarze maski twarzy

$nb_{nonface}$ – ilość pikseli niepoprawnie zaklasyfikowanych do skóry poza obszarem maski twarzy

$n_{nonface}$ – całkowita ilość pikseli poza obszarem maski twarzy

Funkcja przyjmuje wartość minimalną dla przypadku gdy wszystkie piksele należące do twarzy, zostały poprawnie do niej zaklasyfikowane oraz piksele tła zostały zaklasyfikowane do obszaru tła.

Ta sama funkcja celu zastosowana została także do automatycznego tworzenia modeli barwy skóry dla poszczególnych metod. W przypadku metody bezpośredniego określenia zestawu reguł, modelem są wyznaczone progi binaryzacji poszczególnych składowych koloru (Cb-Cr, Y, r-g[1]). Optymalne progi wyznaczono minimalizując funkcję celu (5.2):

$$\min_{\psi} f_s(\psi) \tag{5.2}$$

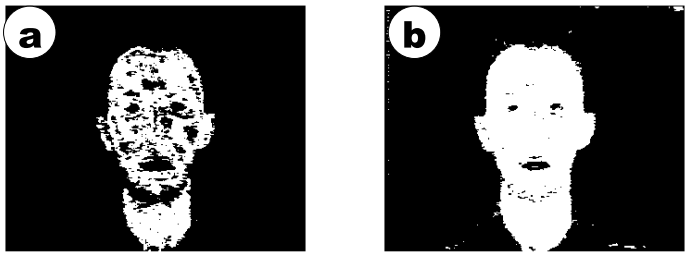
Jako wartości początkowe wektora parametrów przyjęto progi określone jako przedział dwóch odchyleń standardowych od wartości oczekiwanej pikseli w obszarze maski twarzy (5.3):

$$\mu_\xi \pm 2 \cdot std_\xi \tag{5.3}$$

Dodatkowo w funkcji celu następuje sortowanie wartości progów co pozwala uniknąć sytuacji, gdy dolny próg jest większy od górnego. Optymalizacji dokonano przy pomocy funkcji **fmincon** z modułu Optimization Toolbox środowiska MATLAB, która opiera się na metodzie sekwencyjnego programowania SQP (ang. Sequential Quadratic Programming). Na rysunku 5.5 pokazane są wyniki segmentacji dla progów przyjętych jako wartości początkowe oraz dla progów wyznaczonych poprzez optymalizację przyjętej funkcji celu.

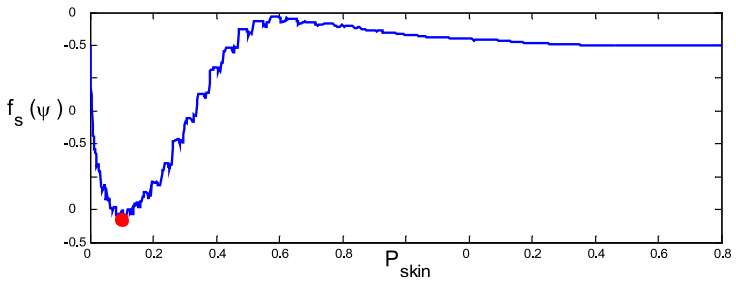
W metodach opartych o tablice przekodowań (LUT) oraz gaussowskiej funkcji rozkładu prawdopodobieństwa, rezultatami są obrazy prawdopodobieństwa — wartościom bliskim 1 odpowiadają piksele należące z dużą pewnością do twarzy,

[1] w rozprawie użyto oznaczenia „r-g” dla składowych znormalizowanej przestrzeni RGB




Rysunek 5.5: Wyniki segmentacji: (a) dla progów przyjętych wg kryterium (5.3), (b) dla progów wyznaczonych metodą minimalizacji funkcji celu (5.2)

wartościom w pobliżu zera — tła. Obrazy te mogą zostać wykorzystane bezpośrednio, np. w dalszej analizie skupień, bądź też można z nich uzyskać obrazy binarne. W tym celu konieczne jest określenie progu pozwalającego na klasyfikację pikseli do jednej z dwu kategorii — twarz i tło. Próg ten jest wyznaczany automatycznie poprzez optymalizację funkcji celu. Ponieważ w tym przypadku funkcja jest jednowymiarowa, wystarczające jest obliczenie jej wartości dla liniowo zmieniającego się progu binaryzacji oraz znalezienie minimum — wykres (rys 5.6).




Rysunek 5.6: Wykres funkcji celu dla przypadku jednowymiarowego.

Porównanie rezultatów segmentacji przeprowadzono dla obrazów sekwencji video pochodzących z różnych kamer, w odmiennych sytuacjach oraz w różnych warunkach oświetleniowych:

- sekwencja 1 — kamera EVI, optymalne warunki akwizycji (jednolite tło, natężenie oświetlenia około 320lux), (rys. 5.7a),
- sekwencja 2 — kamera EVI, słabe oświetlenie (jednolite tło, natężenie oświetlenia około 40lux), (rys. 5.7b),

- sekwencja 3 — kamera USB Creative WebCam, skomplikowana scena i warunki akwizycji (tzn. skomplikowane tło, boczne oświetlenie, okulary), (rys. 5.7c),
- sekwencja 4 — kamera EVI, skomplikowana scena i warunki akwizycji (tzn. skomplikowane tło, boczne oświetlenie, okulary), (rys. 5.7d).

Każdy obraz został poddany operacji usuwania szumów (filtr medianowy o masce 5*5). Segmentację przeprowadzono dla wcześniej opisanych przestrzeni kolorów oraz metod.

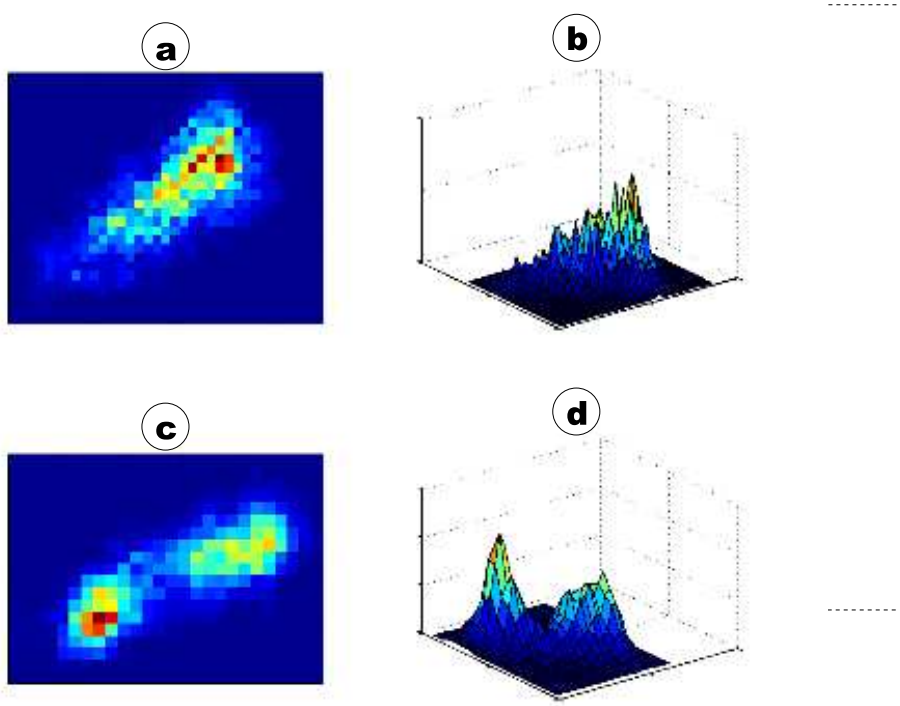

Rysunek 5.7: Przykładowe obrazy z wybranych sekwencji video.

Rezultaty segmentacji twarzy, wraz z wyznaczonymi wartościami funkcji celu, zamieszczone zostały w dodatku D (rysunki D.1 — D.4, tabele D.1 — D.4).

W przypadku dobrych warunków akwizycji i braku na scenie obiektów o barwie skóry, wyniki segmentacji są dobre i pozwalają na wyodrębnienie z obrazu twarzy. Wybór metody oraz przestrzeni kolorów nie jest krytyczny, aczkolwiek metoda wykorzystująca zestaw reguł okazuje się najskuteczniejsza (najmniejsza wartość funkcji celu).

Duży wpływ na skuteczność algorytmu ma nieprawidłowe oświetlenie sceny (sekwencja 2). W tym przypadku wybór odpowiedniej metody oraz przestrzeni kolorów jest istotny dla prawidłowej segmentacji. W szczególności boczne oświetlenie (często spotykane ze względu na typowe ustawienie biurka komputera) ma niekorzystny wpływ. Wynika to z powstawania cieni, które znacząco wpływają na rejestrowaną barwę skóry. Widoczne jest to na dwuwymiarowych histogramach rozkładu prawdopodobieństwa kolorów pikseli (rys 5.8). W przypadku oświetlenia jednorodnego rozkład jest zgodny z teorią [93] — barwa skóry człowieka tworzy teoretycznie w przestrzeni kolorów zwarty, niewielki obszar (rozkład prawdopodobieństwa pikseli obszaru skóry normalny lub skośny). Natomiast dla bocznego oświetlenia rozkład jest wielomodalny.

Podczas eksperymentów natrafiono na sytuację, gdy prawidłowo dobrane progi segmentacji (metody regułowa), były nieskuteczne do segmentacji innych obrazów tej samej sekwencji wideo. Dlatego konieczne okazało się sprawdzenie stabilności progów dla testowanych sekwencji. W tym celu wyznaczono progi binaryzacji na wybranym z sygnału wideo obrazie, a następnie dla pozostałych obrazów sekwencji (kilkadziesiąt obrazów pobranych dla stabilnych globalnych warunków




Rysunek 5.8: Rozkłady prawdopodobieństwa pikseli obszaru skóry dla: (a)(b) sekwencja 1; (c)(d) sekwencja 4. Rozkłady przedstawione w postaci: (a)(c) obrazu będącego odwzorowaniem histogramu 2D, (b)(d) wykresu powierzchniowego histogramu 2D.

oświetleniowych) wyznaczono wartość funkcji celu $f_S(\Psi)$. Na podstawie otrzymanych wyników obliczono wartość oczekiwaną i odchylenie standardowe funkcji celu dla poszczególnych sekwencji oraz przestrzeni barw. Wyniki przedstawia tabela 5.1. Na jej podstawie można zauważyć że najlepsze wyniki uzyskano dla przestrzeni Cb-Cr, natomiast przestrzeń r-g okazuje się cechować największą zmiennością, dla występujących niewielkich wahań oświetlenia sceny.

Tabela 5.1:

| | **r-g** | **Cb-Cr** | **Hue** |
|---|---|---|---|
| **sekwencja 1** | $\mu = -0.83757$<br>$\sigma = 0.051141$ | $\mu = -0.86744$<br>$\sigma = 0.0043139$ | $\mu = -0.81677$<br>$\sigma = 0.005981$ |
| **sekwencja 2** | $\mu = -0.56997$<br>$\sigma = 0.17913$ | $\mu = -0.69882$<br>$\sigma = 0.0089865$ | $\mu = -0.32583$<br>$\sigma = 0.019995$ |
| **sekwencja 3** | $\mu = -0.26569$<br>$\sigma = 0.27182$ | $\mu = -0.50389$<br>$\sigma = 0.018087$ | $\mu = -0.53638$<br>$\sigma = 0.054475$ |
| **sekwencja 4** | $\mu = -0.14738$<br>$\sigma = 0.11159$ | $\mu = -0.22583$<br>$\sigma = 0.057913$ | $\mu = -0.35548$<br>$\sigma = 0.042728$ |

## 5.3 Detekcja i lokalizacja twarzy

Selektywne przetwarzanie informacji (segmentacja) pozwala na określenie obszarów zainteresowań dzięki wykorzystaniu wskazówek takich jak kolor, ruch, tekstura. Nie zapewnia jednak jednoznacznego stwierdzenia istnienia na obrazie twarzy człowieka. Konieczna jest weryfikacja, czy znaleziony obiekt jest twarzą oraz określenie jest położenia i wielkości. Cały proces nazywany jest w literaturze detekcją oraz lokalizacją twarzy. Istniejące metody detekcji i lokalizacji twarzy na obrazach statycznych, można podzielić na kilka grup:

1. metody wykorzystujące wiedzę o budowie morfologicznej twarzy (ang. knowledge based methods),

2. metody oparte na detekcji strukturalnych cech obrazu, niezależnych od zmian oświetlenia oraz punktu widzenia kamery (ang. feature invariant approaches),

3. metody oparte na wyszukiwaniu wzorców twarzy oraz jej morfologicznych elementów, np. oczy (ang. template matching methods, appearance-based methods).

Wyczerpujące ich omówienie można znaleźć w pracach: [96],[37]. Pierwsza grupa metod (punkt -1-), która wykorzystuje wiedzę o budowie morfologicznej twarzy, reprezentuje podejście „od ogółu do szczegółu" (ang. top-down). Detekcja odbywa się na podstawie zestawu reguł przyjętych na podstawie wiedzy o anatomii człowieka [50]. Przykładem może być: symetria elementów twarzy, zależności geometryczne pomiędzy nimi oraz informacje o intensywności pikseli w poszczególnych obszarach.

Przykładem odmiennego podejścia jest strategia „od szczegółu do ogółu" (ang. bottom-up). Jest to grupa algorytmów opartych na ekstrakcji strukturalnych cech obrazu (punkt -2-). Głównym założeniem jest niewrażliwość cech na zmiany położenia twarzy w stosunku do kamery, zmiany oświetlenia bądź też indywidualnego wyglądu człowieka. Detekcja twarzy realizowana jest poprzez

dopasowaniu znalezionych cech (np. własności pikseli obrazu takie jak poziomy szarości, kolor, krawędzie) do modelu twarzy. Model ten może być o różnym stopniu szczegółowości, dwu- lub trójwymiarowy. Przykładem jest publikacja [38], w której wykorzystano informacje o kolorze skóry człowieka do identyfikacji obszaru twarzy oraz jej elementów charakterystycznych (oczy, usta). Autorzy zaproponowali algorytm lokalizacji oczu i ust, oparty na analizie składowych chrominancji Cb-Cr. Poprzez porównanie prostego modelu twarzy ze znalezionymi cechami uzyskano skuteczną detekcję twarzy. Główną wadą metod opartych na detekcji cech jest ich wrażliwość na dobór parametrów oraz problem niejednoznaczności dopasowania cech do modelu. Charakteryzują się jednak małą złożonością obliczeniową, co pozwala na ich stosowanie w systemach pracujących w czasie rzeczywistym.

Osobną grupę (punkt -3-) stanowią metody oparte na wyszukiwaniu wzorców wyglądu twarzy. Zadanie detekcji postawione jest jako problem klasyfikacji nieznanego obrazu do jednej ze znanych kategorii utworzonych w procesie uczenia na przykładach. Jedną z metod częściej wykorzystywanych do detekcji twarzy, jest algorytm nazywany „eigenfaces" [91]. Na podstawie zestawu przykładów oraz analizy składowych głównych PCA (ang. principal component analysis), tworzony jest statystyczny model twarzy. Detekcja na nieznanym obrazie odbywa się poprzez przesuwanie okna analizy (w różnych skalach) i porównanie wycinka obrazu z wzorcem. Porównanie odbywa się poprzez rzutowanie wycinka do przestrzeni bazowej i obliczenie błędu średniokwadratowego określającego miarę podobieństwa do twarzy (ang. DFFS – distance-from-face-space). Metoda ta wykorzystywana jest również do rozpoznawania twarzy. Skuteczność tego podejścia jest większa niż algorytmów opartych na ekstrakcji cech, ale zależy w dużym stopniu od sposobu uczenia oraz wielkości bazy wzorców.

Wyżej przedstawiony podział dotyczy algorytmów pracujących na obrazach statycznych, natomiast detekcja twarzy na sekwencji wideo, wymaga uwzględnienia szeregu dodatkowych czynników. Jednym z nich jest zmienność warunków akwizycji dla sekwencji obrazów wpływająca na jej jakość. W rzeczywistych warunkach interakcji człowieka z komputerem warunki akwizycji mogą zmieniać się dość znacznie w czasie (np. oświetlenie naturalne i sztuczne, pozycja osoby przed kamerą itp.). Drugim czynnikiem — który pozytywnie wpływa na proces detekcji — jest fakt iż w przypadku sekwencji obrazów dostępne są dodatkowe wskazówki, takie jak ruch oraz zależności czasowe pomiędzy poszczególnymi cechami obrazu. Dowiedziono [33], iż dla człowieka wyszukiwanie twarzy na sekwencji wideo jest łatwiejsze niż na losowo prezentowanych obrazach statycznych. Metody detekcji i lokalizacji twarzy na sekwencji obrazów, wykorzystujące zarówno informacje przestrzenne jak i czasowe, rozwinęły się dopiero niedawno i — wg [99] — w dalszym ciągu wymagają dalszych badań w tym kierunku.

### 5.3.1 Algorytm detekcji i lokalizacji twarzy

Segmentacja twarzy na podstawie barwy skóry umożliwia wydzielenie miejsc obrazu w których istnieje duże prawdopodobieństwo wystąpienia twarzy. Nie pozwala jednak na jednoznaczną detekcję i lokalizację twarzy. Konieczne jest wykorzystanie dodatkowych informacji. W niniejszej pracy połączono algorytm wstępnej segmentacji twarzy z metodą detekcji jej elementów charakterystycznych. Lokalizacja oczu i ust wraz z zestawem reguł opracowanych na podstawie wiedzy o budowie morfologicznej twarzy umożliwia bardziej skuteczną jej detekcję. Schemat blokowy algorytmu przedstawiony został na rysunku 5.9.

W pierwszej kolejności odbywa się wstępna filtracja (filtr medianowy), dzięki której uzyskuje się minimalizację wpływu szumów. Operację tą poprzedza zmniejszenie wymiarów obrazu, co pozwala na skrócenie czasu obliczeń. Segmentacja obszaru twarzy odbywa się jedną z metod opisaną w poprzednim rozdziale (por. 5.2.2) — wybór metody, przestrzeni kolorów oraz parametrów algorytmów został zrealizowany na podstawie kryterium najlepszej segmentacji (minimalna wartość funkcji celu dla wybranego obrazu z danej sekwencji video). W wyniku otrzymuje się binarną maskę twarzy (5.4). Maska ta poddawana jest dodatkowo następującym operacjom morfologicznym: usunięcie niewielkich obiektów, zamknięcie oraz zalewanie otworów.

$$BW_{face} = \Gamma(BW) \tag{5.4}$$

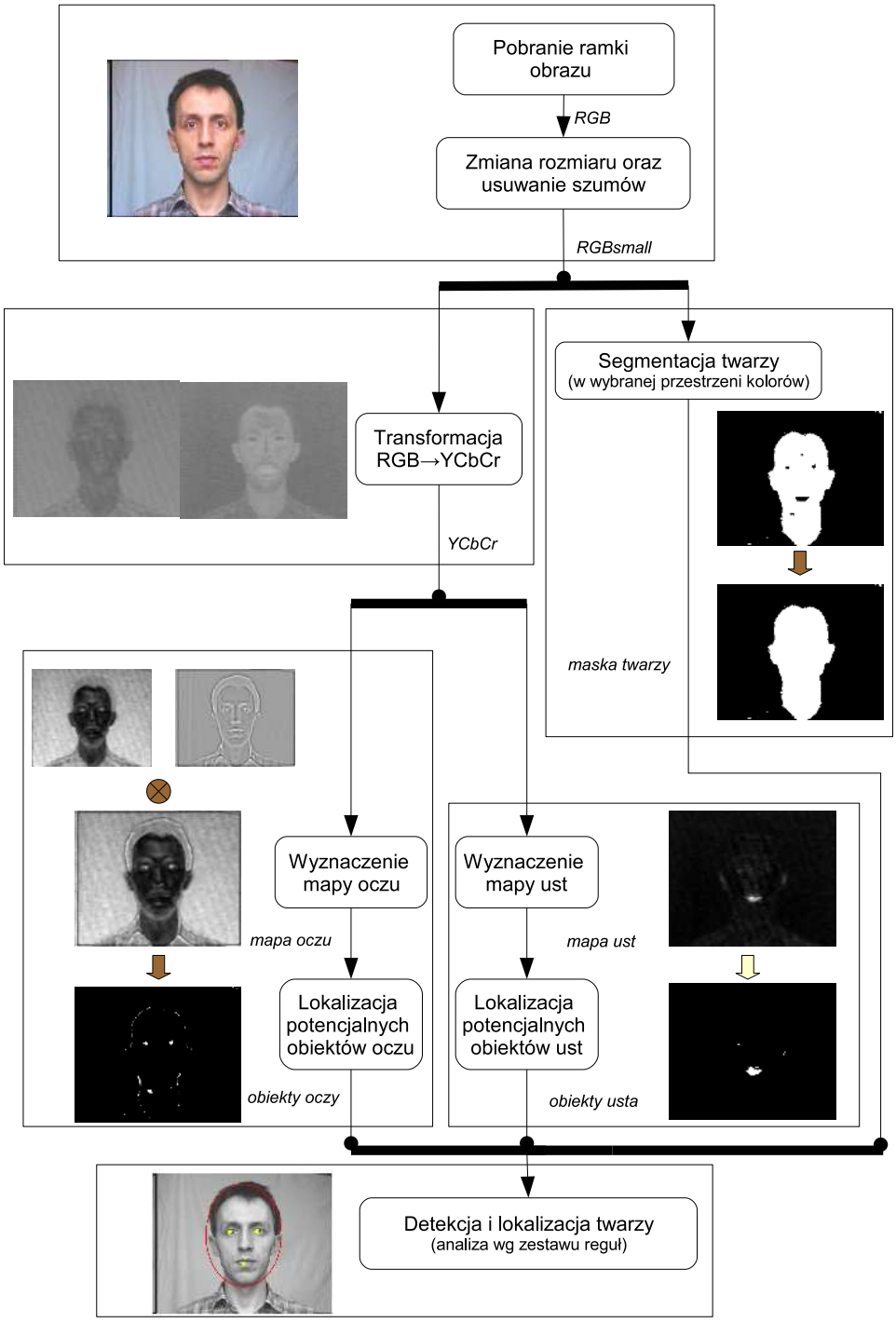
gdzie:

$$BW \in \{BW_{r,g}, BW_{hue}, BW_{cb,cr}\}$$

$\Gamma(BW)$ – operacja lub ciąg operacji morfologicznych na obrazie binarnym (w tym przypadku zalewanie otworów poprzedzone usunięciem niewielkich obiektów)

Do lokalizacji oczu oraz ust wykorzystano algorytm zaproponowany w publikacji [38]. Opiera się on na fakcie, iż obszary oczu i ust charakteryzują się innymi wartościami składowych chrominancji Cb oraz Cr. Na obrazie okolic oczu występują wysokie wartości Cb oraz niskie Cr, natomiast obraz ust zawiera inny stosunek składowych $Cr/Cb$ niż inne obszary twarzy. Na tej podstawie tworzone są mapy prawdopodobieństwa wystąpienia oczu i ust. Algorytm został przystosowany na potrzeby systemu rozpoznawania mimiki poprzez wprowadzenie szeregu zmian, m.in. wykorzystanie algorytmu detekcji plam (ang. blobs) oraz własny zestaw reguł na podstawie zależności geometrycznych elementów twarzy.

Obliczanie mapy prawdopodobieństwa oczu oraz ust zostało szerzej opisane w dodatku A.7. Z binarnych masek twarzy, oczu i ust ($BW_{face}$, $BW_{eye}$, $BW_{mouth}$), po indeksacji (ang. labelling), otrzymywane są obiekty będące potencjalnymi




Rysunek 5.9: Schemat blokowy algorytmu detekcji twarzy.

oczami i ustami. Na podstawie badań okazało się, że ilość znajdowanych elementów jest dość duża (rys. E.1). Dlatego konieczne okazało się opracowanie algorytmu selekcji i weryfikacji, który wybierałby z zestawu kandydatów jedną (najbardziej prawdopodobną) trójkę obiektów „oko-oko-usta". W tym celu wykorzystano głównie zależności geometryczne elementów twarzy. Podsumowuje to następująca lista kryteriów:

- obiekty muszą leżeć w obszarze maski twarzy (poddanej operacji erozji aby usunąć zakłócenia pojawiające się na krawędziach twarzy i tła),
- wielkość obiektów musi być większa niż założone minimum (usunięcie niewielkich obiektów składających się z kilku pikseli będących zakłóceniami),
- odległość lewego oka od ust musi być w przybliżeniu równa odległości prawego oka od ust (w rzeczywistości duży obrót głowy w lewo lub prawo powoduje iż zależność ta nie jest spełniona, jednakże w większości przypadków kryterium to jest spełnione),
- odległość pomiędzy oczami musi być proporcjonalna do średniej odległości oczu od ust (czyli oczy nie mogą być położone zbyt blisko siebie),
- oczy muszą leżeć powyżej ust (zakłada się że niedopuszczalna jest sytuacja odwrócenia kamery do góry nogami).

Powyższe kryteria pozwalają na selekcję oczu i ust. W przypadku gdy dalej pozostanie wiele kandydatów, jako właściwy wybierana jest trójka obiektów, których suma ilości pikseli jest największa. Wyznaczenie położenia oraz wielkości twarzy na podstawie lokalizacji oczu i ust jest już nieskomplikowane.

### 5.3.2 Rezultaty detekcji i lokalizacji twarzy

Skuteczność algorytmu detekcji twarzy określono dla kilku sekwencji wideo:

- sekwencja 1 — kamera EVI, optymalne warunki akwizycji (jednolite tło, natężenie oświetlenia około 320lux), (rys 5.7a),
- sekwencja 2 — kamera EVI, słabe oświetlenie (jednolite tło, natężenie oświetlenia około 40lux), (5.7b),
- sekwencja 3 — kamera USB Creative WebCam, skomplikowana scena i warunki akwizycji (tzn. skomplikowane tło, boczne oświetlenie, okulary), (5.7c).

Jako poprawnie wykrytą i zlokalizowaną twarz przyjęto sytuację gdy zostały spełnione założone kryteria selekcji oczu i ust oraz błąd lokalizacji cech był

mniejszy niż zadany próg (10 pikseli). Na podstawie analizy wyników segmentacji przeprowadzonej w poprzednim rozdziale (por. 5.2.2), odpowiednio dobrano najlepszą metodę oraz przestrzeń kolorów. Była nią segmentacja oparta o zestaw reguł (dwuprogowa binaryzacja składowych) w przestrzeni Cb-Cr.

Przy większych ruchach głowy, może się zdarzyć że niektóre elementy (szczególnie oczy) nie są widoczne. Ponieważ są one niezbędne do prawidłowego działania algorytmu, ich brak uniemożliwia prawidłową detekcję. Z tego powodu badanie skuteczności algorytmu przeprowadzono dla wybranych z sekwencji obrazów, odpowiadających położeniu neutralnemu (twarz frontalnie do kamery) — rys (5.10).

Rysunek 5.10: Skrajne położenia głowy dla twarzy uznanej za frontalną.

Rezultaty detekcji i lokalizacji twarzy dla powyższych sekwencji przedstawiono w dodatku E, tabele: E.1, E.2. Na podstawie wyników można stwierdzić, że w przypadku gdy twarz usytuowana jest względem kamery w przybliżeniu frontalnie, detekcja i lokalizacja jest wystarczająca na potrzeby systemu rozpoznawania mimiki ($> 80\%$). Podczas testów nie uwzględniono również obrazów sekwencji dla których występuje gest mrugania. W takim przypadku poprawna lokalizacja oczu jest trudna, ponieważ cechy je reprezentujące nie są dobrze widoczne na obrazie.

Podsumowując zagadnienia przedstawione w rozdziale — można przyjąć następujące kierunki dalszych prac, które pozwolą na zwiększenie skuteczności algorytmu detekcji twarzy:

- wykorzystanie informacji o ruchu na obrazie do wykrycia sytuacji mrugania, a następnie do lokalizacji oczu,

- użycie współczynników kształtu do selekcji cech (zamiast kryterium wielkości obiektów),
- znalezienie innych kryteriów odległości cech dla położenia twarzy innego niż pozycja neutralna,
- wykorzystanie większej ilości elementów twarzy (np. brwi, nozdrza), widocznych również w przypadku położenia głowy innego niż frontalne,
- użycie algorytmów śledzenia cech (ang. feature tracking).

# Rozdział 6

# Wyodrębnianie z obrazu twarzy elementów mimiki

**Streszczenie**

Zdefiniowane w pierwszej części rozprawy atrybuty elementów mimiki oraz odpowiadające im cechy oczekiwane na obrazie twarzy, stanowią podstawę umożliwiającą rozpoznanie wybranych gestów mimicznych. W pierwszej kolejności odpowiednie cechy (kształt, wygląd...) muszą zostać wyodrębnione z obrazu. W rozdziale przedstawiono wybrane metody i algorytmy pozwalające na ekstrakcję z obrazu twarzy wymaganych na etapie rozpoznawania wektorów cech, reprezentujących elementy mimiki. Wybrane metody to: statystyczne modele kształtu, histogramy orientacji oraz detekcja ruchu.

## 6.1 Wstęp

W niniejszej rozprawie dokonano wyboru trzech metod reprezentujących różne podejścia do wyodrębniania z obrazu cech odpowiadających elementom mimiki:

- Statystyczne modele kształtu (ang. point distributed models) umożliwiające rozpoznawanie kształtu jakie tworzą cechy charakterystyczne wybranych jednostek czynnościowych. Jako wybrane jednostki czynnościowe przyjęto elementy mimiki górnej części twarzy: AU1+2 (unoszenie brwi) oraz AU4 (marszczenie brwi). U większości osób niezależne wywołanie ruchu AU1 lub AU2 jest niemożliwe bez długotrwałego treningu dlatego zrezygnowano z rozpoznawania osobno tych gestów.

- Histogramy orientacji pozwalające na rozpoznawanie lokalnych zmian wyglądu. W tym przypadku skupiono się na rozpoznawaniu mimiki dolnej

części twarzy, ponieważ występuje tam więcej zmarszczek oraz bruzd. Wybrane do rozpoznawania jednostki czynnościowe przyjęto na podstawie kryterium łatwości wykonania oraz ich przydatności do sterowania (patrz rozdział 3). Są to: AU12 (unoszenie kącików ust — obustronne i jednostronne), AU25+AU26 (otwarcie ust).

- Detekcja ruchu wykorzystana do segmentacji dynamicznych atrybutów elementów mimiki, co pozwala na rozpoznawanie gestów takich jak mrugnięcia, zmrużenia lub przymknięcia oczu (AU43, AU45, AU46).

W dalszej części rozprawy, do opisu elementów mimiki, wykorzystano głównie metodykę FACS. Nie oznacza to jednak że informacje z metodyki FAP nie są przydatne — obydwa te sposoby opisu mimiki są w pewien sposób komplementarne. Przykładowo — za odpowiedniki jednostek AU1+2 oraz AU4 w metodyce FAP, można przyjąć elementy FAP31-36. Określają one położenie punktów charakterystycznych brwi oraz sposób ich deformacji. Jest to istotne np. przy wyborze punktów kształtu.

Obszerną systematykę algorytmów wykorzystywanych w rozpoznawaniu mimiki, można znaleźć w pracy [26].

## 6.2 Statystyczne modele kształtu

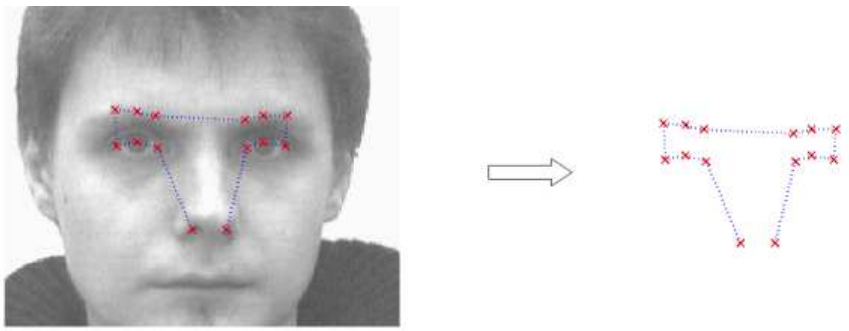
Tworzenie modelu opisującego zmiany kształtu cech twarzy pod wpływem wybranych jednostek czynnościowych (AU1+2, AU4) wymaga określenia szeregu punktów charakterystycznych definiujących analizowane w dalszych etapach kształty. Przy ustalaniu położenia punktów kierowano się kilkoma następującymi kryteriami:

- Punkty charakterystyczne powinny być łatwe do zlokalizowania na dwuwymiarowym statycznym obrazie. Oznacza to, że powinny być położone w obrębie wyraźnych i dobrze odróżnialnych elementów twarzy. Przykładem punktu stwarzającego problem podczas lokalizacji jest wierzchołek nosa, dobrze określony anatomicznie na rzeczywistej twarzy, lecz trudny do odróżnienia od jego otoczenia na zarejestrowanym obrazie.

- Punkty charakterystyczne powinny pozostawać widoczne i rozróżnialne pomimo wykonywanych ruchów mimicznych, przynajmniej w akceptowalnym ich zakresie. Punkty nie spełniające tego kryterium są na przykład położone w obrębie bruzdy powiekowej górnej, która przy wyraźniejszym zmarszczeniu brwi zostaje zasłonięta.

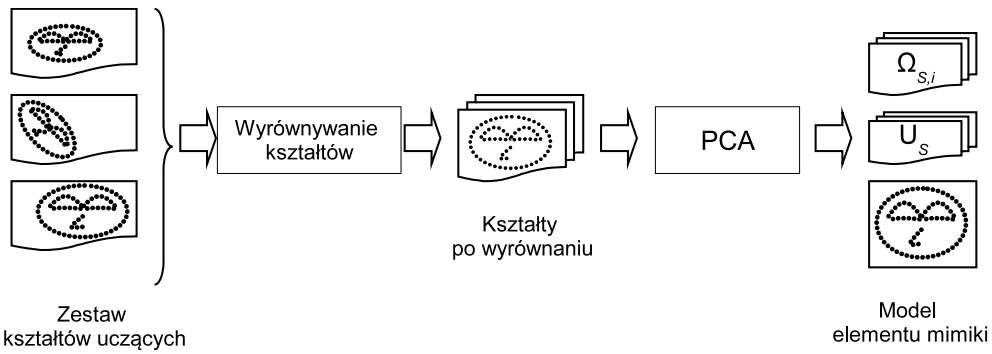
- Punkty charakterystyczne muszą reprezentować trwałe i powtarzalne cechy obrazu. Wyklucza to punkty zlokalizowane w zmiennych szczegółach anatomicznych, na przykład bruzdach, zmarszczkach czy liniach owłosienia.

Biorąc pod uwagę powyższe kryteria ustalono łącznie 14 punktów charakterystycznych, po 7 dla obu stron obserwowanej twarzy. Po każdej stronie wyznaczono 3 punkty położone w obrębie brwi oraz 4 dodatkowe punkty odniesienia. Lokalizacje punktów oraz tworzony przez nie kształt przedstawia rysunek 6.1. Punkty kształtów (oprócz dwóch punktów — nozdrza) odpowiadają lokalizacji zdefiniowanej w standardzie FAP — rys. 2.1, rys. 2.2. Szczegółowe informacje na temat doboru punktów oraz ich lokalizacji znajdują się w artykule autora [71].



Rysunek 6.1: Lokalizacja punktów kształtu jednostek czynnościowych górnej części twarzy (AU1+2, AU4)

Proces tworzenia modelu jednostki czynnościowej jest następujący (rys. 6.2). Na całej serii obrazów uczących, zawierających kształty danej klasy (np. AU0), zaznaczane są manualnie charakterystyczne punkty określające położenie i wzajemną relację wybranych elementów morfologicznych twarzy. Aby możliwe było wyznaczenie parametrów statystycznych, przeprowadzane jest wyrównywanie kształtów (ang. shape alignment), wg algorytmu przedstawionego w artykule [80].




Rysunek 6.2: Tworzenie statystycznego modelu jednostki czynnościowej

Następnie, przy pomocy metody PCA (ang. Principal Component Analysis), tworzony jest statystyczny model, opisujący średni kształt oraz jego poszczególne

deformacje. Metoda PCA została opisana szerzej w dodatku A.4.

W przypadku modeli kształtu składowe główne $U_S$ określają nową przestrzeń, w której każdy kształt (obiekt) reprezentowany jest przez wektor wag $\Omega_{S,i}$. Wektor wag określa stopień deformacji kształtu. Tylko część składowych głównych niesie istotne informacje, pozostałe mogą zostać pominięte, dzięki czemu uzyskuje się redukcję informacji. Ilość istotnych składowych głównych $K_S$ została określona na podstawie kryterium (6.1):

$$\sum_{i=1}^{K_S} \lambda_i \geqslant \frac{p_S}{100} \cdot \sum_{i=1}^{N} \lambda_i \tag{6.1}$$

gdzie:

$i$ – indeks kolejnego kształtu

$\lambda_i$ – wartość własna macierzy kowariancji danych kształtów zestawu uczącego

$N$ – ilość punktów kształtu

$p_S$ – zadana, procentowa wartość określająca proporcję wariancji danych z zestawu uczącego

Przy założeniu odpowiedniej ilości obrazów uczących, na podstawie parametrów tak utworzonego modelu, możliwe jest zsyntetyzowanie dowolnej instancji kształtu obiektu. Odbywa się to poprzez złożenie średniego kształtu oraz liniowej kombinacji poszczególnych deformacji (równanie 6.2).

$$S_i = \bar{S} + U_S \cdot \Omega_{S,i} \tag{6.2}$$

gdzie:

$S_i$ – wektor współrzędnych punktów i-go kształtu

$\bar{S}$ – wektor współrzędnych punktów kształtu średniego

$U_S$ – składowe główne modelu

$\Omega_{S,i}$ – wektor wag i-go kształtu

W przypadku nowego (nieznanego) kształtu, sposób postępowania jest następujący:

1. Nieznany kształt wyrównywany jest względem kształtu średniego otrzymanego podczas budowy modelu

2. Wyrównany kształt jest następnie rzutowany do przestrzeni kształtów modelu (równanie 6.3).

3. Z otrzymanego wektora wag wybierane jest $K_S$ elementów tworząc wektor cech, poddawany klasyfikacji przy pomocy jednej z dostępnych metod.

$$\Omega_{S,tst} = U_S^T \cdot S_{tst} \tag{6.3}$$

gdzie:

$\Omega_{S,tst}$ – wektor wag testowego kształtu
$\bar{S}$ – wektor współrzędnych punktów kształtu średniego
$U_S$ – składowe główne modelu
$S_{tst}$ – wektor współrzędnych punktów testowego kształtu (po wyrównaniu)

Z perspektywy budowy interfejsu, który powinien adaptować się do użytkownika, wyznaczanie położenia punktów charakterystycznych, powinno odbywać się automatycznie. Na potrzeby niniejszej rozprawy punkty wyznaczono w sposób ręczny. Problematyka automatyzacji tego procesu została krótko opisana w rozdziale 8.4.

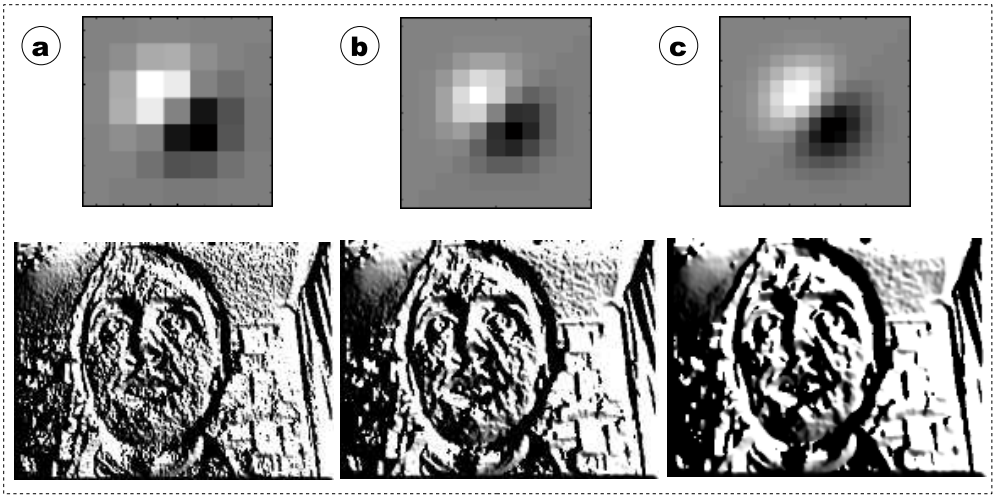
## 6.3 Histogramy orientacji

Histogramy orientacji pozwalają na opisanie lokalnej orientacji krawędzi obrazu. Mogą być wyznaczane na wiele różnych sposobów. W niniejszym rozdziale opisano dwa z nich, różniące się sposobem detekcji krawędzi. Pierwszy z nich, wzorowany na rozwiązaniu stosowanym w pracy [28], realizowany jest następująco:

- dla całego obrazu wyznaczana jest jego reprezentacja w przestrzeni skali (ang. scale-space) $L : \Re^N \times \Re_+ \rightarrow \Re$ (metoda ta została opisana szerzej w dodatku A.5).
- następnie obliczane są gradienty poziome oraz pionowe $L_x, L_y$
- na ich podstawie wyznaczany jest moduł oraz orientacja gradientu dla każdego piksela (6.4)
- dla wybranego regionu zainteresowań obliczany jest histogram orientacji krawędzi $h_{edge}$, poprzez zliczanie pikseli o danej orientacji $L_\theta$.

$$\begin{aligned} L_m &= \sqrt{L_x^2 + L_y^2} \\ L_\theta &= \operatorname{arc\,tg}(L_x, L_y) \end{aligned} \tag{6.4}$$

Rysunek 6.3 przedstawia rezultaty wyznaczania reprezentacji skali dla przykładowych filtrów. Do zalet tego typu histogramów orientacji można zaliczyć: szybkość obliczeń, małą wrażliwość na zmiany oświetlenia sceny, niezależność od translacji i skali na obrazie.

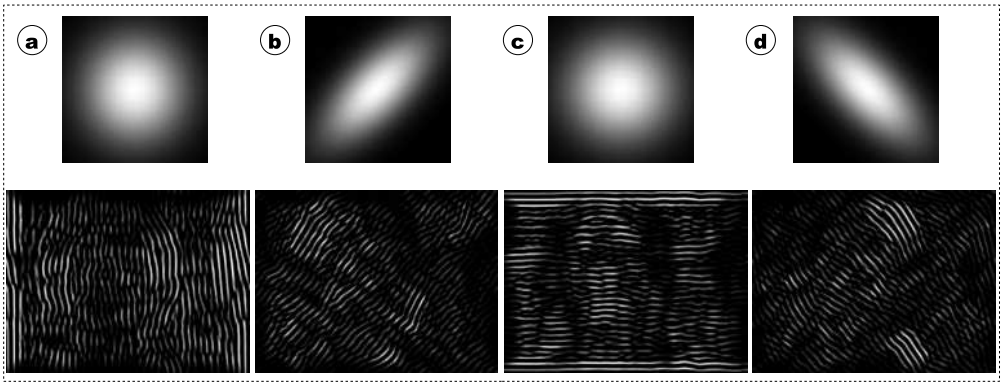



Rysunek 6.3: Maski filtrów przestrzeni skali i odpowiadające im obrazy po filtracji dla różnych skal: (a) $\sigma = 1$,(b) $\sigma = 2$,(c) $\sigma = 4$.

Drugi sposób wyznaczania histogramów orientacji wykorzystuje reprezentację obrazu powstałą przez konwolucję z zestawem filtrów Gabora. Umożliwiają one detekcję cech w różnych skalach i orientacjach [89]. Motywacją do ich użycia jest teoretyczna niewrażliwość cech wyznaczonych w ten sposób od globalnych zmian oświetlenia sceny, połączona z małą wrażliwością na deformacje afiniczne obrazu. Dodatkową zaletą jest to, iż dobrze zachowane są informacje o lokalnej geometrii i teksturze obrazu. Filtry Gabora zostały opisane szerzej w dodatku A.6. W celu wydobycia z obrazu interesujących informacji o obiektach, niezależnie od ich skali i orientacji, tworzony jest zestaw filtrów (tzw. bank filtrów). Zestaw ten jest określany dla różnych parametrów — z reguły są to częstotliwość oraz kąt. Więcej informacji na temat zastosowania filtrów Gabora w przetwarzaniu obrazów oraz doboru ich parametrów, można znaleźć w pracy [39]. Algorytm wyznaczania histogramów orientacji jest następujący:

- dla całego obrazu wyznaczana jest jego reprezentacja dla wybranego banku filtrów Gabora (kilka częstotliwości i kątów): $L_{\psi_G}(i)$, i - numer filtru
- w wybranym regionie zainteresowań obliczane są histogramy dla każdej reprezentacji - $h_{gabor}$,
- następnie histogramy łączone są (kolejno) w jeden wektor cech reprezentujący dany obszar zainteresowań.

Rysunek 6.4 przedstawia rezultaty wyznaczania reprezentacji Gabora dla przykładowych filtrów.




Rysunek 6.4: Maski filtrów Gabora i odpowiadające im obrazy po filtracji. Parametry: $f_G = 0.125$, (a) $\theta_G = 0°$,(b) $\theta_G = 45°$,(c) $\theta_G = 90°$,(d) $\theta_G = 135°$

## 6.4 Detekcja ruchu

Istnieje wiele metod detekcji ruchu [52]. Metody podstawowe umożliwiają uzyskanie informacji o zaistnieniu zmian na scenie oraz ich lokalizacji. Zaliczyć do nich można: odejmowanie obrazu od tła, obrazy różnicowe, analizę histogramów sekwencji, testy statystyczne, potok optyczny (ang. optical flow). Jednym z prostszych sposobów detekcji zmian na obrazie jest odejmowanie obrazu tła (ang. background subtraction). Ponieważ oprócz poruszających się obiektów mogą wystąpić zmiany oświetlenia sceny, sposób ten jest mało skuteczny bez zastosowania metod automatycznej generacji tła [79]. Algorytmy generacji tła poprawiają w widoczny sposób rezultaty detekcji. W przypadku jednak obserwacji pracującego przed komputerem człowieka, który nie tylko wykonuje ruchy głową ale też często pozostaje nieruchomo, mogą wystąpić trudności w prawidłowym doborze parametrów szybkości uaktualniania tła (zbyt szybkie uaktualnianie spowoduje iż twarz człowieka zostanie uwzględniona jako tło). Innym sposobem detekcji zmian na obrazie jest obliczanie różnicy jasności piksela w kolejnych chwilach czasowych (ang. image difference). Z matematycznego punktu widzenia odpowiada to wyznaczaniu pochodnej cząstkowej funkcji jasności względem czasu $\frac{\partial I(x,y,t)}{\partial t}$. Ponieważ w przypadku przetwarzania obrazów sygnał jest dyskretny, w praktyce wyznaczaniu pochodnej czasowej odpowiada zwykłe odejmowanie wartości pikseli obrazu aktualnego od obrazu poprzedniego: $I(x, y, k) - I(x, y, k–1)$. Często stosuje się również obliczanie wartości bezwzględnej dla różnicy.

Bardziej złożone metody pozwalają na rozpoznanie obiektów ruchomych oraz ich śledzenie, estymację parametrów ruchu (translacja, orientacja, skala). Jedną

z ciekawszych metod są tzw. szablony ruchu (ang. temporal templates) . Wśród nich wyróżnić można:

- obrazy historii ruchu MHI (ang. motion history images),
- obrazy energii ruchu MEI (ang. motion energy images).

Obydwie techniki wykorzystują analizę historii zmian intensywności pikseli w czasie do określenia obszarów, w których występuje ruch. W pracy [10] zastosowano powyższe metody jako wstępny etap rozpoznawania akcji wykonywanych przez człowieka (siadanie, ćwiczenia aerobik). Interesującym zastosowaniem rozpoznawania akcji jest użycie metody do analizy zachowania dzieci w interaktywnym pokoju zabaw [9]. Wyczerpujące omówienie komputerowej analizy ruchu można znaleźć w pracy Kuriańskiego [52].

Dużym problemem podczas detekcji ruchu, są zmiany oświetlenia sceny, które mogą zostać zinterpretowane błędnie jako ruch. Zmiany oświetlenia mogą być wprowadzane nie tylko przez np. zaświecenie światła w pomieszczeniu, ale również mogą pochodzić od zmian jasności monitora przed którym znajduje się użytkownik systemu. Jednym ze sposobów ominięcia tego problemu, jest wykrywanie globalnych zmian jasności sceny. Z kolei zakłócenia powodowane przez monitor komputera mogą zostać wykorzystane do wstępnej segmentacji twarzy, która z reguły jest obiektem pierwszoplanowym — najbliższym monitora i kamery obserwującej scenę. Wykorzystanie tego efektu opisano w rozdziale 8.3.

W systemie rozpoznawania mimiki detekcja ruchu może być wykorzystana do wstępnej segmentacji obszarów zainteresowań takich jak:

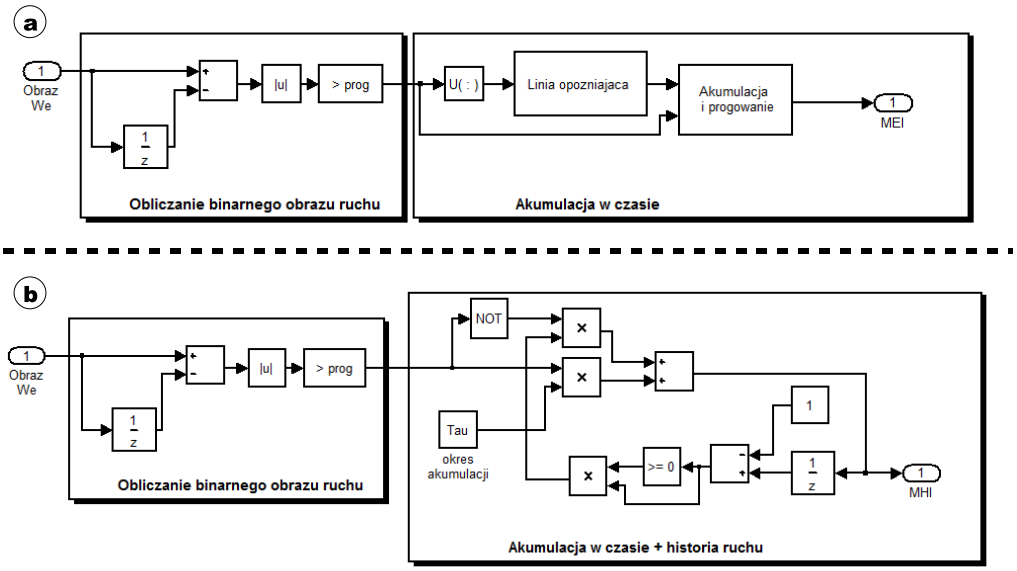
- obręb oczu — mruganie, ruchy oczu oraz jednostki mimiki wpływające na tą część twarzy,
- obszar głowy — ruchy głowy powodują zazwyczaj duże zmiany intensywności pikseli obrazu,
- miejsca w których występują zmiany wyglądu wywołane ruchami mimicznymi.

Lokalizacja powyższych zmian umożliwia nie tylko zawężenie obszaru poszukiwań do odpowiedniego regionu twarzy, ale również uzyskanie bezpośrednich wskazówek użytecznych do rozpoznania niektórych gestów mimicznych (np. mruganie). Przykładem może być praca [56], której autorzy wykorzystali detekcję mrugnięć do inicjalizacji algorytmu rozpoznawania ruchu brwi.

W badaniach opisanych w niniejszej rozprawie wykorzystano obrazy historii ruchu MHI oraz energii ruchu MEI. W odróżnieniu od prostej różnicy obrazów, pozwalają one na uwzględnienie kontekstu historii zmian intensywności piksela.

Daje to w efekcie większą ilość użytecznych do dalszej analizy danych — nie tylko informacje gdzie wystąpił ruch, ale również jego charakterystykę.

Rysunek 6.5 przedstawia schemat blokowy algorytmu wyznaczenia MHI oraz MEI.




Rysunek 6.5: Model algorytmu (Simulink) obliczania obrazów: (a) energii ruchu MEI, (b) historii ruchu MHI.

Pierwszym krokiem jest obliczenie binarnego obrazu ruchu — piksele gdzie wystąpił ruch przyjmują wartość 1. Dla każdego piksela obrazu wyznaczany jest moduł różnicy kolejnych wartości w czasie, a otrzymana tablica wartości poddawana jest operacji binaryzacji z progiem $M_{tresh}$przyjętym doświadczalnie (6.5).

$$I_{diff}(x,y,k) = \begin{cases} 0; & |I(x,y,k) - I(x,y,k-1)| \leqslant M_{tresh} \\ 1; & |I(x,y,k) - I(x,y,k-1)| > M_{tresh} \end{cases} \tag{6.5}$$

gdzie:

$x, y$ – koordynaty pikseli

$k$ – numer obrazu z sekwencji video

$I(x,y,k)$ – jasność piksela obrazu o współrzędnych $x, y$ w chwili $k$

$M_{tresh}$ – próg binaryzacji dla wyznaczania obrazów energii ruchu

$I_{diff}(x,y,k)$ – moduł różnicy piskeli w chwili $k$

Obraz MEI powstaje poprzez akumulowanie w czasie binarnych obrazów ruchu (6.6).

$$MEI_{\tau}(x,y,k) = \bigcup_{i=0}^{\tau-1} BW_{diff}(x,y,k-1) \tag{6.6}$$

gdzie:

$\tau$ – wartość okresu akumulacji dla obrazów energii i historii ruchu
$BW_{diff}(x,y,k)$ – binarny obraz ruchu dla piksela o współrzędnych $x, y$ w chwili $k$
$MEI_{\tau}(x,y,k)$ – obraz energii ruchu dla piksela o współrzędnych $x, y$ w chwili $k$

Natomiast reprezentacja MHI powstaje poprzez przyjęcie wartości okresu akumulacji $\tau$ dla pikseli binarnego obrazu ruchu $BW_{diff}$ równych 1, i sukcesywne zmniejszanie dla kolejnych obrazów sekwencji video tej wartości, jeśli piksel $BW_{diff}$ będzie miał wartość 0 (6.7):

$$MHI_{\tau}(x,y,k) = \left\{ \begin{array}{cc} \tau; & BW_{diff}(x,y,k)=1 \\ \max\left[0, MHI_{\tau}(x,y,k-1)\right]; & BW_{diff}(x,y,k)=0 \end{array} \right\} \tag{6.7}$$

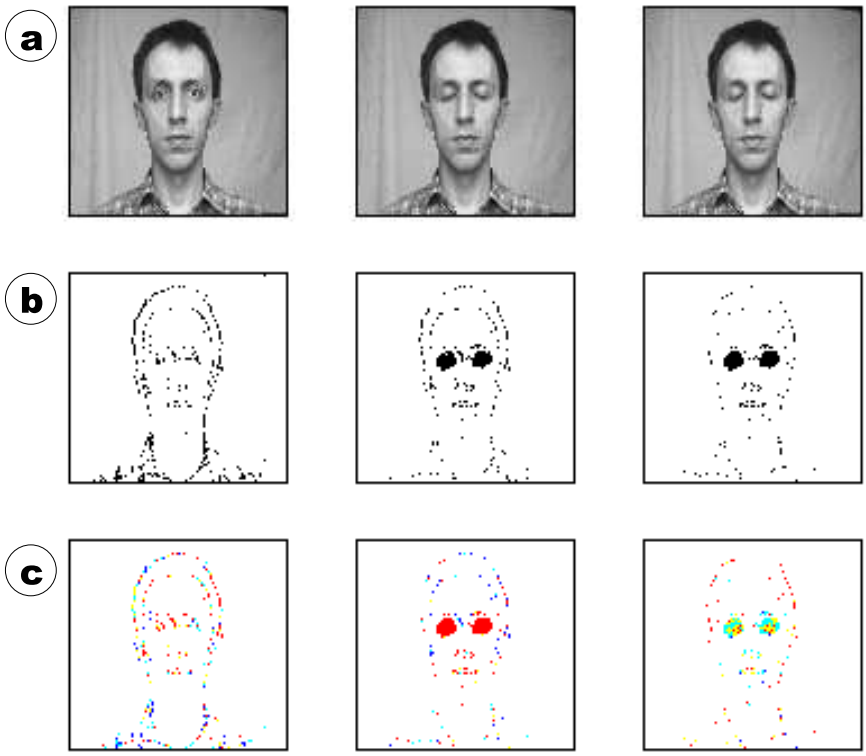
gdzie:

$MHI_{\tau}(x,y,k)$ – obraz historii ruchu dla piksela o współrzędnych $x, y$ w chwili $k$

Wartość okresu akumulacji $\tau$ wpływa na czasowy kontekst wykonywanego ruchu mimicznego i została dobrana tak, aby odpowiednio wykryć obszary w których występuje gest mimiczny lub ruch głowy.

Na rysunku 6.6 zamieszczone zostały rezultaty wyznaczania reprezentacji MEI, MHI dla sekwencji ramek obrazu zawierającej akcję mrugania oczu.

W przypadku gdy nie występuje ruch głowy, reprezentacja MEI może zostać użyta do zidentyfikowania obszarów oczu podczas wykonywania np. gestów mrugania (AU45, AU46). Również reprezentacja MHI niesie informacje przydatne do identyfikacji gestu. Obrazy MEI pozwalają stwierdzić „gdzie" wystąpił ruch, natomiast MHI udostępnia informacje o charakterystyce ruchu. Piksele MHI o dużej wartości odpowiadają miejscom, w którym aktualnie występują zmiany jasności, natomiast wartości niewielkie sygnalizują że ruch w tym miejscu odbywał się wcześniej w czasie. Przy pomocy parametru $\tau$ (okres akumulacji) możliwy jest wpływ na detekcję obszarów charakteryzujących się założoną szybkością wykonywanego ruchu. Przykładowo — gdy celem jest wykrycie szybkich ruchów (odróżnienie zamykania oczu od ich otwierania), parametr $\tau$ powinien mieć niewielką wartość. W przeciwnym przypadku szybkie zmiany mimiki zostaną zinterpretowane jak jeden ciągły ruch.




Rysunek 6.6: Kolejne obrazy sekwencji zawierającej gest mrugania: (a) obrazy z kamery EVI, (b) reprezentacja MEI, (c) reprezentacja MHI.

Na rysunku 6.6, oprócz miejsc w których występuje gest mimiczny, widoczne są również zakłócenia. Powstają one w wyniku istnienia szumów oraz mimowolnych poruszeń głowy (widoczny zarys głowy). Szumy występujące na obrazie mogą być usunięte w znacznym stopniu poprzez zastosowanie wstępnej filtracji medianowej, która skutecznie zmniejsza lokalny szum, równocześnie pozostawiając nienaruszoną główną składową ruchu (lub zmieniając ją w pomijalnym stopniu). Problem estymacji i usuwania szumów z sekwencji video został szerzej opisany w dodatku A.3.

Z kolei usunięcie zakłóceń od mimowolnych ruchów głowy, może być zrealizowane poprzez wybór tych obiektów, które spełniają założone dla danego gestu lub ruchu własności (np. kształt, orientacja, wzajemne położenie, czas trwania). W tym celu wykorzystane mogą być współczynniki kształtu. Zostało to szerzej opisane w rozdziale 8.4.1.

# Rozdział 7

# Rozpoznawanie wybranych elementów mimiki

**Streszczenie**

W zadaniu rozpoznawania celem jest określenie przynależności różnego typu obiektów (w tym przypadku elementów mimiki) do pewnych klas na podstawie znanych wcześniej przynależności do klas innych obiektów. Jest to zatem typowe zadanie klasyfikacji, w którym obiekty opisane są przy pomocy zestawu atrybutów/cech. W niniejszym rozdziale przedstawiono metodykę oraz rezultaty rozpoznawania wybranych elementów mimiki, wykorzystując dane będące wynikiem etapu wyodrębniania cech. Na podstawie analizy wyników wybrano najlepszą metodę rozpoznawania oraz zaproponowano dalsze kierunki prac.

## 7.1 Wstęp

Uzyskane na etapie ekstrakcji cech charakterystycznych informacje, pozwalają na opisanie elementów mimiki poprzez zbiór wartości (cechy) i na tym zbiorze można dokonywać obliczeń w celu podjęcia decyzji o przynależności obiektu do określonej klasy. Istnieje wiele metod klasyfikacji obrazów — ich szerszą systematykę i matematyczne podstawy można znaleźć np. w pracy [83]. Wśród najpopularniejszych algorytmów, wykorzystywanych w rozpoznawaniu obrazów, wymienić można:

- Metody minimalno-odległościowe (np. klasyfikator kNN), opierające się na przesłankach związanych z geometrią przestrzeni cech i wykorzystujący różne metryki [92].

- Metody probabilistyczne (np. naiwny klasyfikator bayesowski), pozwalające przewidzieć prawdopodobieństwo przynależności obiektu do klasy w oparciu o twierdzenie Bayesa [75].

- Statystyczna analiza dyskryminacyjna — pozwalająca na klasyfikację obiektów do dwóch lub większej liczby klas przy pomocy funkcji dyskryminujących, utworzonych w oparciu o zbiór uczący[98].

- Maszyny wektorów wspierających[1] (SVM - ang. suport vector machines) [17][42]. Metoda ta poszukuje hiperpłaszczyzny w wielowymiarowej przestrzeni, która rozdziela przykłady ze zbioru treningowego, rzutowane do tej przestrzeni przez odpowiednią funkcję zwaną jądrem.

- Sieci neuronowe pozwalające na przetwarzanie informacji w sposób równoległy, analogicznie do ludzkiego mózgu. Jednym z zastosowań jest klasyfikacja wzorców [82].

W literaturze dotyczącej problematyki rozpoznawania formułuje się również komplementarne w stosunku do klasyfikacji — zadanie grupowania (ang. clustering). W odróżnieniu od klasyfikacji, w której celem jest określenie przynależności pojedynczego obiektu do znanego zbioru klas, zadanie klasteryzacji polega na automatycznym wyznaczeniu ilości oraz charakterystyki klas na podstawie danej zbiorowości obiektów. W kontekście systemu rozpoznawania mimiki, to drugie podejście jest również istotne, ponieważ nie wymaga etapu uczenia, który zazwyczaj jest realizowany z udziałem człowieka („ręczne" przyporządkowanie elementów zbioru uczącego do przyjętych klas). Wśród metod grupowania można wymienić:

- Metody grupowania hierarchicznego (ang. hierarchical clustering), które generują sekwencję podziałów zbiorów obiektów w procesie grupowania.

- Metoda k-średnich (ang. k-means) należąca do grupy algorytmów iteracyjno-optymalizacyjnych [90].

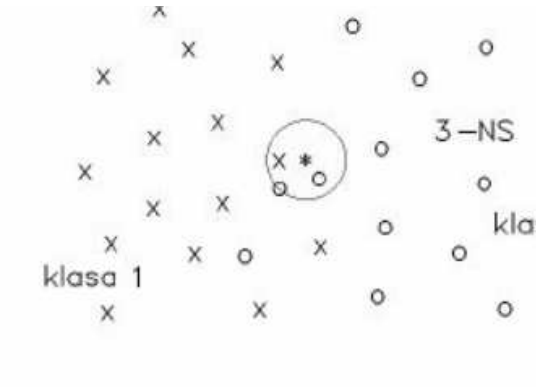
## 7.2 Opis wybranych metod klasyfikacji

W niniejszej pracy, do rozpoznawania elementów mimiki, wykorzystano dwie metody klasyfikacji — klasyfikator kNN oraz wielowymiarową analizę dyskryminacyjną.

Metoda kNN (k najbliższych sąsiadów) polega na tym, że klasyfikowany obiekt jest zaliczany do klasy najliczniej reprezentowanej wśród jego k „najbliższych sąsiadów". Jeżeli w tej samej odległości, co k-ty „sąsiad" znajdą się jeszcze inne elementy, to wszyscy ci „sąsiedzi" biorą udział w głosowaniu. Działanie tej reguły

[1] spotykana jest również nazwa „metoda wektorów nośnych"

dla k=3 i sztucznego małego dwuwymiarowego zbioru odniesienia zilustrowano na rysunku (rys. 7.1). Zaletą tej metody jest jej prosta implementacja, możliwość zrównoleglenia obliczeń, duża skuteczność. Wadą — konieczność zapamiętania całego zbioru uczącego.




Rysunek 7.1: Ilustracja reguły k-NN dla k=3. Źródło: [43]

Wielowymiarowa analiza dyskryminacyjna opiera się na podziale zbioru danych uczących na obszary ograniczone funkcjami dyskryminującymi, które najlepiej rozróżniają dwie lub więcej klas obiektów. Funkcje te mogą być liniowe (liniowa analiza dyskryminacyjna), kwadratowe (kwadratowa analiza dyskryminacyjna), itp. Mogą one być następnie wykorzystane do klasyfikacji nowych danych do poszczególnych klas na podstawie prostego kryterium wyboru (rys. 7.2) — przypisz wektor danych $\mathbf{X}$ do klasy $\omega_i$ jeśli $df_i(\mathbf{X}) > df_j(\mathbf{X}) \quad \forall j \neq i$

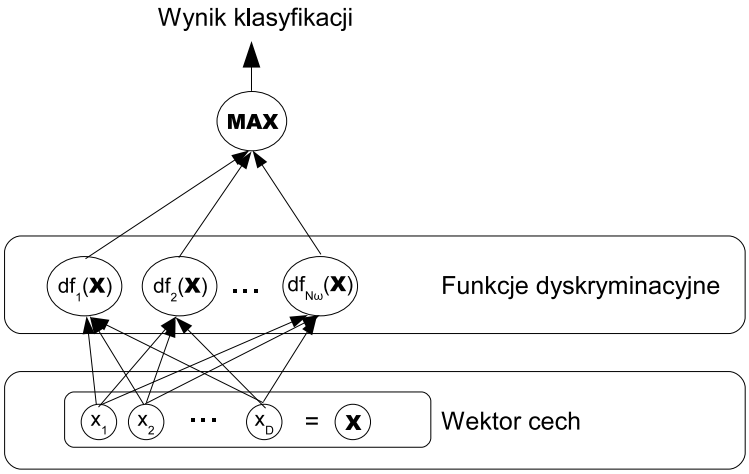
Optymalne parametry funkcji dyskryminacyjnych wyznaczane są na podstawie zbioru obserwacji (dane uczące) poprzez minimalizację prawdopodobieństwa błędu z wykorzystaniem reguły maksimum prawdopodobieństwa a posteriori Bayesa (MAP) . W przypadku parametrycznej analizy dyskryminacyjnej przyjmuje się, że funkcje gęstości prawdopodobieństwa zbioru posiadanych obserwacji są opisane rozkładem normalnym. Reguła MAP prowadzi do następującego ogólnego wzoru (7.1):

$$df_i = -\frac{1}{2} \cdot (\mathbf{X} - \mu_{X_i})^T \cdot \Sigma_{X_i}^{-1} \cdot (\mathbf{X} - \mu_{X_i}) - \frac{1}{2} \cdot \log(|\Sigma_{X_i}|) + \log(P(\omega_i)) \quad (7.1)$$

gdzie:

$\mu_{X_i}, \Sigma_{X_i}$ – wartość oczekiwana oraz macierz kowariacji dla elementów zbioru danych uczących należących do i-tej klasy

$P(\omega_i)$ – prawdopodobieństwo apriori dla i-tej klasy




Rysunek 7.2: Ilustracja klasyfikacji przy pomocy analizy dyskryminacyjnej.

W przypadku gdy macierze kowariancji są takie same dla każdej z klas ($\Sigma_{X_i} = \Sigma_X$), jest to liniowa analiza dyskryminacyjna (LDA – ang. Linear Discriminant Analysis). Gdy macierze kowariancji estymowane są osobno dla każdej z grup — kwadratowa analiza dyskryminacyjna (QDA – Quadratic Discriminant Analysis). W porównaniu do metody kNN, analiza dyskryminacyjna wymaga pamiętania niewielu parametrów, a nie całego zbioru uczącego. Posiada jednak silne założenie normalności i unimodalności rozkładów zbioru elementów należących do poszczególnych klas. Więcej informacji na temat powyższych metod klasyfikacji można znaleźć w pracy [21].

## 7.3 Rozpoznawanie elementów mimiki — rezultaty i wnioski

Testy rozpoznawania elementów mimiki przeprowadzono niezależnie dla dwóch z opisanych wcześniej metod (tj. modeli kształtu oraz histogramów orientacji). W przypadku informacji uzyskanych z algorytmów detekcji ruchu, rozpoznawanie realizowane jest w odmienny sposób. Z tego powodu detekcja mrugnięć została odrębnie opisana w rozdziale 8.4.1. Przyjęto następującą ogólną procedurę testową:

1. Przygotowanie danych do rozpoznawania. Dla każdego obrazu sekwencji video:
   a) ekstrakcja cech charakterystycznych wg odpowiedniej metody (modele kształtu, histogramy orientacji...),
   b) utworzenie wektorów cech.

2. Podział danych na sekwencje uczące oraz testowe, standaryzacja danych (zerowa wartość oczekiwana oraz odchylenie standardowe).

3. Tworzenie modelu — uczenie klasyfikatora i dobór jego parametrów (na danych ze zbioru uczącego).

4. Ocena skuteczności klasyfikacji na danych ze zbioru testowego.

Etapy pierwszy i drugi, zostały omówione dokładniej w kolejnych podrozdziałach poświęconych poszczególnym metodom (por. 7.3.1, 7.3.2).

Uczenie i dobór parametrów klasyfikatora zrealizowano z wykorzystaniem jednej z technik testowania jakości klasyfikacji — algorytm K-krotnej walidacji krzyżowej (ang. K-fold crossvalidation) [49]. W metodzie tej oryginalna próba jest dzielona na K możliwie równych, wzajemnie niezależnych podzbiorów. Następnie kolejno każdy z nich brany jest jako zbiór testowy, a pozostałe razem jako zbiór uczący i dokonywana jest klasyfikacja. Klasyfikacja jest wykonywana K-krotnie, a rezultaty są następnie uśredniane w celu uzyskania jednego wyniku.

Jako K przyjęto 3 (ze względu na niewielką liczność poszczególnych klas zbioru uczącego). Aby zmniejszyć wariancję oceny jakości algorytm walidacji był powtarzany 10-krotnie. Ponadto każdy podzbiór danych uczących został podzielony losowo na część trenującą i walidującą w stosunku 70% do 30% (tzw. stratified sampling). Sumaryczny błąd klasyfikacji wyznaczony został jako średnia ze wszystkich klasyfikatorów i prób. Jako najlepszy klasyfikator został wybrany ten, który zapewniał największą dokładność klasyfikacji — przyjęte kryterium dokładności jest określone równaniem 7.2.

$$\arg\max_{i} \left( \sum_{j}^{N_\omega} PPV + NPV \right), \quad i = 1 \ldots Kfold \tag{7.2}$$

gdzie:

$PPV, NPV$ – wartość predykcyjna dodatnia i ujemna klasyfikacji (ang. positive predictive value, negative predictive value)

$Kfold$ – ilość podzbiorów dla walidacji krzyżowej K-fold pomnożona przez ilość powtórzeń walidacji

$N_\omega$ – ilość klas

Oceny skuteczności klasyfikacji dokonano na zbiorze testowym dla najlepszego klasyfikatora. Wyznaczono następujące parametry jakości klasyfikacji: wartość predykcyjna dodatnia, wartość predykcyjna ujemna, swoistość, czułość. Objaśnienia znajdują się w dodatku C.1.

### 7.3.1 Statystyczne modele kształtu

Do testów skuteczności działania przedstawionego algorytmu, wykorzystane zostały sekwencje obrazów zawierających wybrane jednostki czynnościowe górne części twarzy, wykonywane przez różne osoby:

- AU0 — położenie neutralne,
- AU1+2 — unoszenie brwi,
- AU4 — marszczenie brwi.

Jak wspomniano w rozdziale 6.1, za odpowiedniki jednostek AU1+2 oraz AU4 w metodyce FAP, można przyjąć elementy FAP31-36. Określają one położenie punktów charakterystycznych brwi oraz sposób ich deformacji.

Podczas manualnego wyznaczania punktów charakterystycznych kształtów ważne jest zachowanie powtarzalności umieszczania punktów na kolejnych obrazach. W celu ułatwienia tego procesu utworzona została aplikacja pomocnicza pozwalająca na: edycję punktów kształtu, obserwację profili obrazu w wybranych miejscach oraz tworzenie zestawu kształtów [71].

Przygotowane w ten sposób dane kształtów, podzielono następnie na zbiory uczące oraz testowe, użyte w kolejnych testach:

- Test 1 — sprawdzenie skuteczności rozróżniania jednostki AU0 od AU1+2 dla danych jednej osoby.
- Test 2 — sprawdzenie skuteczności rozróżniania jednostki AU0 od AU1+2 dla danych kilku osób.
- Test 3 — sprawdzenie skuteczności rozróżniania jednostek AU0, AU1+2 oraz AU4 dla danych jednej osoby.
- Test 4 — sprawdzenie skuteczności rozróżniania jednostek AU0, AU1+2 oraz AU4 dla danych kilku osób.

Parametry klasyfikatora dobrano z wykorzystaniem kryterium dokładności (7.2) dla sekwencji uczącej. Są one następujące:

- dla klasyfikatora kNN: metryka euklidesowa, K=2,
- rodzaj analizy dyskryminacyjnej: liniowa analiza dyskryminacyjna LDA.

Wyniki testów przedstawiono w dodatku C, tabele: C.15 — C.22, wykresy: C.12 — C.19.

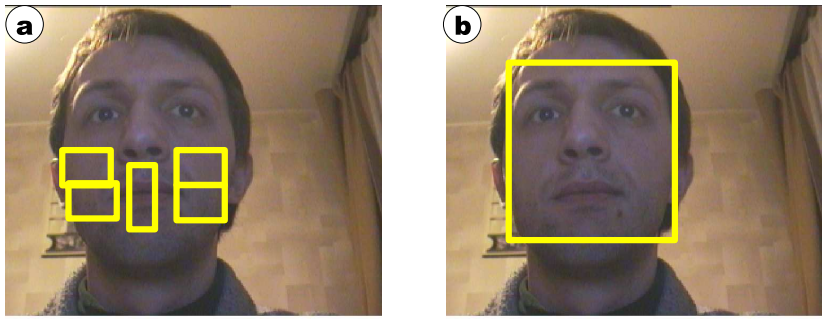
Z przeprowadzonych testów wynika, że rozpoznawanie jednostek czynnościowych przy pomocy statystycznych modeli kształtów jest skuteczne. Testy nr 1 i 2 dla wszystkich gestów: $> 90\%$. Testy nr 3 i 4 $> 86\%$ (oprócz jednego z gestów dla którego wyniki były niższe). Wybór rodzaju klasyfikatora ma wpływ na rezultaty, przy czym dla większej ilości rozpoznawanych jednostek czynnościowych lepszy okazał się klasyfikator kNN.

Istotną obserwacją jest stosunkowo niewielka wrażliwość algorytmu rozpoznawania opartego o kształty na cechy osobnicze (dobre rezultaty dla odróżniania kształtów dla różnych osób). Daje to niezaprzeczalną zaletę w porównaniu do algorytmu opartego na histogramach orientacji, które z definicji są wrażliwe na cechy osobnicze. Należy jednak pamiętać, że badania przeprowadzono na niewielkiej grupie osób, w związku z czym przy budowie docelowego systemu rozpoznawania mimiki konieczne jest przeprowadzenie dodatkowych testów.

Ręczna adnotacja punktów, w badaniach opisanych w rozprawie, pozwala na odróżnienie wpływu skuteczności lokalizacji punktów na wyniki rozpoznawania, a co za tym idzie ocenę jakości samej klasyfikacji kształtów.

### 7.3.2 Histogramy orientacji

Przygotowanie danych do rozpoznawania w przypadku histogramów orientacji wymaga określenia obszaru zainteresowań na obrazie (ROI), w którym spodziewane są zmiany wynikające z mimiki (zmarszczki, bruzdy, fałdy). Najprostszym obszarem, który może zostać wybrany, jest cała twarz (rys. 7.3a). Powoduje to jednakże utrudnienia w klasyfikacji, ze względu na konieczność odróżnienia od siebie wielu gestów. Ponadto określenie ROI dla całej twarzy nie zawsze jest konieczne — przykładowo jeśli do sterowania ruchem kursora wykorzystywana będzie mimika dolnej części twarzy, a do emulacji przycisków gesty mimiczne górnej części, możliwe jest zdefiniowanie kilku ROI (rys. 7.3b) dla których rozpoznawany będzie określony zestaw gestów. ROI te powinny być umieszczone w miejscach, w których występują największe zmiany wizualne wywołane mimiką.



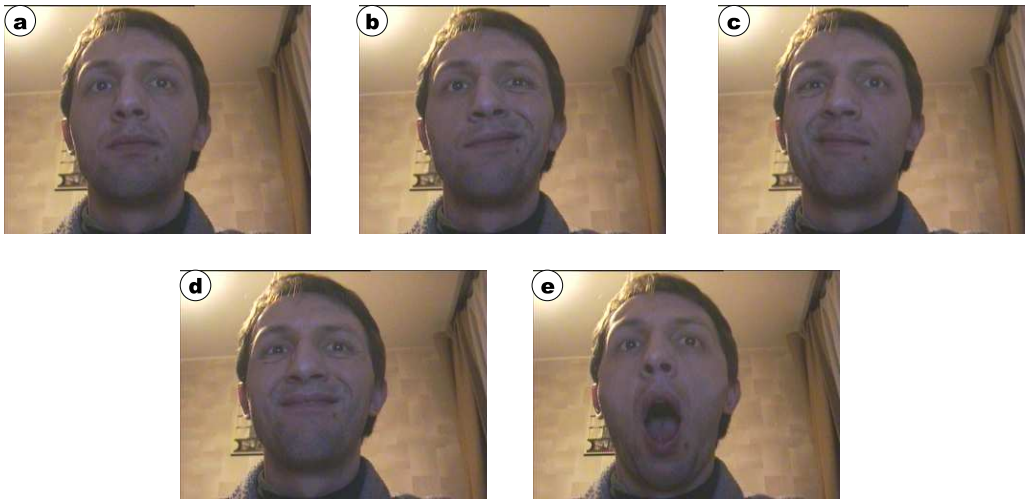

Rysunek 7.3: Wybrane obszary zainteresowań dla rozpoznawania gestów: (a) wybrane ROI, (b) cała twarz.

Ręczne zaznaczanie obszarów zainteresowań jest bardzo czasochłonne i nie spełnia założenia automatycznej pracy systemu (bez udziału operatora). Dlatego istotną kwestią jest automatyczne pozycjonowanie ROI na poszczególnych obrazach sekwencji video. Na potrzeby testów skuteczności algorytmu rozpoznawania, położenie ROI zostało określone na podstawie względnego położenia w stosunku do nozdrzy, zlokalizowanych ręcznie. Automatyzacja tego procesu została opisana w rozdziale 8.4.

Przeprowadzono badanie skuteczności rozpoznawania następujących gestów mimicznych dolnej części twarzy:

- AU0 — położenie neutralne,
- AU12 — unoszenie kącików ust - obustronne i jednostronne,
- AU25+AU26 — otwarcie ust.

Na rysunku 7.4 przedstawiono poszczególne gesty mimiczne (zdjęcie dla ich maksymalnej intensywności).




Rysunek 7.4: Rozpoznawane gesty mimiczne dolnej części twarzy: (a) AU0, (b) AU12P, (c) AU12L, (d) AU12, (e) AU25+26.

Sekwencje filmowe pobrano przy pomocy kamery Sony-EVI, z następującymi parametrami: rozdzielczość – 320*240, ilość ramek na sekundę FPS=25 (ang. frames per second). Obrazy zostały przekonwertowane z przestrzeni RGB do skali szarości (8 bitów). Sekwencje zawierają kolejno:

- Sekwencja 1. Na początku pracy użytkownika z systemem dokonywana jest kalibracja, podczas której użytkownik proszony jest o wykonanie wybranych gestów. Jednokrotne powtórzenie danego gestu, przy powyższych ustawieniach kamery skutkuje około kilkunastoma obrazami zawierającymi

poszczególne gesty mimiczne. Pozostałe obrazy sekwencji zawierają położenie neutralne twarzy oraz elementy przejściowe pomiędzy poszczególnymi gestami mimiki.

- Sekwencja 2. Następnie, przy tych samych ustawieniach i położeniu głowy pobierana jest sekwencja zawierająca trzykrotnie powtórzony każdy z gestów.

- Sekwencja 3. W celu określenia wpływu oświetlenia kolejna sekwencja zawiera obrazy mimiki (dwukrotnie powtórzenie gestów) z włączonym dodatkowym oświetleniem bocznym.

- Sekwencja 4. Ostatnie dwie sekwencje zawierają dwukrotnie powtórzone gesty dla innego położenia głowy (oddalenie od kamery oraz przechylenie głowy). Na tej podstawie określono wpływ zmian skali i rotacji twarzy na wyniki rozpoznawania.

Jako zbiór danych uczących przyjęto obrazy z sekwencji zawierającej kalibrację systemu (nr 1). Pozostałe dane tworzyły osobne zbiory testowe. Parametry klasyfikatora dobrano z wykorzystaniem kryterium dokładności (7.2) dla sekwencji nr 1. Są one następujące:

- dla klasyfikatora kNN: metryka korelacyjna, K=2,

- rodzaj analizy dyskryminacyjnej: kwadratowa analiza dyskryminacyjna QDA z diagonalną macierzą kowariancji (naiwny klasyfikator Bayesa).

Aby ocenić wpływ wyboru obszaru zainteresowań (rys. 7.3a, rys. 7.3b) oraz typu histogramów orientacji na skuteczność rozpoznawania wykonano testy z wykorzystaniem sekwencji nr 2 dla następujących konfiguracji:

- konfiguracja 1: obszar zainteresowań – wybrane ROI, histogramy orientacji – gradienty w przestrzeni skali (ang. scale-space),

- konfiguracja 2: obszar zainteresowań – wybrane ROI, histogramy orientacji – filtry Gabora,

- konfiguracja 3: obszar zainteresowań – cała twarz, histogramy orientacji – gradienty w przestrzeni skali,

- konfiguracja 4: obszar zainteresowań – cała twarz, histogramy orientacji – filtry Gabora.

Rezultaty rozpoznawania dla powyższych konfiguracji przedstawiono w dodatku C, tabele: C.4 — C.11, wykresy: C.1 — C.8.

Porównując metody klasyfikacji, lepszym okazał się klasyfikator kNN. Natomiast klasyfikator QDA pozwala na lepsze odróżnienie gestu AU0 (neutralnego), co może zostać wykorzystane do stworzenia hierarchicznej struktury klasyfikacji — najpierw rozpoznawana jest sytuacja wykonania gestu przez użytkownika, później rozróżniane gesty.

Biorąc pod uwagę sposoby wyznaczania histogramów orientacji, lepsze rezultaty otrzymano dla histogramów opartych o gradienty w przestrzeni skali (ang. scale-space).

Oceniając wpływ wyboru obszaru zainteresowań, można zauważyć dużo słabsze wyniki dla przypadku ROI obejmującego całą twarz. Potwierdza to hipotezę że wybór większego obszaru utrudnia rozpoznawanie. Jest to najprawdopodobniej spowodowane tym, iż dla większego obszaru jakim jest cała twarz lokalne zmiany wyglądu od gestów są niewielkie i trudne do rozróżnienia. Potwierdza to porównanie wyników dla gestów AU12 i AU25+26 — gest otwarcia ust (AU25+26) powoduje większe zmiany na obrazie i jest lepiej rozpoznawany. Inną przyczyną są również — możliwość wystąpienia zmian obrazu pochodzących od innych gestów, wpływ ruchu całej głowy.

Oceny wpływu zmian oświetlenia oraz ruchów głowy dokonano dla metody klasyfikacji oraz typu histogramów orientacji dających najlepsze rezultaty — tj. wybrane ROI, klasyfikator kNN, histogramy orientacji — gradienty w przestrzeni skali). Wykonano następujące testy:

- ocena wpływu bocznego oświetlenia sztucznego, sekwencja 2,
- ocena wpływu zmian skali (odsunięcie głowy od kamery), sekwencja 3,
- ocena wpływu zmian skali i rotacji (odsunięcie głowy połączone z jej przechyleniem), sekwencja 4.

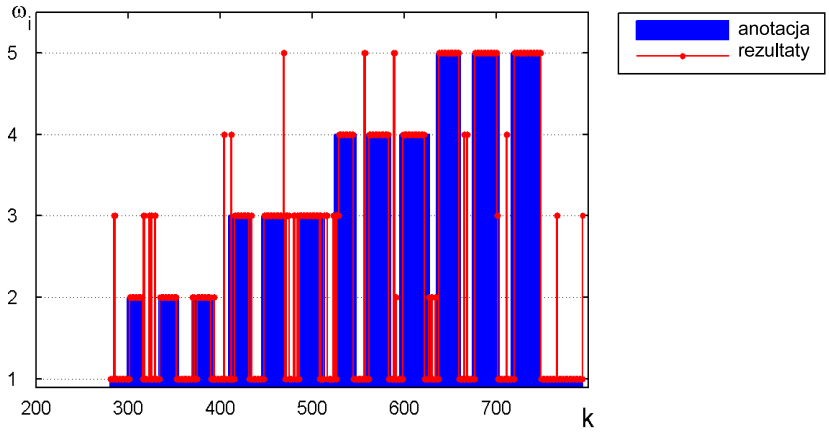
Wyniki przedstawiono w dodatku C, tabele: C.12 — C.14, wykresy: C.9 — C.11.

Podana analiza wpływu czynników zakłócających takich jak dodatkowe oświetlenie, zmiany skali i rotacji wywołane ruchami głowy, widoczne jest iż w takich przypadkach skuteczność klasyfikacji w dużym stopniu spada. Najmniejszy wpływ mają zmiany skali — tylko gest AU12R był błędnie rozpoznawany. Dodatkowa analiza wyników, wskazuje jako przyczynę błędne klasyfikowanie AU12R do AU12. Jest tak dlatego, że gesty te są bardzo podobne do siebie (unoszenie obu kącików i unoszenie jednego z nich). W przypadku zmian oświetlenia, widoczne jest że największy wpływ ma ono na gest AU12L. Jest to zrozumiałe, ponieważ dodatkowe oświetlenie było skierowane na tą część twarzy. Największe zakłócenia powoduje rotacja — zmienia się wtedy orientacja wszystkich krawędzi na obrazie twarzy.

Autor pracy proponuje następujące kierunki dalszych badań, mające na celu poprawę jakości klasyfikacji:

- Wykorzystanie informacji czasowej do rozpoznawania (idea omówiona poniżej).
- Wykrywanie zmian skali i rotacji, automatyczna rekalibracja klasyfikatorów (uwzględnienie w klasyfikacji nowych obserwacji).
- Śledzenie twarzy i estymacja parametrów ruchu (skala, rotacja). Korekta histogramów orientacji wg otrzymanych parametrów.
- Wykorzystanie metod doboru zmiennych diagnostycznych (np. analiza czynnikowa, metoda Helwinga) do selekcji orientacji niosących informacje, najistotniejsze z punktu widzenia klasyfikacji.

Niska jakość klasyfikacji w przypadku czynników zakłócających niekoniecznie musi prowadzić do niepoprawnego działania systemu. Podstawowym zadaniem systemu rozpoznawania mimiki jest generacja poleceń np. dla systemu operacyjnego komputera. Reagowanie na zmiany wykrywane dla każdego obrazu sekwencji video, może powodować generację szeregu błędnych poleceń (każde zakłócenie). Dlatego konieczne jest uwzględnienie kontekstu czasowego — np. wykrycie kilku lub kilkunastu obrazów dla których otrzymano ten sam wynik rozpoznania gestu. Potwierdzeniem skuteczności takiego podejścia jest następujący wykres (rys. 7.5), na którym przedstawiono wyniki klasyfikacji gestów dla całej sekwencji testowej. Dodatkowo nałożono informację o rzeczywistym geście występującym na danej ramce (ręczna anotacja ramek dokonana przez człowieka).




Rysunek 7.5: Rezultaty klasyfikacji dla konfiguracji 1 przedstawione w formie wykresu czasowego. Kolorem niebieskim oznaczono ręczne przyporządkowanie kolejnych ramek do poszczególnych klas, niebieskim – rezultaty klasyfikacji. $\omega_i$ - i-ta klasa, $k$ - numer obrazu sekwencji

# Rozdział 8

# Adaptacja systemu rozpoznawania mimiki

**Streszczenie**

Duża zmienność wyglądu twarzy oraz elementów mimiki, wynikająca z wpływu różnych czynników stanowi istotne utrudnienie w automatycznym rozpoznawaniu gestów. W rozdziale usystematyzowano czynniki wpływające na skuteczność rozpoznawania oraz zaproponowano metody adaptacji systemu rozpoznawania mimiki do człowieka oraz zmieniających się warunków otoczenia.

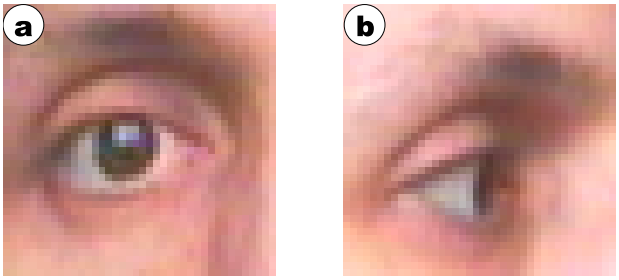
## 8.1 Wstęp

Czynniki, mające wpływ na wygląd twarzy oraz elementów mimiki, można podzielić na dwie grupy [29]:

- zewnętrzne źródła zmienności (ang. extrinsic sources of variation) oraz
- wewnętrzne źródła zmienności wyglądu twarzy (ang. intrinsic sources of variation in facial appearance).

Do zewnętrznych źródeł zmienności zalicza się przede wszystkim geometrię sceny (ang. viewing geometry). Obraz jest dwuwymiarowym odwzorowaniem trójwymiarowej sceny, dlatego też następuje utrata części informacji szczególnie widoczna w przypadku dużych ruchów głowy — obroty, przechylenia (rys. 8.1). Istotny wpływ ma również oświetlenie sceny (ang. illumination), które powoduje powstawanie cieni, odblasków oraz skutkuje zmianami kolorystyki obrazu. Wśród pozostałych zewnętrznych czynników można wyróżnić przysłanianie przez inne obiekty (np. ręka użytkownika, okulary), a także samą akwizycję obrazu która

wiąże się z ograniczoną rozdzielczością przetwornika oraz powstawaniem szumów i zakłóceń.




Rysunek 8.1: Obraz oka: (a) dla twarzy widzianej frontalnie, (b) twarzy z profilu.

Do wewnętrznych źródeł zmienności wyglądu twarzy zaliczyć można tożsamość — różne osoby mają odmienny wygląd, mimo iż ogólna budowa morfologiczna twarzy jest taka sama. Inne czynniki, które mają wpływ na wygląd twarzy, to wiek i płeć osoby, elementy charakterystyczne (np. włosy, broda...) oraz indywidualny sposób wykonywania gestów mimicznych.

Podczas interakcji człowieka z maszyną, w każdej chwili może wystąpić zmiana oświetlenia sceny, położenia osoby względem kamery lub innych parametrów. Z systemu mogą ponadto korzystać różne osoby posiadające indywidualne cechy tożsamości wynikające z wyglądu, cech osobniczych, itp. **Z tego powodu, kluczowym zagadnieniem staje się konieczność adaptacji systemu do indywidualnych cech człowieka oraz zmieniających się warunków otoczenia (ang. visual learning and adaptation).**

Zagadnienia adaptacji interfejsu stanowią bardzo obszerny temat badań. W niniejszej pracy przedstawione zostały wybrane problemy badawcze oraz zaproponowane przez autora rozwiązania:

- Estymacja położenia głowy człowieka względem kamery.
- Wpływ parametrów kamery oraz zmian oświetlenia sceny — automatyzacja tworzenia modelu barwy skóry.
- Uwzględnienie indywidualnego wyglądu człowieka oraz osobniczych cech mimiki — metoda kalibracji systemu w oparciu o detekcję mrugnięć.

Uwzględnienie zewnętrznych źródeł zmienności (oświetlenie, pozycja...) powinno odbywać się w sposób ciągły podczas pracy systemu. Z kolei wpływ tożsamości, wieku, cech osobniczych może zostać uwzględniony rzadziej — na etapie kalibracji systemu. Kalibracja ta powinna odbywać się każdorazowo gdy nowa osoba rozpocznie korzystanie z systemu, bądź też w momencie gdy wystąpią istotne zmiany w wyglądzie tej osoby (np. założenie okularów).

## 8.2 Estymacja położenia głowy człowieka względem kamery

Opisane w poprzednich rozdziałach algorytmy rozpoznawania elementów mimiki, są dostosowane do sytuacji gdy twarz osoby jest w położeniu frontalnym. Dlatego też, w trakcie pracy systemu, konieczne jest określenie położenia głowy człowieka w stosunku do pozycji neutralnej (twarz widziana frontalnie). Dokładna estymacja położenia obiektu na scenie trójwymiarowej przy użyciu jednej kamery, jest co prawda możliwa ale bardzo trudna i skomplikowana obliczeniowo. Z reguły przyjmowany jest model kamery o uproszczonej perspektywie (ang. weak camera model). Jest on poprawny, w przypadku gdy odległości pomiędzy elementami twarzy są niewielkie w stosunku do odległości pomiędzy głową a kamerą. Przyjmuje się że środek układu współrzędnych obserwowanego obiektu znajduje się w środku osi głowy. W takim założeniu oddalanie i przybliżanie kamery skutkuje zmianą skali twarzy na obrazie, przechylenia głowy — rotacją w płaszczyźnie obrazu, obroty głowy (góra-dół, prawo-lewo) — rotacją w płaszczyźnie prostopadłej do płaszczyzny obrazu.

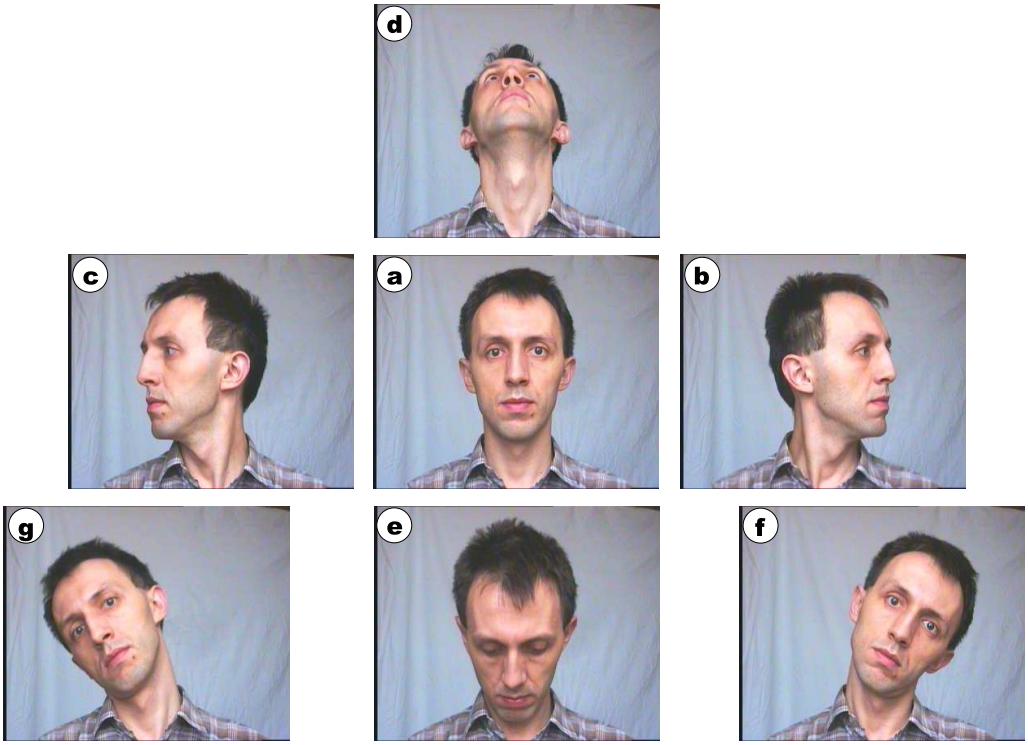
Można założyć, iż w normalnych warunkach pracy z komputerem, typowe ruchy głowy ograniczają się do (rys. 8.2):

- oddalania-przybliżania,
- przechylenia głowy prawo-lewo,
- obrotów w lewo-prawo oraz góra-dół.

Dla kamery umieszczonej naprzeciwko twarzy człowieka oraz przyjęciu środka układu współrzędnych obserwowanego obiektu (twarz) w środku osi głowy, powyższe ruchy skutkują odpowiednio:

- zmianą skali twarzy na obrazie,
- rotacją w płaszczyźnie obrazu,
- rotacją w płaszczyznach prostopadłych do płaszczyzny obrazu (zniekształcenia perspektywiczne oraz przysłanianie części elementów twarzy).

Problem określenia położenia głowy w stosunku do pozycji neutralnej sprowadza się zazwyczaj do wyznaczenia dopasowania między dwoma obrazami (ang. correspondence problem). Na jednym z nich twarz znajduje się w położeniu frontalnym, na drugim w położeniu które jest nieznane. Istnieje wiele metod pozwalających na rozwiązanie problemu dopasowania obrazów. Ogólnie można je podzielić na:




Rysunek 8.2: Głowa widziana (a) frontalnie, (b)(c) obroty prawo-lewo, (d)(e) obroty góra-dół, (f)(g) przechylenia prawo-lewo.

- metody wykorzystujące znalezione strukturalne cechy obrazu, niezależne od zmian oświetlenia oraz punktu widzenia kamery (ang. feature invariant approaches),
- metody oparte na wyszukiwaniu wzorców (ang. template matching methods, correlation based methods).

Przykładem algorytmów z pierwszej grupy jest publikacja [97], w której autorzy przedstawili algorytm estymacji pozycji głowy wykorzystujący zestaw punktów charakterystycznych twarzy (oczy, usta, nos). Cechy te są śledzone na poszczególnych obrazach sekwencji wideo z użyciem korelacyjnej metody dopasowania do wzorca. Na podstawie ich położenia na wzorcowej ramce oraz położenia na obrazie aktualnym, obliczane są parametry transformacji afinicznej (ang. affine transform). Transformacja ta charakteryzuje się najmniejszym błędem dla przyjętego modelu kamery o uproszczonej perspektywie (ang. weak camera model). Model ten jest poprawny, w przypadku gdy odległości pomiędzy elementami twarzy są niewielkie w stosunku do odległości pomiędzy głową a kamerą. Na podstawie wyznaczonych parametrów oraz reprezentacji twarzy w postaci elipsy, określane są: położenie twarzy, rotacja w płaszczyźnie obrazu oraz rotacje w płaszczyznach prostopadłych.

Odmienne podejście [59] wykorzystuje model oparty o zapamiętane wzorce wyglądu twarzy dla różnych pozycji. Wzorce te poddawane są analizie składowych głównych PCA. Autorzy udowodnili, że tylko pierwsze kilka składowych głównych, zawiera informację o zmianach położenia twarzy — występuje wtedy największa wariancja pikseli obrazu. Dla kolejnych obrazów sekwencji wideo przedstawiającej ruch głowy, w przestrzeni składowych tworzy się trajektoria odpowiadająca kolejnym położeniom (ang. pose manifold). Ponowny ruch w tym samym kierunku skutkuje podobną trajektorią. W oparciu o analizę otrzymanych trajektorii w przestrzeni cech, autorzy McKenna, Gong opracowali algorytm estymacji położenia twarzy.

### 8.2.1 Algorytm estymacji położenia głowy

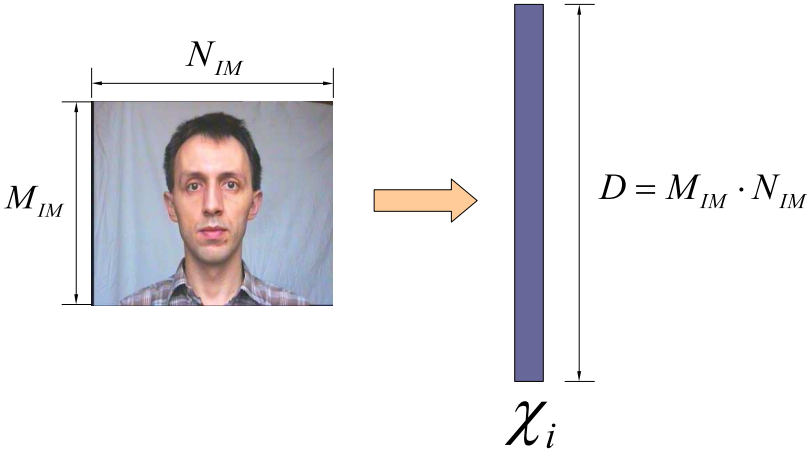
Obrazy twarzy charakteryzują się dużą zmiennością wywołaną wieloma czynnikami. Najistotniejsze z nich (pomijając ruchy głowy) to mimika oraz oświetlenie. Uwzględnienie zmienności będącej skutkiem ruchów głowy przy jednoczesnym pominięciu wpływu innych czynników jest możliwe dzięki analizie statystycznej. Spośród kilku metod wybrano algorytm oparty o analizę składowych głównych (PCA – ang. Principal Component Analysis) określany w literaturze jako „eigenfaces" [59]. Algorytm ten został zmodyfikowany oraz przystosowany dla systemu rozpoznawania mimiki twarzy. Modyfikacje obejmują m.in.:

- wykorzystanie modularnych przestrzeni dla rozróżnienia poszczególnych pozycji głowy,
- sposób porównywania obrazów ze wzorcami oraz
- metodę korekcji globalnych zmian oświetlenia.

Podstawą algorytmu określania położenia głowy jest przygotowana wcześniej baza, zawierająca wzorce twarzy pobrane podczas ruchów głowy względem kamery. Dla każdego rodzaju ruchu głowy (przechylenia, obroty...), uzyskiwana jest sekwencja obrazów zawierająca widok twarzy w jej kolejnych położeniach (rys. 8.3). Z widoków tych pobierany jest tylko wycinek zawierający twarz. Każdy obraz z bazy wzorców zamieniany jest na wektor cech poprzez odpowiednie ułożenie kolejnych pikseli np. kolumnami — rys. 8.4. Następnie wektory te są normalizowane poprzez odjęcie „średniego obrazu twarzy" (ang. mean face). Ponadto każdy obraz testowy jest wstępnie przetwarzany w celu usunięcia szumów i zmniejszenia wymagań obliczeniowych (m.in. filtracja medianowa, zmiana rozmiaru).

Na rysunku 8.5 przedstawiono poszczególne kroki algorytmu określania położenia. W pierwszym etapie, przy pomocy metody PCA, wyznaczana jest reprezentacja wzorców twarzy w nowej przestrzeni cech. Są to: średni obraz twarzy $\bar{\chi}$, składowe główne oraz wagi dla wzorców. Użyta w metodzie analiza składowych głównych PCA została opisana dokładniej w dodatku A.4.

Rysunek 8.3: Wzorce twarzy dla obrotu głowy w prawo i z powrotem (dla uproszczenia rysunek zawiera co czwarty wzorzec).




Rysunek 8.4: Ilustracja zasady tworzenia wektorów cech odpowiadających wzorcom twarzy.

Składowe główne $U_\Phi$ (ang. principal components), określane również jako „eigen-obiekty", stanowią nową bazę przestrzeni, do której rzutowane są wzorce twarzy. W procesie rzutowania wyliczany jest wektor wag $\Omega_{\chi_i}$ reprezentujący dany wzorzec.

W przypadku określania położenia, największa zmienność jest wynikiem ruchów głowy. Pozostałe kierunki (osie nowej przestrzeni) mogą zostać pominięte ponieważ kodują zmienność pochodzącą np. od szumów, zakłóceń, ruchów mimicznych. Dlatego ze wszystkich składowych głównych, wybierane jest $K_\Phi$ pierwszych elementów. Wartość $K_\Phi$ obliczana jest na podstawie kryterium (8.1).

$$\sum_{i=1}^{K_\Phi} \lambda_i \geqslant \frac{p_\Phi}{100} \cdot \sum_{i=1}^{D} \lambda_i \tag{8.1}$$

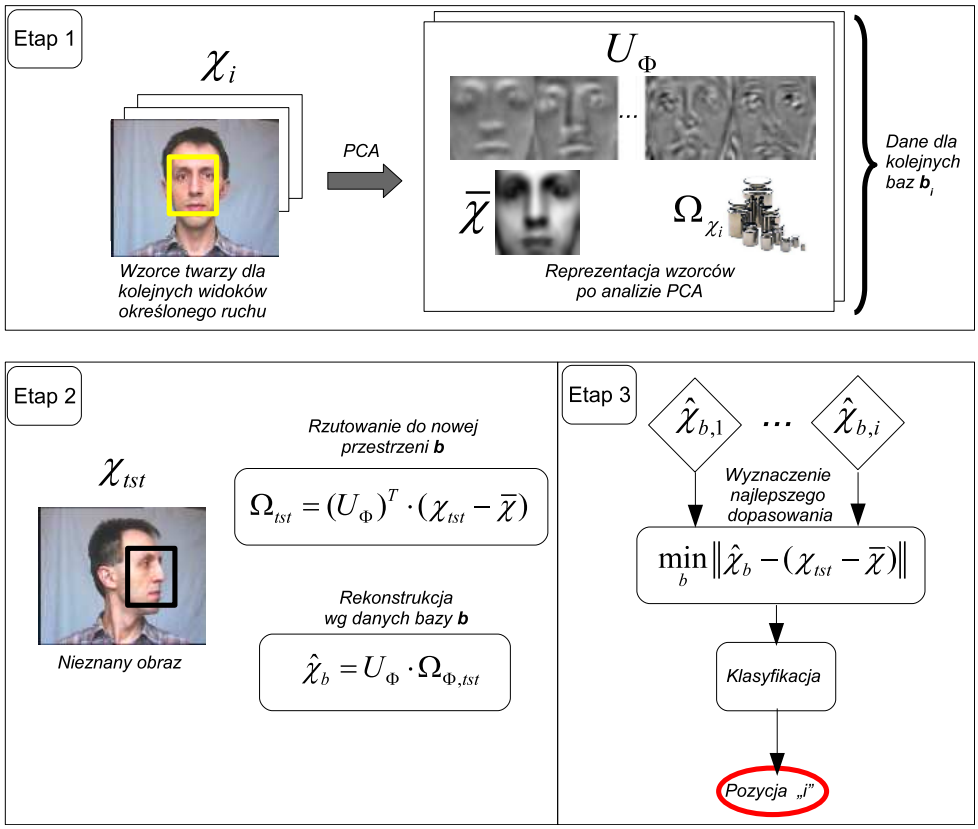
gdzie:

$i$ – indeks kolejnego wzorca

$\lambda_i$ – i-ta wartość własna macierzy kowariancji danych wzorców

$D$ – ilość pikseli obrazu wzorca

$p_\Phi$ – zadana, procentowa wartość określająca proporcję wariancji danych z zestawu wzorców




Rysunek 8.5: Etapy algorytmu określania położenia głowy.

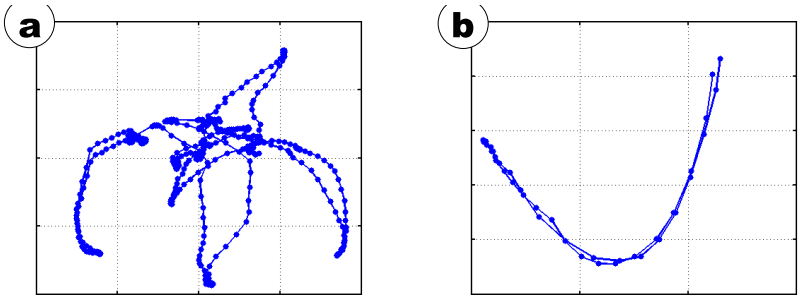
Każdy wektor cech (a co za tym idzie obraz wzorca) możne zostać odtworzony poprzez kombinację liniową składowych głównych (8.2) oraz dodanie średniego wektora cech.

$$\tilde{\chi}_i = \bar{\chi} + U_\Phi \cdot \Omega_{\Phi,i} \tag{8.2}$$

gdzie:

$\tilde{\chi}_i$ – odtworzony wektor cech (obraz i-go wzorca)
$\bar{\chi}$ – wektor średniego obrazu twarzy
$U_\Phi$ – składowe główne
$\Omega_{\Phi,i}$ – wektor wag i-go wzorca

W wyniku rzutowania wzorców odpowiadających kolejnym położeniom głowy do nowej przestrzeni, otrzymuje się wektory cech tworzące w tej przestrzeni trajektorię odpowiadającą zmianie pozycji (ang. pose manifold) — rys. 8.6.




Rysunek 8.6: Wykres trajektorii (dwa pierwsze elementy wektora wag): (a) dla wszystkich ruchów, (b) dla obrotu głowy w prawo.

Określenie pozycji w rozwiązaniu zaproponowanym w artykule [59], odbywa się poprzez obliczenie odległości pomiędzy wagami badanego obrazu, a zapamiętaną trajektorią dla wszystkich ruchów (8.4) — rys. 8.7. Badany obraz jest przygotowywany tak samo jak wzorce (pobranie wycinka twarzy, usuwanie szumów). Znalezienie najlepszego dopasowania pomiędzy wzorcem a nieznanym obrazem jest realizowane poprzez wyznaczenie minimum odległości (8.3).

$$\min_i \|\Omega_{\chi_i} - \Omega_{tst}\| \tag{8.3}$$

$$\Omega_{tst} = (U_\Phi)^T \cdot (\chi_{tst} - \bar{\chi}) \tag{8.4}$$

gdzie:

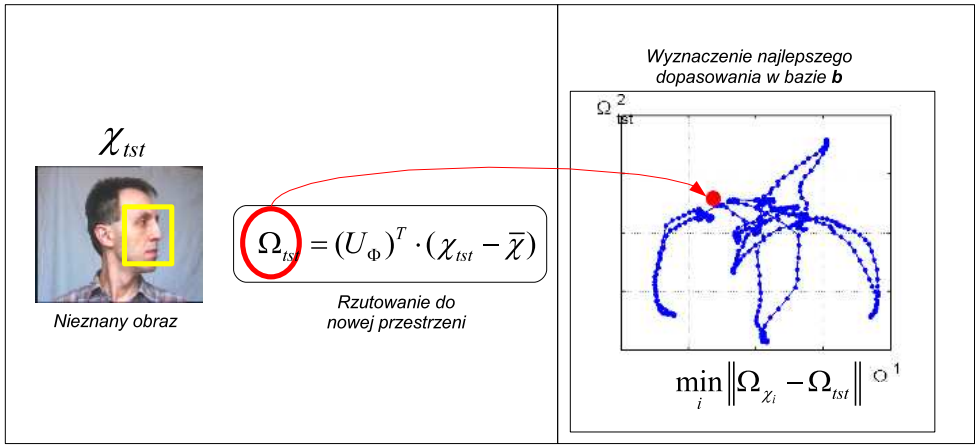
$\chi_{tst}$ – nieznany obraz (wektor cech)
$\Omega_{tst}$ – wektor wag dla nieznanego obrazu (po rzutowaniu do nowej przestrzeni)
$\bar{\chi}$ – wektor średniego obrazu twarzy
$U_\Phi$ – składowe główne

Opisane rozwiązanie wykorzystuje jedną bazę wzorców (ang. general pose eigenspace), w której umieszczone są widoki twarzy dla wszystkich pozycji. Może to powodować błędne rozpoznanie w przypadku gdy baza zawiera elementy odstające (ang. outliers) lub trajektorie dla poszczególnych ruchów leżą blisko siebie.

Dlatego w niniejszej pracy użyto alternatywna metodę określania pozycji głowy, wykorzystującą osobne bazy dla każdego ruchu (ang. modular pose eigenspace). Można ją przyrównać do obserwacji sceny przez kilku obserwatorów rów-

nocześnie, z których każdy zwraca uwagę na co innego. Bazy, reprezentujące poszczególne pozycje głowy, zostały utworzone poprzez podział sekwencji wzorcowej na części odpowiadające następującym ruchom głowy:

- pozycja neutralna (w przybliżeniu),
- przechylenie głowy w prawo,
- przechylenie głowy w lewo,
- obrót w lewo,
- obrót w prawo,
- obrót w górę,
- obrót w dół.




Rysunek 8.7: Określanie stopnia intensywności ruchu.

Jedną z zalet tej modyfikacji jest wymagana mniejsza ilość komponentów głównych konieczna do opisania położenia oraz mniejsze skomplikowanie trajektorii (rys. 8.6b). Inna, przydatna w systemie rozpoznawania mimiki cecha, to możliwość uzyskania w pierwszej kolejności informacji o rodzaju ruchu (np. prawo, lewo), a w późniejszym etapie określenia stopnia intensywności ruchu.

Klasyfikacja badanego obrazu $\chi_{tst}$ do jednej z baz (odpowiadających poszczególnym pozycjom głowy) odbywa się następująco:

- rzutowanie badanego obrazu do każdej z modularnych przestrzeni, w wyniku otrzymywane są wektory wag $\Omega_{tst}$ dla danej przestrzeni $b$ (8.4),

- rekonstrukcja badanego obrazu przy pomocy $K_\Phi$ składowych głównych danej bazy (8.5),
- klasyfikacja obrazu wg kryterium minimalnego błędu rekonstrukcji (8.6),

Dla czytelności, we wzorach nie wprowadzono dodatkowych indeksów oznaczających bazę.

$$\hat{\chi}_b = U_\Phi \cdot \Omega_{\Phi,tst} \tag{8.5}$$

gdzie:

$\hat{\chi}_b$ – zrekonstruowany wektor cech badanego obrazu
$U_\Phi$ – składowe główne przestrzeni $b$
$\Omega_{\Phi,tst}$ – wektor wag badanego obrazu w przestrzeni $b$

$$\min_b \|\hat{\chi}_b - (\chi_{tst} - \bar{\chi})\| \tag{8.6}$$

gdzie:

$\hat{\chi}_b$ – zrekonstruowany wektor cech badanego obrazu
$\bar{\chi}$ – wektor średniego obrazu twarzy z bazy $b$
$\chi_{tst}$ – wektor cech badanego obrazu

### 8.2.2 Rezultaty estymacji położenia głowy

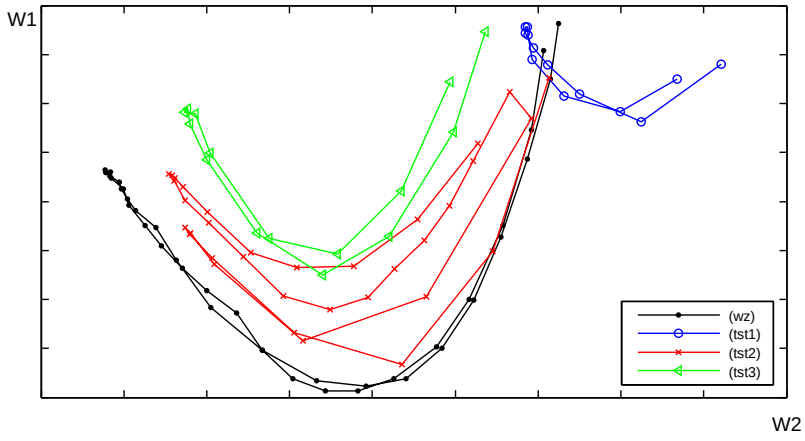
Istotnym parametrem, które ma bezpośredni wpływ na skuteczność określania pozycji, jest wrażliwość algorytmu na zmiany oświetlenia sceny oraz inne czynniki (np. jednocześnie wykonywane ruchy mimiczne). Aby to sprawdzić wyznaczono trajektorie ruchów dla następujących sekwencji video:

- sekwencja 1 — kamera EVI, jednolite tło, słabe oświetlenie (natężenie oświetlenia około 40lux), (rys. 8.8b),
- sekwencja 2 — kamera EVI, jednolite tło, optymalne warunki oświetlenia (natężenie oświetlenia około 320lux), w pozycji frontalnej wykonany szereg gestów mimicznych (rys. 8.8c),
- sekwencja 3 — kamera EVI, jednolite tło, optymalne warunki oświetlenia (natężenie oświetlenia około 320lux), inna mimika twarzy (rys. 8.8d),

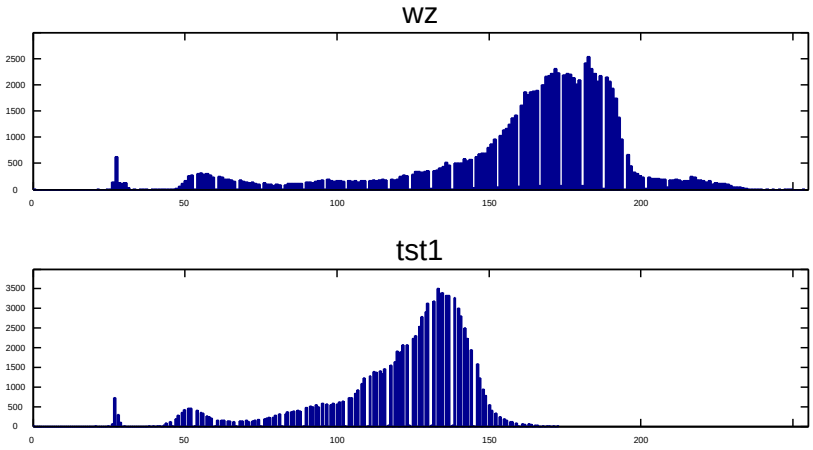

Rysunek 8.8: Przykładowe obrazy z sekwencji video: (a) sekwencja wzorcowa, (b) sekwencja nr 1, (c) sekwencja nr 2, (d) sekwencja nr 3

Jako sekwencję wzorcową (inną niż sekwencje testowe) przyjęto obrazy pobrane kamerą EVI dla następujących parametrów: jednolite tło, optymalne warunki oświetlenia (natężenie oświetlenia około 320lux), (rys. 8.8a). Każdy obraz testowy przetwarzany był wstępnie analogicznie jak obrazy wzorcowe z bazy (m.in. pobranie wycinka obrazu zawierającego twarz, zmiana rozmiaru).

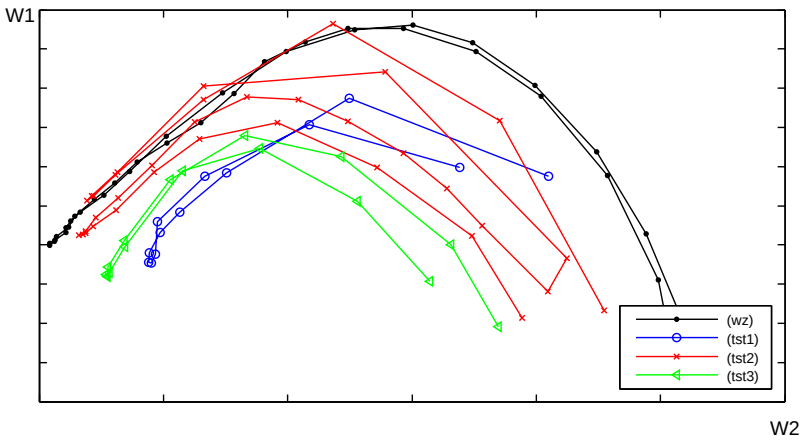
Podczas badań stwierdzono niską skuteczność klasyfikacji, dla sekwencji nr 1 (słabe oświetlenie). Prawie wszystkie obrazy testowe charakteryzowały się błędem rekonstrukcji powyżej założonego progu. Przyczynę tego obrazują rysunki 8.9 i 8.10. Można zauważyć, iż globalna zmiana oświetlenia sceny (np. zmierzch) skutkuje mniejszą dynamiką obrazu. Powoduje to znaczne przesunięcie trajektorii i uniemożliwia znalezienie optymalnego dopasowania pomiędzy wzorcem a nieznanym obrazem. Rozwiązaniem tego problemu jest zastosowanie algorytmu wyrównywania histogramu, który koryguje lokalizację trajektorii w przestrzeni wag — rys. 8.11.




Rysunek 8.9: Wykresy trajektorii (ruch głowy – obrót w prawo) dla poszczególnych sekwencji przed wyrównaniem histogramu: (wz) sekwencja wzorcowa, (tst1) sekwencja nr 1, (tst2) sekwencja nr 2, (tst3) sekwencja nr 3




Rysunek 8.10: Histogramy dla obrazu: (wz) z sekwencji wzorcowej, (tst) oraz sekwencji nr 1)




Rysunek 8.11: Wykresy trajektorii (ruch głowy – obrót w prawo) dla poszczególnych sekwencji po wyrównaniu histogramu: (wz) sekwencja wzorcowa, (tst1) sekwencja nr 1, (tst2) sekwencja nr 2, (tst3) sekwencja nr 3

Rezultaty estymacji położenia głowy przedstawione zostały w dodatku F. Dla położenia frontalnego (czyli najbardziej istotnego z punktu widzenia systemu) skuteczność rozpoznawania jest wystarczająca (większa od 98%). Pozostałe pozycje głowy są gorzej rozpoznawane.

Opisana powyżej metoda estymacji położenia głowy, może służyć również do detekcji i lokalizacji twarzy. W tym celu dla badanego obrazu wejściowego wyznaczana jest mapa dopasowania do modularnych baz pozycji. Poniżej opisano pokrótce procedurę postępowania, powtarzaną dla jednej z kilku założonych skal (wielkość obrazu twarzy):

1. Dla każdego piksela obrazu pobierane jest jego otoczenie o rozmiarze proporcjonalnym do skali (tak aby otrzymany wycinek ROI miał rozmiar równy rozmiarom wzorców w bazie).
2. Dla każdego ROI powtarzana jest procedura wyznaczania najlepszego dopasowania do bazy, analogicznie do opisanej wyżej klasyfikacji nieznanego obrazu do jednej z kategorii pozycji.
3. W wyniku tego, dla każdego piksela są wyznaczane: numer bazy do której został zaklasyfikowany oraz wartości błędu rekonstrukcji dla tej bazy (8.6). Wartość błędów rekonstrukcji tworzą mapę dopasowania.
4. Poprzez znalezienie współrzędnych piksela dla którego występuje minimum mapy dopasowania, określona jest najbardziej prawdopodobna lokalizacja twarzy na badanym obrazie oraz identyfikowana jest baza odpowiadająca pozycji głowy.
5. Na podstawie kryterium minimum odległości pomiędzy wagami ROI rzutowanego do zidentyfikowanej modularnej przestrzeni, a zapamiętaną trajektorią (8.3) możliwe jest określenie pozycji głowy w bazie.

Powyższą procedurę można powtarzać dla różnych skal uzyskując w ten sposób kolejną istotną dla systemu informację — skalę czyli wielkość twarzy na obrazie.

Przedstawione badania nie wyczerpują całkowicie tematyki estymacji pozycji głowy. Można wskazać następujące dalsze kierunki prac:

- automatyczna akwizycja sekwencji video połączona z pomiarem pozycji głowy w przestrzeni,
- badanie skuteczności określania położenia dla różnych reprezentacji obrazu (np. filtry kierunkowe, własności tekstury),
- badanie odporności utworzonych baz na czynniki zakłócające (np. zmiana kąta oświetlenia twarzy, elementy takie jak okulary, zarost),
- badanie wpływu zmian skali na skuteczność detekcji i lokalizacji twarzy.

## 8.3 Wpływ parametrów kamery oraz zmian oświetlenia sceny

Popularny sprzęt wizyjny instalowany w zestawach komputerowych (np. internetowe kamery USB) nie został zaprojektowany dla potrzeb zaawansowanego przetwarzania i analizy obrazu. Wiąże się z tym szereg problemów wymagających uwzględnienia przy opracowywaniu wizyjnego interfejsu człowiek-komputer o niewygórowanych wymaganiach sprzętowych. Problemy te skupiają się wokół następujących aspektów akwizycji obrazu:

- szum,
- ostrość i rozdzielczość obrazu,
- balans bieli,
- oświetlenie i parametry ekspozycji.

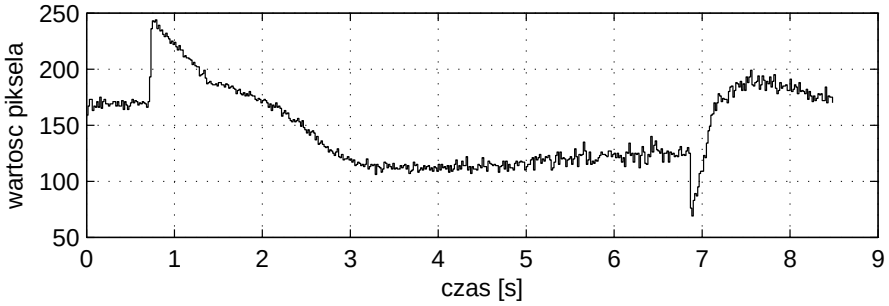
Obecność szumu, generowanego już przez sam sensor kamery, wprowadza stochastyczną zmienność koloru pikseli w następujących po sobie w sekwencji obrazach. W przypadku tanich kamer internetowych stosunek sygnału do szumu jest mały i pogarsza się szybko w miarę zmniejszania oświetlenia, gdy układ automatycznej regulacji wzmocnienia dostosowuje kamerę do pracy w ciemniejszych warunkach. Wymusza to wprowadzenie do opracowywanych algorytmów pewnych metod oddzielających szum od istotnych zmian intensywności pikseli. Zagadnienia te zostały opisane w dodatku A.3.

Także pozostałe parametry (ostrość obrazu, fizyczna rozdzielczość) w przypadku niedrogich kamer, nie są zbyt wysokie. Układ optyczny jest niejednokrotnie bardzo uproszczony i pozbawiony możliwości regulacji. Stosowana jest także często interpolacja obrazu na różnych etapach jego akwizycji. Z punktu widzenia segmentacji barwnej ważna jest również rozdzielczość barwna obrazu wynikowego (głębia kolorów), która w przypadku niektórych kamer, może być ograniczona do 5 bitów na kanał koloru, czyli 32 poziomów jasności (kamera Creative USB WebCam).

Typowe kamery internetowe, wykorzystywane powszechnie przez użytkowników komputerów, są wyposażone w układ regulacji balansu bieli, pomagający unikać przebarwień obrazu w przypadku zmian temperatury barwowej oświetlenia. Działanie tego układu nie zawsze prowadzi jednak do pożądanych rezultatów i nie zapewnia powtarzalności kolorów rejestrowanych w kadrze obiektów szczególnie przy oświetleniu mieszanym (np. światło żarowe i światło monitora komputerowego). Z tego powodu nie jest możliwe z góry dokonanie założenia o wartości koloru skóry twarzy i wykorzystanie go w segmentacji barwnej, gdyż w zależności od kontekstu i oświetlenia sceny odcień skóry przesuwa się w szerokim zakresie widma barw. Z powodu wspomnianego wyżej układu automatycznego wzmocnienia kamery, lokalne zmiany jasności obrazu powodują globalne zmiany parametrów układu przetwornika wizyjnego kamery. Skutkuje to przesunięciem poziomu średnich wartości wszystkich pikseli w czasie. Zostało to zilustrowane na rysunku 8.12.

### 8.3.1 Automatyzacja tworzenia modelu barwy skóry

Jednym z problemów, związanych z segmentacją twarzy, jest ustalenie modelu barwy skóry charakteryzującego twarz aktualnie obserwowanego użytkownika.



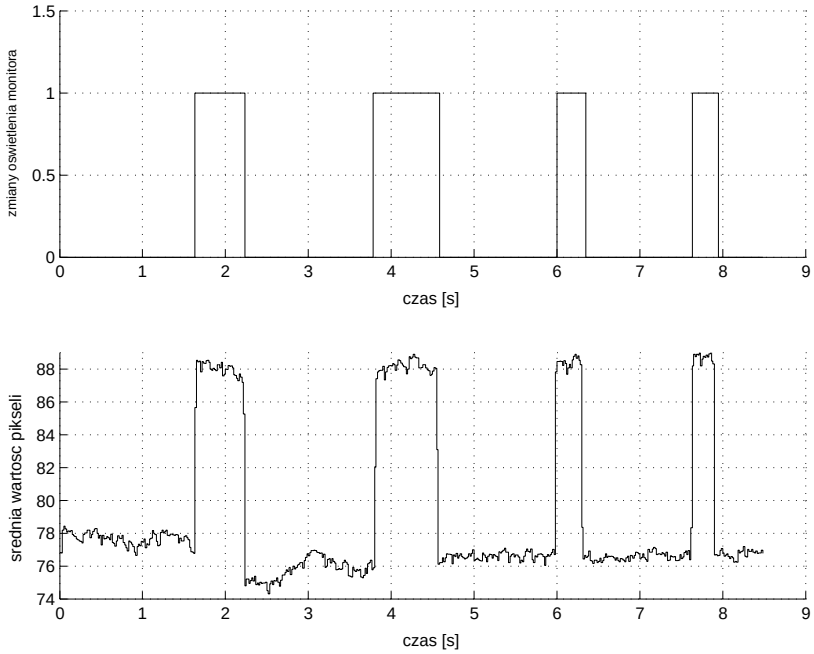

Rysunek 8.12: Zmiana intensywności wybranego piksela wynikająca z działania układu automatycznego wzmocnienia kamery.

Aby wyznaczyć charakterystykę kolorystyczną skóry w sposób automatyczny, konieczne jest wyodrębnienie z obrazu pikseli należących do skóry twarzy, a więc dokonanie segmentacji tego obiektu. Segmentacja ta nie powinna opierać się na jeszcze nie znanym modelu barwy, ale na innym kryterium, nie wykorzystywanym później podczas przetwarzania kolejnych obrazów sekwencji.

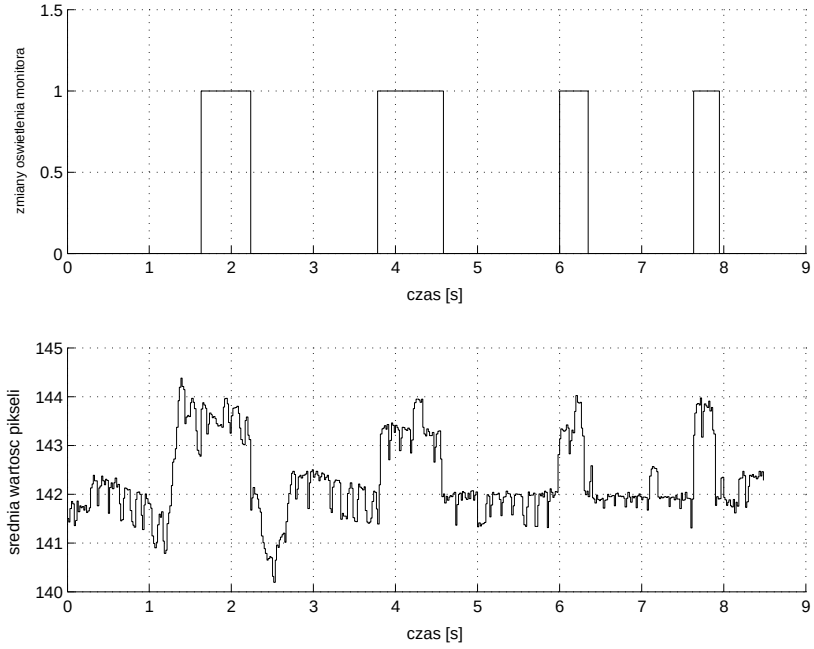
Dodatkowym kryterium segmentacji, które może zostać wykorzystane, jest zmienność oświetlenia obiektów wywołana intencjonalnymi zmianami emisji światła przez monitor komputera, przed którym znajduje się użytkownik. Impulsowe oświetlenie twarzy przez ekran monitora stwarza możliwość odróżnienia jej od pozostałych obiektów znajdujących się w kadrze. Twarz użytkownika jest obiektem znajdującym się bezpośrednio przed ekranem w polu widzenia kamery zamontowanej na monitorze. Programowo sterowane rozjaśnienie i wygaszenie ekranu monitora powoduje zmiany jasności znajdujących się przed nim obiektów, tym większe, im bliżej monitora znajduje się dany obiekt. Efekt ten jest wyraźny, gdyż strumień światła emitowanego przez ekran maleje z kwadratem odległości od niego. Efekt ten obrazują wykresy (rys. 8.13, rys. 8.14) zmian intensywności pikseli należących do niewielkiego wycinka twarzy (np. 3x3).

Uwzględniając szerszy kontekst (ergonomię rozwiązania, bezpieczeństwo użytkownika) — migotanie ekranu monitora powinno zostać ograniczone do jednorazowego etapu wstępnego (oraz ewentualnie rzadko powtarzanej rekalibracji) i nie może być powtarzane wielokrotnie. Te zastrzeżenia dotyczą przede wszystkim użytkowników z zaburzeniami neurologicznymi, u których migotanie ekranu może sprowokować napad padaczkowy. Nie stanowi to jednak przeszkody w wykorzystaniu tej techniki do jednorazowej kalibracji systemu, pomiędzy rozjaśnieniem i wygaszeniem monitora może upłynąć czas rzędu sekundy, zaś użytkownik może w tym czasie mieć zamknięte oczy.

Zagadnienia budowy modelu barwy skóry na podstawie analizy zmian wywołanych sztucznym oświetleniem monitora, zostały szerzej opisane w publikacji autora [41].




Rysunek 8.13: Zmiany średniej intensywności pikseli twarzy podczas impulsowych zmian jaskrawości monitora.




Rysunek 8.14: Zmiany średniej intensywności pikseli tła podczas impulsowych zmian jaskrawości monitora.

## 8.4 Metoda kalibracji systemu

Opisane w poprzednich rozdziałach metody lokalizacji twarzy, rozpoznawania elementów mimiki oraz określenia położenia głowy, do poprawnego działania wymagają szeregu informacji określonych a priori:

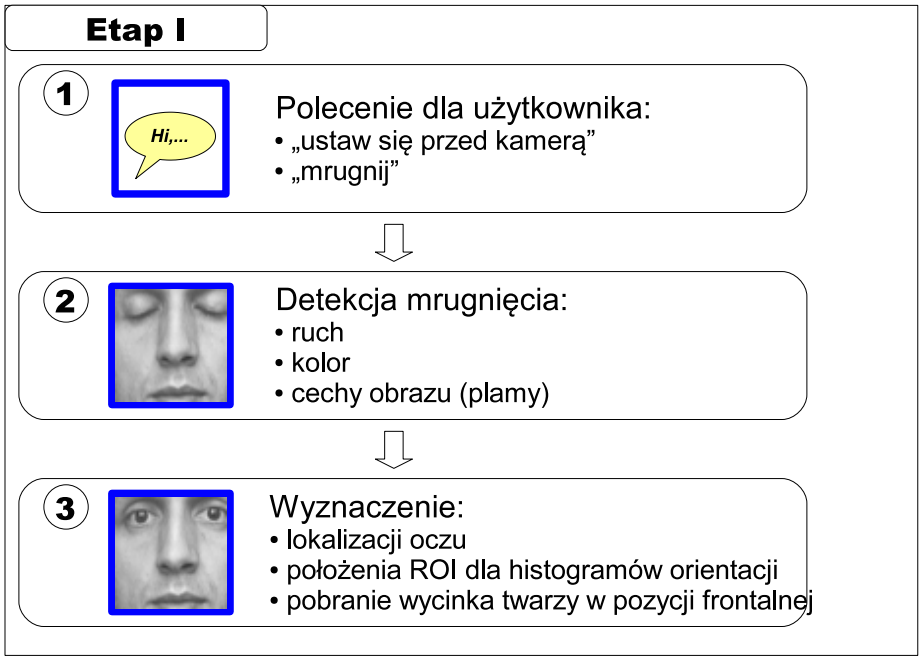
- Używane do detekcji i lokalizacji twarzy (por. 5.3) położenie cech odpowiadających elementom twarzy (oczy, nozdrza, usta) na obrazie oraz ich wzajemne odległości.
- Położenie punktów charakterystycznych kształtów (statystyczne modele kształtu – por. 6.2).
- Dokładne pozycjonowanie obszarów zainteresowań (ROI) na poszczególnych obrazach sekwencji video dla histogramów orientacji (por. 7.3.1).
- Baza wzorców widoków twarzy używanych przez algorytm estymacji położenia głowy do porównywania nieznanego obrazu.

Część z wyżej wymienionych informacji może zostać oszacowana na podstawie ogólnie dostępnej wiedzy. Przykładem mogą być ogólne zależności geometryczne pomiędzy elementami twarzy (np. wzajemne położenie oczu, ust, nosa...), które są znane z anatomii. Pozostałe natomiast — powinny zostać pobrane na etapie kalibracji systemu.

Biorąc pod uwagę wygodę korzystania i ergonomię rozwiązania, kalibracja nie może być czasochłonna i skomplikowana. Ręczne wskazywanie elementów na obrazie (np. wybieranie wycinka obrazu), jest rozwiązaniem które nie spełnia założeń ergonomii użytkowania (szczególnie w przypadku osób niepełnosprawnych ruchowo). Dlatego istotnym celem jest automatyzacja tego procesu.

Implementacja pełnej automatyzacji (nie zakładającej udziału człowieka) jest zadaniem trudnym do realizacji. Z pomocą przychodzi możliwość wykorzystania interakcji z użytkownikiem. System może generować szereg poleceń dla człowieka takich jak np. „ustaw się przed kamerą w pozycji frontalnej", „mrugnij oczami dwa razy", „obróć głowę w prawo", itp. Dzięki temu zostaje znacznie zredukowany stopień skomplikowania obserwowanej sceny, a także uzyskiwane są dodatkowe informacje. Przykładowo — ustawienie głowy w pozycji neutralnej zapewnia, iż elementy twarzy takie jak oczy i usta są dobrze widoczne. Z kolei mruganie oczami ułatwia lokalizację oczu. Proponowany algorytm kalibracji składa się z trzech etapów:

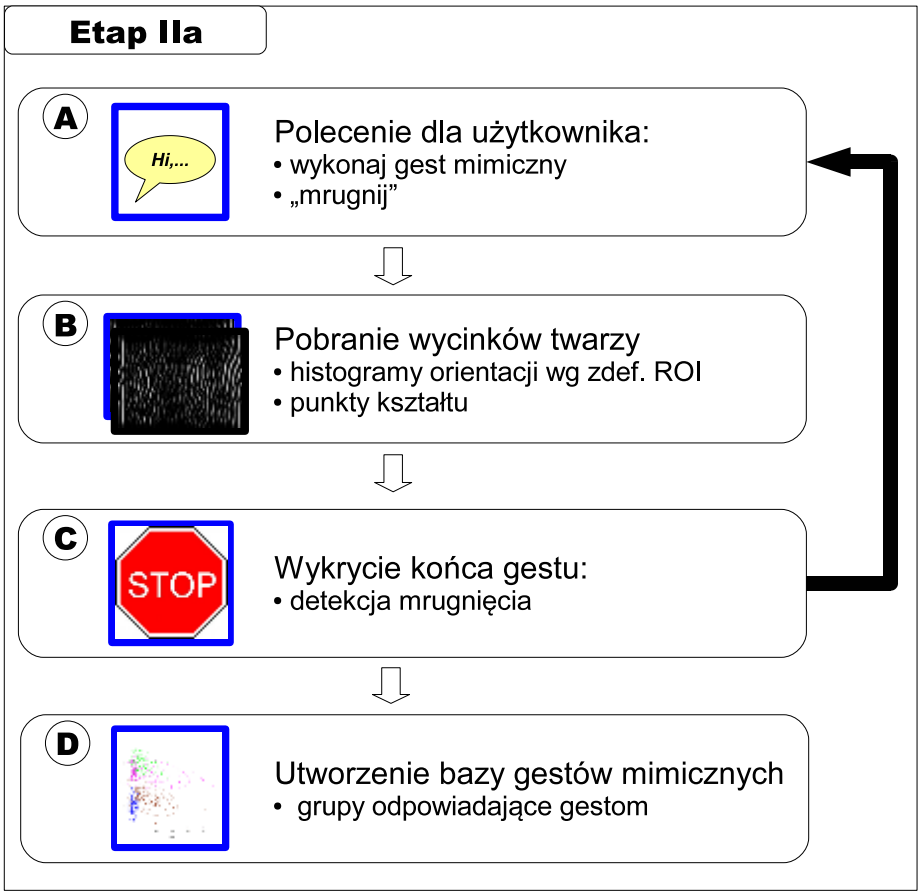
- Etap I (rys. 8.15) — lokalizacja oczu oraz innych elementów twarzy.
- Etap II (rys. 8.16) — tworzenie bazy danych gestów mimicznych.
- Etap III (rys. 8.17) — pobieranie wzorców twarzy dla algorytmu estymacji położenia głowy.




Rysunek 8.15: Etap 1 kalibracji — lokalizacja oczu.

Kalibracja w dużej części opiera się na algorytmie detekcji mrugnięć, dzięki którym uzyskiwana jest informacja o lokalizacji oczu. Zidentyfikowanie położenia obydwu oczu pozwala na określenie pozycji oraz rozmiaru twarzy. Ponadto, w połączeniu z wiedzą o jej budowie anatomicznej, możliwe jest przybliżone określenie położenia pozostałych elementów twarzy (usta, nozdrza...). Upraszcza to w znaczący sposób algorytmy detekcji tych elementów. Informacje te, w połączeniu z modelem barwy skóry określonym poprzez impulsowe oświetlenie od monitora, stanowią podstawę wyznaczania parametrów algorytmu detekcji twarzy.

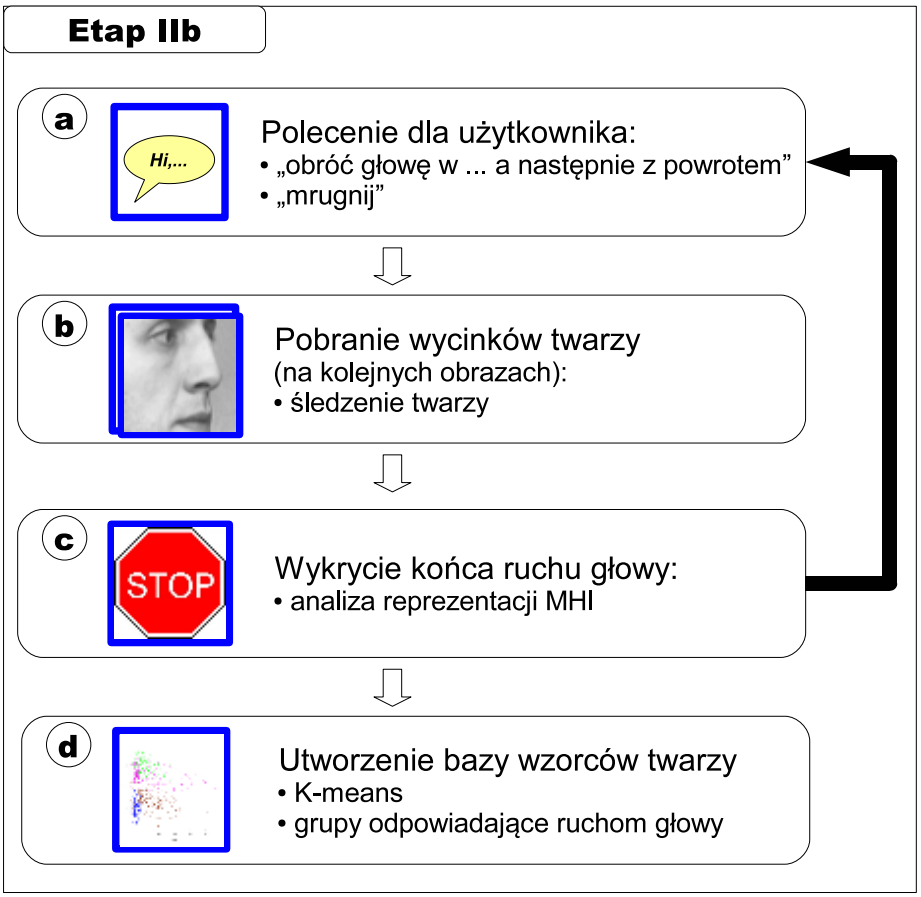
Tworzenie bazy danych gestów mimicznych (etap II) odbywa się poprzez generowanie odpowiednich komunikatów dla użytkownika, który proszony jest o wykonanie wybranych gestów mimicznych. Po wykonaniu każdego z nich, system pobiera i analizuje dane stanowiące zbiór uczący dla algorytmów rozpoznawania. Sygnałem końca wykonania gestu jest detekcja mrugnięcia. W przypadku histogramów orientacji, dane pobierane są ze zdefiniowanych obszarów zainteresowań (ROI). Położenie ROI wyznaczone jest względem niezmiennych części twarzy — w tym przypadku były to nozdrza, które są dobrze widoczne nawet dla większych ruchów głowy. Separację danych należących do różnych gestów mogą zapewnić algorytmy grupowania bez nauczyciela (ang. clustering), takie jak metoda k-średnich (ang. k-means).




Rysunek 8.16: Etap 2 kalibracji — tworzenie bazy elementów mimiki.

Idea trzeciego etapu kalibracji, realizującego pobieranie wzorców twarzy, jest podobna do etapu drugiego — system generuje odpowiednie polecenia dla użytkownika. Po każdym ruchu głową system ponawia polecenie „mrugnij”, co zapewnia informację o początku oraz końcu sekwencji ruchu. Z etapu pierwszego system dysponuje wzorcem twarzy w położeniu frontalnym. Z jego użyciem może być zrealizowane śledzenie twarzy na kolejnych obrazach sekwencji wideo — konieczne ze względu na wymaganą dokładność pobierania wzorców twarzy oraz możliwe przemieszczenia głowy. Podział otrzymanego zestawu wzorców do określonych kategorii pozycji (frontalne, obrót lewo...) odbywa się już bez udziału użytkownika, przy pomocy algorytmu k-średnich (ang. k-means).

Podsumowując zagadnienia przedstawione w rozdziale. Przedstawiona propozycja metody kalibracji systemu wymaga opracowania i implementacji szeregu




Rysunek 8.17: Etap 3 kalibracji — tworzenie bazy wzorców twarzy.

algorytmów. Część z nich została przedstawiona w kolejnych rozdziałach:

- algorytm detekcji mrugnięć (por. 8.4.1),
- algorytm lokalizacji nozdrzy (por. 8.4.2).

Automatyczne wyznaczanie położenie punktów charakterystycznych kształtów oraz ich śledzenie zostało w niniejszej pracy pominięte. Podobnie jak śledzenie twarzy. Prowadzone w tym kierunku prace omawiają następujące artykuły autora:

- „Detekcja markerów dla celów automatycznej adnotacji mimiki twarzy” [69],

- „Śledzenie cech charakterystycznych twarzy w systemie rozpoznawania mimiki" [68].

Prowadzone były również badania nad alternatywną metodą kalibracji systemu, wykorzystującą aktywne metody detekcji charakterystycznych cech twarzy. Ideą była detekcja efektu refleksu siatkówkowego (ang. bright pupil) przy pomocy klasycznych metod przetwarzania obrazu (ang. feature-based). Efekt ten uzyskany został dzięki wykorzystaniu konstrukcji aktywnego oświetlacza podczerwieni. Prace w tym kierunku przedstawione zostały w publikacji [70].

Integracja całości algorytmu kalibracji nie została w pracy uwzględniona — stanowi to (w rozumieniu autora) interesujący kierunek dalszych prac.

### 8.4.1 Detekcja mrugnięć

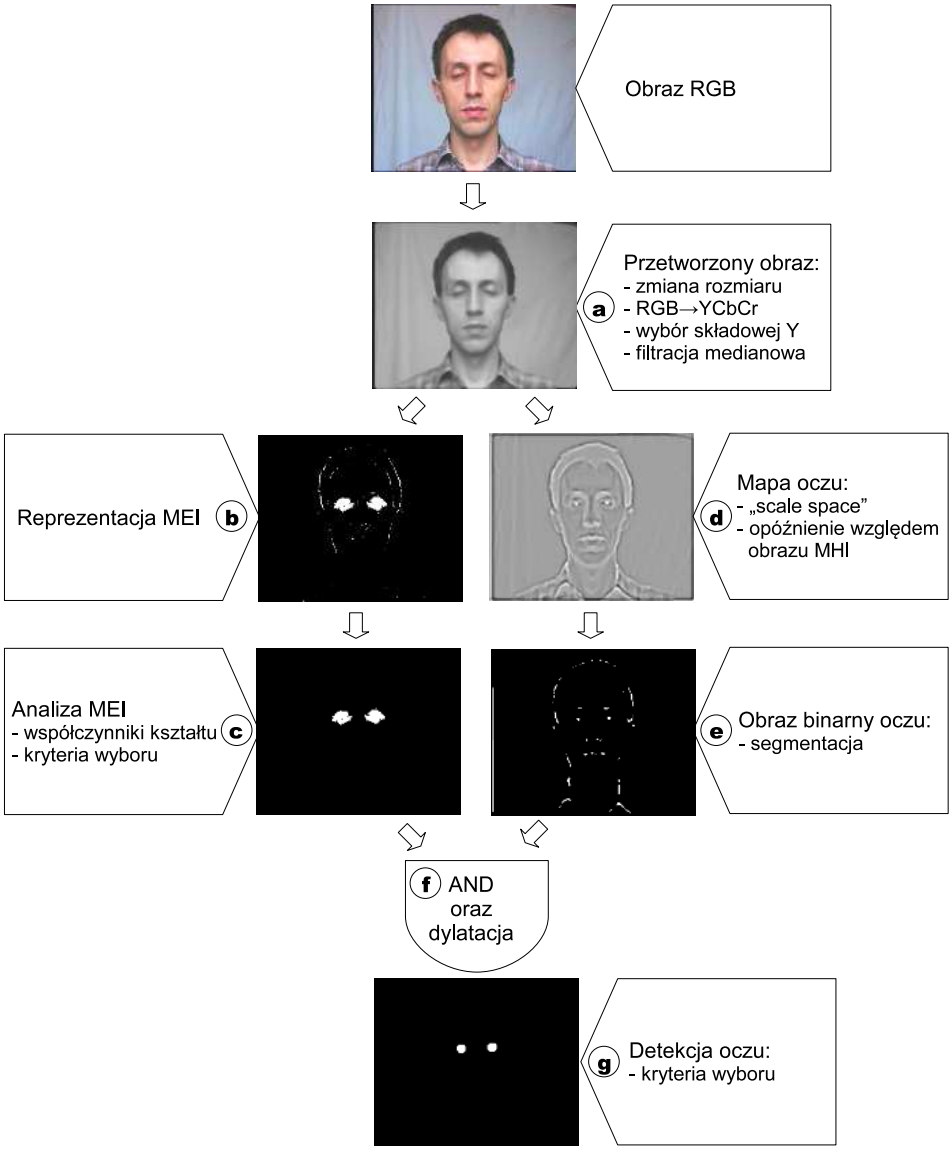
Ideę algorytmu detekcji mrugnięć przedstawia rysunek 8.18.

W pierwszej kolejności (a) następuje wstępne przetwarzanie obrazów — usunięcie szumów przy pomocy filtru medianowego oraz zmniejszenie rozmiaru obrazu. Następnie wyznaczana jest reprezentacja MEI (b) w wyniku czego otrzymywany jest binarny obraz. Zawiera on obiekty powstałe w wyniku zamykania/otwierania oczu oraz inne obiekty będące wynikiem zakłóceń (np. niewielkie drgania głowy powodują powstawanie długich wąskich obiektów będących zarysem głowy). W celu wyboru właściwych obiektów dokonywana jest analiza współczynników kształtu (c), polegająca na pozostawieniu obiektów spełniających następujące kryteria:

- Usuniecie obiektów których wielkość jest poza zadanym przedziałem — celem jest likwidacja drobnych zakłóceń (obiekty złożone z pojedynczych pikseli lub niewielkich grup) oraz wykrycie sytuacji dużego ruchu głową (obiekty o bardzo dużej ilości pikseli).

- Usunięcie obiektów dla których stosunek długiej osi elipsy opisanej na obiekcie do osi krótkiej mniejszy od zadanego progu — celem jest likwidacja dużych podłużnych obiektów odpowiadających zarysowi głowy.

Na podstawie przeprowadzonych badań stwierdzono, że po analizie współczynników kształtu zdarza się, że nadal pozostaje dużo obiektów nie odpowiadających oczom. Dlatego też dodano kolejny etap algorytmu, jakim jest wyznaczanie mapy prawdopodobieństwa oczu (d) z wykorzystaniem metody „detekcji plam" w reprezentacji skali (por. A.5). Możliwe również jest użycie mapy prawdopodobieństwa opartej o kolor, ponieważ w odróżnieniu od modelu koloru skóry, spełniony jest warunek iż obszary oczu charakteryzują się innymi wartościami składowych chrominancji Cb oraz Cr. W wyniku binaryzacji z odpowiednim progiem mapy oczu powstaje obraz, na którym widoczne są obiekty odpowiadające obszarom oczu (e).




Rysunek 8.18: Schemat algorytmu detekcji mrugnięć.

Obydwa binarne obrazy poddawane są operacji AND oraz dylatacji (f) przy czym obraz powstały z analizy „plam” jest opóźniony o kilka obrazów z sekwencji. Opóźnienie to wprowadzone zostało ponieważ przy zamkniętych oczach plamy są mniej widoczne niż przy otwartych. W końcowym etapie (g) następuje lokalizacja oczu poprzez wybór obiektów spełniających następujące kryteria:

- wybór dwóch obiektów o największej powierzchni,

- wielkości obydwu obiektów muszą być porównywalne,
- odległość między obiektami nie może być mniejsza lub większa niż zadany próg.

Powyższe kryteria pozwalają na prawidłową detekcję i lokalizację oczu na większości testowanych sekwencji video. Rezultaty zostały zamieszczone w dodatku G.

Analiza wyników wskazuje niewielką ilość błędów pierwszego rodzaju (wyniki fałszywie dodatnie – $FP < 1$), co jest istotne z punktu widzenia algorytmu kalibracji. Oznacza to, że ilość obrazów błędnie zaklasyfikowanych jako mrugnięcie a nie będących nimi, jest minimalna.

Dość duża ilość błędów drugiego rodzaju (wyniki fałszywie ujemne – $16 < FN < 37$) posiada mniejszy wpływ na procedurę kalibracji. W kalibracji bardziej istotne jest wykrycie samego faktu i momentu mrugnięcia, niż dokładne zliczenie ilość obrazów zaklasyfikowanych jako gest mrugania. Rysunki zamieszczone w dodatku G (rys. G.2 – G.5), obrazują przyczynę dużej ilość błędów drugiego rodzaju. Ze względu na specyfikę algorytmu MHI, pojedyncze mrugnięcie jest rozpoznawane jako dwa osobne gesty — zamykanie i otwieranie oczu. Może to być wykorzystane do odróżniania od siebie tych gestów.

Dla pozostałych przypadków błędnych detekcji, konieczna będzie w przyszłości rozbudowa algorytmu analizy o dodatkowe kryteria.

Aby ocenić dokładność lokalizacji oczu wyznaczony został średni błąd (8.7) dla wszystkich obrazów, dla których detekcja była prawidłowa (dla uproszczenia współrzędne obydwu oczu zostały zsumowane). Wyniki przedstawia tabela G.3. Na jej podstawie można stwierdzić że lokalizacja jest dokładna ($\varepsilon_{eye} < 9px$). Niewielkie rozbieżności są dopuszczalne i mogą wynikać z niedokładności ręcznego wskazania referencyjnego położenia oczu.

$$\varepsilon_{eye} = mean\left(\sqrt{\left(x^i_{eye} - x^i_{ref}\right)^2 + \left(y^i_{eye} - y^i_{ref}\right)^2}\right) \tag{8.7}$$

gdzie:

$\varepsilon_{eye}$ – średni błąd lokalizacji oczu
$x^i_{eye}, y^i_{eye}$ – wyznaczone współrzędne $x$ i $y$ oczu dla i-go obrazu
$x^i_{ref}, y^i_{ref}$ – referencyjne współrzędne $x$ i $y$ oczu dla i-go obrazu

### 8.4.2 Lokalizacja nozdrzy

Algorytmy rozpoznawania mimiki oparte na histogramach orientacji wymagają dokładnego pozycjonowania obszarów zainteresowań (ROI). Położenie ROI wyznaczone może być względem niezmiennych części twarzy — nozdrzy. Nozdrza są

elementem twarzy łatwym do zlokalizowania w przypadku gdy wcześniej zostało określone położenie oczu (detekcja mrugnięć).

Do lokalizacji nozdrzy wykorzystany może zostać fakt, iż stanowią one na obrazie ciemniejsze plamy. Stąd do ich detekcji dobrze nadaje się metoda „detekcji plam" w reprezentacji skali (ang. scale-space) opisana w dodatku A.5.

Algorytm detekcji nozdrzy jest następujący:

- Wstępne przetwarzanie obrazów: usunięcie szumów przy pomocy filtru medianowego, konwersja obrazu z przestrzeni RGB do obrazu w 256 odcieniach szarości.
- Pobranie z obrazu wycinka ROI zawierającego nozdrza (rys. 8.20a). Położenie ROI wyznaczane jest na podstawie znanej wcześniej (detekcja mrugnięć) lokalizacji oczu (rys. 8.19) oraz zależności anatomicznych twarzy.
- Wyznaczenie reprezentacji skali w celu uwypuklenia obrazu nozdrzy — (rys. 8.20b).
- Binaryzacja reprezentacji skali z progiem wyznaczonym automatycznie — wzór (8.8), rys. 8.20c.
- Indeksacja i analiza pozostałych na obrazie obiektów przy pomocy współczynników kształtu. Wyznaczenie współrzędnych nozdrzy — środki ciężkości pozostałych obiektów.

$$TH_{Snostril} = \mu_{Snostril} + 2 \cdot std_{Snostril} \tag{8.8}$$

gdzie:

$TH_{Snostril}$ – wartość progu binaryzacji reprezentacji skali powyżej którego piksel uznawany jest za należący do nozdrza

$\mu_{Snostril}$ – średnia pikseli reprezentacji skali obszaru ROI nozdrzy

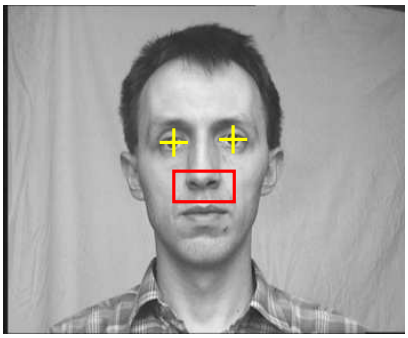
$std_{Snostril}$ – odchylenie standardowe pikseli reprezentacji skali obszaru ROI nozdrzy

W celu sprawdzenia skuteczności algorytmu detekcji nozdrzy, przeprowadzono testy dla tych samych sekwencji video, dla których realizowana była lokalizacja oczu. Na potrzeby testów lokalizacja oczu została wyznaczona ręcznie. Błąd lokalizacji nozdrzy został obliczony jako odległość euklidesowa pomiędzy wyznaczoną pozycją, a położeniem wskazanym przez człowieka (8.9). Rezultaty przedstawiają wykresy H.1 – H.4, zamieszczone w dodatku H.1.

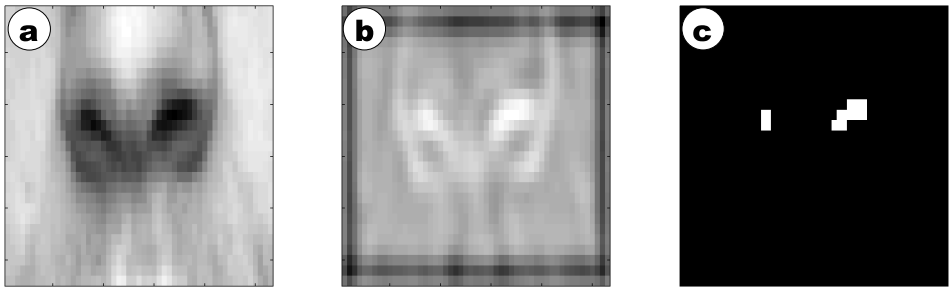
$$\varepsilon^i_{nostril} = \sqrt{\left(x^i_{nostril} - x^i_{ref}\right)^2 + \left(y^i_{nostril} - y^i_{ref}\right)^2} \tag{8.9}$$

gdzie:

$\varepsilon^i_{nostril}$ – błąd lokalizacji nozdrza dla i-go obrazu
$x^i_{nostril}, y^i_{nostril}$ – wyznaczone współrzędne $x$ i $y$ nozdrza dla i-go obrazu
$x^i_{ref}, y^i_{ref}$ – referencyjne współrzędne $x$ i $y$ nozdrza dla i-go obrazu



Rysunek 8.19: Przykładowy obraz wraz z pozycją oczu i ROI dla wyszukiwania nozdrzy.




Rysunek 8.20: Rezultaty kolejnych etapów algorytmu lokalizacji nozdrzy: (a) wycinek obrazu zawierający nozdrza, (b) obraz w reprezentacji skali (ang. scale-space), (c) mapa prawdopodobieństwa po binaryzacji

Analiza rezultatów lokalizacji potwierdza skuteczność algorytmu ($\varepsilon^i_{nostril} < 4px$). Tylko dla niewielu obrazów testowych błąd lokalizacji jest większy od zakładanego lub nozdrza nie zostały znalezione.

# Rozdział 9

# Podsumowanie

## 9.1 Główne rezultaty i wnioski

Postawione na początku niniejszej rozprawy cele zostały w pełni zrealizowane.

1. **Określenie sposobu pomiaru oraz precyzyjnego opisu elementów mimiki.** Przeprowadzono analizę istniejących sposobów opisu mimiki oraz informacji niesionych przez ten kanał komunikacji. Wybrano dwie metody opisu przydatne w kontekście rozpoznawania. Wnioski:

   - Istniejące sposoby opisu pozwalają na wybór elementów mimiki z uwzględnieniem psychofizjologicznych uwarunkowań człowieka (m.in. możliwości wykonania danego gestu przez daną osobę).
   - Odróżnianie mimiki spontanicznej od celowych, znaczących ruchów jest zadaniem spoczywającym na oprogramowaniu interfejsu.

2. **Określenie możliwości jakie oferuje mimika w kontekście typowych scenariuszy interakcji człowieka z maszyną i własności istniejących urządzeń. Wybór elementów mimiki przydatnych w komunikacji człowiek-maszyna.** Na podstawie analizy własności istniejących urządzeń wejściowych oraz poprzez uwzględnienie typowych scenariuszy interakcji, określono możliwości jakie oferuje mimika. Zaproponowano nowe sposoby sterowania komputerem oraz dokonano wyboru elementów mimiki potencjalnie przydatnych w komunikacji człowiek-maszyna. Wnioski:

   - Interfejs człowiek-komputer oparty o rozpoznawanie mimiki może udostępniać użytkownikowi bogatszy „alfabet", pozwalający na elastyczny wybór sposobu interakcji w zależności od umiejętności lub preferencji

człowieka. Nie każda osoba potrafi wykonać te same gesty — jest to sprawa bardzo indywidualna.

- Możliwość adaptacji interfejsu odgrywa szczególną rolę dla ludzi z porażeniem czterokończynowym, ponieważ mimika takich osób może różnić się znacznie od mimiki człowieka w pełni sprawnego ruchowo.

3. **Zdefiniowanie atrybutów elementów mimiki oraz odpowiadających im cech charakterystycznych na obrazie, pozwalających na rozpoznanie gestów mimicznych.** Zdefiniowane zostały podstawowe grupy atrybutów stanowiących abstrakcyjną reprezentację elementów mimiki. Rozpatrując mimikę w kontekście sterowania komputerem, wzięta pod uwagę została nie tylko statyczna reprezentacja elementów mimiki, do której należą atrybuty stałe (np. brwi) oraz zmienne (np. zmarszczki). Uwzględniono także dynamikę zmian wyglądu cech twarzy — atrybuty dynamiczne takie jak np. szybkość wykonania gestu. Określone zostały odpowiadające atrybutom cechy charakterystyczne (kształt, lokalne zmiany wyglądu, ruch) stanowiące ewidencję wystąpienia na obrazie danego atrybutu. Wnioski:

   - Atrybuty elementów mimiki stanowią abstrakcyjną ich reprezentację. Aby były one użyteczne do rozpoznawania, konieczne jest zgromadzenie ewidencji o wystąpieniu danego atrybutu na obrazie . Taką ewidencję stanowią cechy charakterystyczne — przykładowo, jeśli obiekt jest reprezentowany przez jego kształt, to na obrazie oczekuje się istnienia krawędzi odpowiadających konturowi obiektu.
   - Wybór cech jest istotnym zagadnieniem, od którego zależy skuteczność rozpoznawania oraz złożoność obliczeniowa algorytmów.

Opracowanie metody automatycznego rozpoznawania wybranych elementów mimiki wymagało zrealizowania następujących celów:

1. **Zaproponowanie struktury i elementów składowych systemu automatycznego rozpoznawania gestów mimicznych.** Rezultatem przeprowadzonych prac jest opracowana struktura systemu automatycznego rozpoznawania gestów mimicznych. Elementy składowe systemu to: selektywne przetwarzanie informacji, detekcja i lokalizacja twarzy, ekstrakcja cech charakterystycznych mimiki, klasyfikacja, generacja sygnałów sterujących oraz adaptacja systemu. Wśród istotnych rezultatów wymienić można opracowaną metodę doboru przestrzeni barw dla algorytmu segmentacji oraz algorytm lokalizacji twarzy. Wnioski:

   - W kontekście rozpoznawania obrazów przez maszynę selektywne przetwarzanie informacji ma istotne znaczenie, ponieważ pozwala na zawężenie obszaru poszukiwań, przyspieszenie obliczeń oraz uzyskanie dodatkowych wskazówek dla pozostałych elementów składowych systemu

(np. algorytmów rozpoznawania). Szczególnie istotne są informacje o kolorze (pozwalające na wstępną segmentację twarzy).

- Duży wpływ na skuteczność zaproponowanego algorytmu lokalizacji twarzy ma prawidłowe oświetlenie sceny. Szczególnie istotny jest odpowiedni wybór przestrzeni kolorów oraz algorytmu segmentacji. Istotne jest również, aby twarz człowieka była obserwowana w przybliżeniu frontalnie — przy większych ruchach głowy może się zdarzyć ze niektóre elementy (szczególnie oczy) nie będą widoczne.

2. **Opracowanie metod wyodrębniania z obrazu twarzy cech charakterystycznych, odpowiadających atrybutom rozpoznawanych elementów mimiki.** Opracowano i przetestowano trzy metody wyodrębniania cech charakterystycznych, odpowiadających atrybutom rozpoznawanych elementów mimiki. Wybrane metody reprezentują odmienne podejścia analizy obrazów. Są to: statystyczne modele kształtu (grupa metod opartych o ekstrakcję cech), histogramy orientacji (grupa metod bezpośredniej analizy obrazu) oraz detekcja ruchu (metody analizy uwzględniające kontekst czasowo-przestrzenny). Dla histogramów orientacji wykorzystano dwie reprezentacje obrazu — przestrzeń skali (ang. scale-space) oraz opartą o filtry kierunkowe Gabora. Wnioski:

   - Z perspektywy budowy interfejsu, który powinien adaptować się do użytkownika, wyznaczanie położenia punktów charakterystycznych dla statystycznych modeli kształtu powinno odbywać się automatycznie.
   - Detekcja ruchu może być wykorzystana do wstępnej segmentacji obszarów zainteresowań takich jak: okolice oczu, obszar głowy oraz miejsca gdzie występują ruchy mimiczne. Algorytm ten jest jednakże wrażliwy na zmiany oświetlenia.
   - W przypadku histogramów orientacji istotne jest precyzyjne określenie obszaru dla którego są one wyznaczane (ROI). Powinny to być obszary w których spodziewane są zmiany wynikające z ruchów mimicznych (zmarszczki, bruzdy, fałdy).

3. **Opracowanie metod rozpoznawania gestów mimicznych wykorzystujących wyodrębnione cechy. Badanie skuteczności algorytmów dla typowych sytuacji występujących podczas interakcji człowieka z maszyną.** Dokonano wyboru dwóch metod klasyfikacji (klasyfikator kNN oraz wielowymiarowa analiza dyskryminacyjna) pozwalających na rozpoznawanie wybranych elementów mimiki. Zaproponowano metodykę oceny skuteczności algorytmów rozpoznawania. Testy przeprowadzono niezależnie dla dwóch wybranych metod wyodrębniania cech, tj. statystycznych modeli kształtu oraz histogramów orientacji. W przypadku modeli kształtu badano

m.in. skuteczność rozróżniania gestów mimicznych górnej części twarzy dla jednej i wielu osób. Z kolei dla histogramów orientacji przeprowadzono testy rozpoznawania mimiki dolnej części twarzy. Sprawdzono wpływ wyboru obszaru zainteresowań (ROI) oraz rodzaju reprezentacji obrazu (scale-space, filtry Gabora) na skuteczność algorytmu. Badano również wpływ czynników zakłócających — zmiany skali rotacji oraz oświetlenia. Wnioski:

- Rozpoznawanie jednostek czynnościowych przy pomocy statystycznych modeli kształtów jest skuteczne. Istotną obserwacją jest stwierdzenie stosunkowo niewielkiej wrażliwości algorytmu rozpoznawania opartego o kształty na zmienność cech osobniczych (dobre rezultaty dla odróżniania kształtów elementów mimiki różnych osób).
- W przypadku histogramów orientacji, porównując reprezentacje obrazu lepsze rezultaty otrzymano dla histogramów opartych o gradienty w przestrzeni skali (ang. scale-space).
- Czynniki zakłócające takie jak zmiany oświetlenia, skali i rotacji twarzy, mają negatywny wpływ na skuteczność rozpoznawania dla obydwu reprezentacji (scale-space oraz filtrów Gabora). Mimo teoretycznej niewrażliwości na zmiany oświetlenia sceny, skuteczność rozpoznawania maleje. Największe błędy powodują zmiany rotacji twarzy. Algorytmy oparte na histogramach orientacji są również wrażliwe na zmienność cech osobniczych. Wymagają więc każdorazowej kalibracji w przypadku zmiany użytkownika lub warunków oświetlenia sceny.
- Spośród dwóch metod klasyfikacji, lepszym okazał sie klasyfikator kNN.
- Duży wpływ czynników zakłócających na rozpoznawanie w przypadku pojedynczych obrazów sekwencji video, może zostać ograniczony poprzez uwzględnienie kontekstu czasowego. Przykładowo — prawidłowa detekcja gestu wymaga wystąpienia kilku lub kilkunastu następujących po sobie obrazów, dla których otrzymano ten sam wyniki rozpoznania gestu.

4. **Usystematyzowanie czynników wpływających na skuteczność rozpoznawania oraz opracowanie metody adaptacji systemu do człowieka oraz zmieniających sie warunków otoczenia.** Usystematyzowano czynniki wpływające na skuteczność rozpoznawania elementów mimiki. Są to: zewnętrzne źródła zmienności (geometria sceny, oświetlenie, przysłanianie twarzy przez inne obiekty, szumy i zakłócenia) oraz wewnętrzne źródła zmienności (tożsamość osoby, wiek, płeć, indywidualne cechy osobnicze oraz sposób wykonywania gestów). Dokonano oceny wpływu parametrów kamery oraz zmian oświetlenia sceny na segmentację twarzy. Zaproponowano metodę automatycznej kalibracji istotną dla adaptacji systemu do

człowieka oraz zmieniających się warunków otoczenia. Opracowano szereg algorytmów składowych metody kalibracji — algorytm estymacji położenia głowy człowieka względem kamery, metoda automatyzacji tworzenia modelu barwy skóry, algorytm detekcji mrugnięć, algorytm lokalizacji nozdrzy. Rozszerzone opracowanie wybranych zagadnień adaptacji przedstawionych w rozprawie znajduje się w publikacjach autora. Wnioski:

- Podczas interakcji człowieka z maszyną występuje wiele czynników powodujących dużą zmienność wyglądu twarzy oraz elementów mimiki. Z tego powodu, kluczowym zagadnieniem staje sie konieczność adaptacji systemu do indywidualnych cech i umiejętności człowieka oraz zmieniających sie warunków otoczenia (ang. visual learning and adaptation).
- Uwzględnienie zewnętrznych źródeł zmienności (oświetlenie, pozycja...) powinno odbywać się w sposób ciągły podczas pracy systemu. Z kolei wpływ tożsamości, wieku, cech osobniczych może zostać uwzględniony rzadziej — na etapie kalibracji systemu.
- Wykorzystanie możliwości interakcji z użytkownikiem upraszcza procedurę kalibracji systemu i zwiększa skuteczność działania algorytmów.
- Algorytm estymacji położenia głowy może służyć również do lokalizacji twarzy oraz rozpoznawania następujących gestów mimicznych — AU51-56 (ruchy głowy).
- Metoda detekcji mrugnięć nadaje się również do rozpoznawania następujących gestów mimicznych — AU43-AU46 (zamkniecie oczu, mruganie, przymrużenie oka).

## 9.2 Kierunki dalszych badań

Przedstawione w rozprawie badania oraz ich rezultaty nie wyczerpują tematu wykorzystania elementów mimiki w interakcji człowiek-maszyna. Według autora wymagane są następujące dalsze prace — dla zachowania czytelności zostały one pogrupowane wg celów rozprawy:

1. **Określenie sposobu pomiaru oraz precyzyjnego opisu elementów mimiki.**
   - Uwzględnienie sposobów opisu mimiki związanych z emocjami człowieka np. MAX (ang. Maximally Discriminative Affect Coding System) lub FACSAID (ang. Facial Action Coding System Affect Interpretation Dictionary).

2. **Określenie możliwości jakie oferuje mimika w kontekście typowych scenariuszy interakcji człowieka z maszyną i własności istniejących urządzeń. Wybór elementów mimiki przydatnych w komunikacji człowiek-maszyna.**

    - W celu weryfikacji przeprowadzonej analizy i wyciągniętych wniosków konieczne jest dokonanie badań użyteczności zaproponowanych sposobów sterowania. W pracy element ten został pominięty, ponieważ wymaga implementacji całości systemu rozpoznawania mimiki. System taki powinien działać w czasie rzeczywistym umożliwiając przeprowadzenie badań na większej populacji ludzi.

3. **Zdefiniowanie atrybutów elementów mimiki oraz odpowiadających im cech charakterystycznych na obrazie, pozwalających na rozpoznanie gestów mimicznych.**

    - Określenie listy atrybutów oraz odpowiadających im cech charakterystycznych dla pełnego zestawu gestów mimicznych (nie tylko wybranych).

4. **Zaproponowanie struktury i elementów składowych systemu automatycznego rozpoznawania gestów mimicznych.** Rozbudowa algorytmu detekcji i lokalizacji twarzy poprzez:

    - Wykorzystanie dodatkowych informacji o ruchu do skuteczniejszej lokalizacji twarzy.
    - Wykorzystanie szerszego zestawu cech, widocznych również w przypadku położenia głowy innego niż frontalne, do lokalizacji twarzy.
    - Określenie dodatkowych kryteriów selekcji cech oraz użycie algorytmów śledzenia (ang. feature tracking).

5. **Opracowanie metod wyodrębniania z obrazu twarzy cech charakterystycznych, odpowiadających atrybutom rozpoznawanych elementów mimiki.**

    - Opracowanie metody automatycznego wyznaczania punktów charakterystycznych kształtów.
    - Badanie innych reprezentacji obrazu potencjalnie przydatnych dla algorytmu opartego o histogramy orientacji.
    - Opracowanie algorytmów pozwalających na wyodrębnianie cech dynamicznych mimiki takich jak trajektoria, szybkość wykonania gestu.

6. **Opracowanie metod rozpoznawania gestów mimicznych wykorzystujących wyodrębnione cechy. Badanie skuteczności algorytmów**

**dla typowych sytuacji występujących podczas interakcji człowieka z maszyną.**

- Rozbudowa algorytmu o rozpoznawanie pełnego zestawu gestów.
- Uniezależnienie algorytmu od wpływu czynników zakłócających — wykrywanie zmian skali i rotacji, automatyczna rekalibracja algorytmu i uczenie klasyfikatorów w trakcie pracy systemu.
- Wykorzystanie informacji czasowej do rozpoznawania gestów oraz algorytmów śledzenia elementów twarzy do estymacja parametrów ruchu głowy (rotacja, skala...).
- Testowanie innych metod klasyfikacji (sieci neuronowe, klasyfikatory bayesa). Poprawa jakości klasyfikacji poprzez wykorzystanie metod wzmacniania (ang. boosting) np. AdaBoost oraz doboru zmiennych diagnostycznych (np. analiza czynnikowa, metoda Helwinga).

7. **Usystematyzowanie czynników wpływających na skuteczność rozpoznawania oraz opracowanie metody adaptacji systemu do człowieka oraz zmieniających się warunków otoczenia.**

   - Rozbudowa, integracja oraz implementacja całości metody kalibracji pozwalającej na automatyczne: tworzenie bazy gestów mimicznych, tworzenie bazy wzorców twarzy, dobór parametrów algorytmów rozpoznawania, itp.
   - Zaproszenie większej grupy osób do badań, których celem jest określenie najwygodniejszych dla człowieka sposobów kalibracji systemu.
   - Rozbudowa i przystosowanie algorytmu estymacji położenia głowy do rozpoznawania gestów mimicznych. Podobnie dla algorytmu detekcji mrugnięć.

# Bibliografia


[1] P. Augustyniak. Adaptacja oprogramowania interpretacyjnego do stanu pacjenta i celów diagnostycznych. Krajowa Konferencja Biocybernetyka i inżynieria Biomedyczna, 2007. [cytowanie na str. 6, 49]

[2] P. Augustyniak and Z. Mikrut. Badanie reguł postrzegania naturalnego w celu ich wykorzystania w inteligentnych systemach wizyjnych. Krajowa Konferencja: Sztuczna Inteligencja w Inżynierii Biomedycznej, 2004. [cytowanie na str. 6]

[3] D.H. Ballard and C.M. Brown. Computer Vision (section 8.2.2). Prentice Hall Professional Technical Reference, 1982. [cytowanie na str. 38]

[4] M.S. Bartlett, G. Littlewort, B. Braathen, T.J. Sejnowski, and J.R. Movellan. A prototype for automatic recognition of spontaneous facial actions. Advances in Neural Information Processing Systems, 15:1295–1302, 2002. [cytowanie na str. 18]

[5] Y. Bellik and Burger D. The potential of multimodal interfaces for the blind : an exploratory study. 1995. [cytowanie na str. 5]

[6] M. Betke, J. Gips, and P. Fleming. The camera mouse: Visual tracking of body features to provide computer access for people with severe disabilities. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 10(1):1–10, Mar 2002. [cytowanie na str. 6, 39]

[7] I. Biederman. Human image understanding: recent research and a theory. In Papers from the second workshop on Human and Machine Vision II, volume 13, pages 13–57, San Diego, CA, USA, - 1986. Academic Press Professional, Inc. [cytowanie na str. 38]

[8] M.J. Black and A.D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In H. Burkhardt and B. Neumann, editors, European Conf. on Computer Vision, ECCV-98, volume 1406 of LNCS-Series, pages 909–924, Freiburg, Germany, - 1998. Springer-Verlag. [cytowanie na str. 40]

[9] A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schutte, and A. Wilson. The kids room: A perceptually-based interactive and immersive story environment. Technical Report 398, -, E15, 20 Ames Street, Cambridge, MA 02139, December 1999. [cytowanie na str. 70]

[10] A.F. Bobick and J.W. Davis. Action Recognition Using Temporal Templates (Chapter 6). - 1997. [cytowanie na str. 70]

[11] R.A. Bolt. „Put-that-there": Voice and gesture at the graphics interface. In SIGGRAPH '80: Proceedings of the 7th annual conference on Computer graphics and interactive techniques, pages 262–270, New York, NY, USA, 1980. ACM. [cytowanie na str. 5]

[12] A. Broniec. Reprezentacja ruchu i zamiaru ruchu w sygnale eeg oraz możliwości jej wykorzystania w interfejsie człowiek-maszyna, to appear. [cytowanie na str. 6]

[13] W. Buxton. A three-state model of graphical input. In D. Diaper et al. (Eds), Human-Computer Interaction - INTERACT '90 Amsterdam: Elsevier Science Publishers B.V. (North-Holland), pages 449–456, 1990. [cytowanie na str. 24]

[14] D. Chai and K.N. Ngan. Face segmentation using skin-color map in videophone applications. CirSysVideo, 9(4):551, June 1999. [cytowanie na str. 48, 49]

[15] T. Chen. Audio-visual integration in multi-modal communication. Proceedings of the IEEE 86, pages 837–852, 1998. [cytowanie na str. 33]

[16] T. Cootes and C. Taylor. Active shape models- - smart snakes. Proc. British Machine Vision Conference, page 266, 1992. [cytowanie na str. 39]

[17] C. Cortes and V. Vapnik. Support-vector networks. Machine Learning, 20(3):273–297, 1995. [cytowanie na str. 76]

[18] L. da F. Costa and R.M. Cesar J. Shape Analysis and Classification: Theory and Practice. CRC Press, Inc., Boca Raton, FL, USA, 2000. [cytowanie na str. 38]

[19] S. Dickinson. Object representation and recognition. 1999. [cytowanie na str. 38]

[20] R.O. Duda and P.E. Hart. Use of the hough transformation to detect lines and curves in pictures. Commun. ACM, 15(1):11–15, 1972. [cytowanie na str. 38]

[21] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. Wiley-Interscience Publication, 2000. [cytowanie na str. 78]

[22] P. Ekman. About brows: Emotional and conversational signals. In J. Aschoff, M. von Carnach, K. Foppa, W. Lepenies, & D. Plog (Eds.) Human ethology. Cambridge: Cambridge University Press, 1979. [cytowanie na str. 14]

[23] P. Ekman and W.V. Friesen. Unmasking the face. A guide to recognizing emotions from facial clues. Englewood Cliffs, New Jersey: Prentice-Hall., 1975. [cytowanie na str. 14]

[24] P. Ekman and W.V. Friesen. Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists, pages –, 1978. [cytowanie na str. 15]

[25] P. Ekman, J. Hager, CH. Methvin, and W. Irwin. Ekman-hager facial action exemplars (unpublished data). Human Interaction Laboratory, Univ. of California, 1978. [cytowanie na str. 17]

[26] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. Pattern Recognition, 36(1):259–275, 2003. [cytowanie na str. 44, 45, 64]

[27] J.D. Foley, V.L. Wallace, and P. Chan. The human factors of computer graphics interaction techniques. IEEE Computer Graphics and Applications, 4 (11):13–48, 1984. [cytowanie na str. 22]

[28] W. Freeman. Orientation histogram for hand gesture recognition. In Int'l Workshop on Automatic Face- and Gesture-Recognition, 1995. [cytowanie na str. 39, 67]

[29] S. Gong, S.J. McKenna, and A. Psarrou. Dynamic Vision: From Images to Face Recognition. Imperial College Press, London, UK, UK, 2000. [cytowanie na str. 87]

[30] T. Goto, M. Escher, Ch. Zanardi, and N. Magnenat-Thalmann. Mpeg-4 based animation with face feature tracking. In Computer Animation and Simulation 1999, pages 89–98, - -. [cytowanie na str. 20]

[31] S. Grossberg. How does the cerebral cortex work? [cytowanie na str. 46]

[32] M. Hachet, J. Pouderoux, and P. Guitton. A camera-based interface for interaction with mobile handheld computers. In I3D'05 - Symposium on Interactive 3D Graphics and Games., 2005. [cytowanie na str. 33]

[33] B.V. Hancock and A.M Burton. Human face perception and identification. Face Recognition: From Theory to Applications, Berlin: Springer, pages 51–72, 1998. [cytowanie na str. 56]

[34] M. Held. Voronoi diagrams and offset curves of curvilinear polygons. In Computer-Aided Design, pages 287–300, 30(4) 1998. [cytowanie na str. 38]

[35] K. Hinckley. Input technologies and techniques. The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications, pages 151–168, 2003. [cytowanie na str. 22]

[36] K. Hinckley, R. Jacob, and C. Ware. Input/output devices and interaction techniques. In CRC Computer Science and Engineering Handbook, A. B. Tucker, ed. CRC Press LLC: Boca Raton, FL., to appear. (Microsoft). [cytowanie na str. 23]

[37] E. Hjelmas and B.K. Low. Face detection: A survey. Computer Vision and Image Understanding, 83(3):236–274, Sept. 2001. [cytowanie na str. 44, 48, 55]

[38] Rein-Lien Hsu, M. Abdel-Mottaleb, and A.K. Jain. Face detection in color images. IEEE Trans. Pattern Anal. Mach. Intell., 24(5):696–706, 2002. [cytowanie na str. 48, 49, 56, 57]

[39] J. Ilonen, J.K. Kamarainen, and H. Kalviainen. Efficient computation of gabor features. Research Report 100, Lappeenranta University of Technology, Department of Information Technology, pages –, 2005. [cytowanie na str. 68]

[40] C.E. Izard. The maximally discriminative facial movement coding system max. Unpublished manuscript available from Instructional Resource Center, University of Delaware, 1979. [cytowanie na str. 14]

[41] M. Jabłoński, J. Przybyło, and P. Wołoszyn. Automatyczna segmentacja twarzy dla potrzeb interfejsu człowiek-komputer. AGH Automatyka, 9/3:587–600, 2005. [cytowanie na str. 48, 101]

[42] W. Jakuczun. Lokalne klasyfikatory jako narzędzie analizy i klasyfikacji sygnałów. PhD thesis, IPI PAN, promotor: dr hab. Jerzy Cytowski, 2006. [cytowanie na str. 76]

[43] A. Jozwik, S. Serpico, and F. Roli. A parallel network of modified 1-nn and k-nn classifiers: Application to remote sensing image classification. PRL, 19(1):57–62, January 1998. [cytowanie na str. 77]

[44] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. The 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), pages –, March 2000. [cytowanie na str. 17]

[45] A. Kapoor, Y. Qi, and W. Picard-Rosalind. Fully automatic upper facial action recognition. Analysis and Modeling of Faces and Gestures, pages 195–202, Oct. 2003. [cytowanie na str. 17]

[46] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. International Journal of Computer Vision, 1:321–331, 1988. [cytowanie na str. 38]

[47] Y. Keselman and S. Dickinson. Bridging the representation gap between models and examplars. In IEEE Conf. on Comp. Soc. Work. on Models versus Exemplars in Comp. Vis., 2001. [cytowanie na str. 38]

[48] K. Koffka. Principles of gestalt psychology (International library of psychology, philosophy, and scientific method) (International library of psychology, philosophy, and scientific method). Routledge & K. Paul, January 1962. [cytowanie na str. 47]

[49] J. Koronacki and J. Ćwik. Statystyczne systemy uczące się. Warszawa: Wydawnictwa Naukowo-Techniczne, 2005. [cytowanie na str. 79]

[50] C. Kotropoulos and I. Pitas. Rule-based face detection in frontal views. In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 97), IV:2537–2540, April 21-24 1997. [cytowanie na str. 55]

[51] C. Kouadio, P. Poulin, and P. Lachapelle. Real-time facial animation based upon a bank of 3d facial expressions. In CA '98: Proceedings of the Computer Animation, page 128, Washington, DC, USA, - 1998. IEEE Computer Society. [cytowanie na str. 20]

[52] A. Kuriański. Detekcja i śledzenie ruchu przy użyciu czasowo-przestrzennego modelowania obrazów za pomocą pól losowych. PhD thesis, IPPT PAN, 1992. [cytowanie na str. 69, 70]

[53] F. Lavagetto and R. Pockaj. The facial animation engine : towards a high-level interface for the design of mpeg-4 compliant animated faces. IEEE Transactions on Circuits and Systems for Video Technology, pages –, 1998. [cytowanie na str. 18]

[54] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. J. of Applied Statistics, 21(2):224–270, 1994. (Supplement on Advances in Applied Statistics: Statistics and Images: 2). [cytowanie na str. 140]

[55] Ch.L. Lisetti and D.J. Schiano. Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect. Pragmatics and Cognition (Special Issue on Facial Information Processing: A Multidisciplinary Perspective), 8(1):185–235, 2000. [cytowanie na str. 7, 22]

[56] J. Lombardi and M. Betke. A self-initializing eyebrow tracker for binary switch emulation. 2002. [cytowanie na str. 70]

[57] J.M. Marnik. Rozpoznawanie znaków Polskiego Alfabetu Palcowego z wykorzystaniem morfologii matematycznej i sieci neuronowych. PhD thesis, AGH Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki, 2002. [cytowanie na str. 48]

[58] L. Mazurek. Modelowanie początkowych etapów przetwarzania informacji wzrokowej. PhD thesis, AGH, wydz. EAIiE, promotor: dr inż. Zbigniew Mikrut, 2001. [cytowanie na str. 44]

[59] S.J. McKenna and S. Gong. Real-time face pose estimation. Real-Time Imaging, 4(5):333–347, 1998. [cytowanie na str. 91, 94]

[60] S.J. McKenna, S.G. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. PR, 31(12):1883–1892, December 1998. [cytowanie na str. 49]

[61] Z. Mikrut and P. Augustyniak. Lokalizacja kluczowych cech rozpoznawanych obiektów na podstawie analizy sciezki wzrokowej obserwatora. Bio-Algorithms and Med-Systems, 1(1/2):307–310, 2005. [cytowanie na str. 6]

[62] MPEG-4. ISO/IEC MPEG-4 Part 2 (Visual). [cytowanie na str. 18, 19]

[63] S. Oviatt. Multimodal interfaces. In Handbook of Human-Computer Interaction. J. Jacko and A. Sears (Eds.) Mahwah NJ Lawrence Erlbaum), 2002. [cytowanie na str. 5]

[64] M. Pantic, P. Ioannis, and M. Valstar. Learning spatiotemporal models of facial expressions. Proceedings of International Conf. Measuring Behaviour, pages 7–10, September 2005. [cytowanie na str. 17]

[65] P. Pawlik, D. Iwaniec, and M. Iwaniec. Analiza obrazu z kamery jako podstawa interfejsu człowiek niepełnosprawny-komputer. AGH Automatyka, 10/3:399–405, 2006. [cytowanie na str. 39]

[66] P. Peer, J. Kovac, and F. Solina. Human skin colour clustering for face detection. 1998. [cytowanie na str. 49]

[67] D.A. Pollen and S.F. Ronner. Phase relationships between adjacent simple cells in the visual cortex. Science, 212:1409–1411, jun 1981. [cytowanie na str. 39]

[68] J. Przybyło. Śledzenie cech charakterystycznych twarzy w systemie rozpoznawania mimiki. AGH Automatyka, 11/3:257–266, 2007. [cytowanie na str. 107]

[69] J. Przybyło, M. Jabłoński, and P. Wołoszyn. Detekcja markerów dla celów automatycznej anotacji mimiki twarzy. AGH Automatyka, 10/3:413–425, 2006. [cytowanie na str. 106]

[70] J. Przybyło, P. Wołoszyn, and M. Jabłoński. Wizyjny interfejs człowiek-komputer przeznaczony dla użytkowników niepełnosprawnych. AGH Automatyka, 7/3:385–398, 2003. [cytowanie na str. 107]

[71] J. Przybyło, P. Wołoszyn, and M. Jabłoński. Rozpoznawanie jednostek czynnościowych mimiki twarzy na potrzeby interfejsu człowiek-komputer. AGH Automatyka, 8/3:378–379, 2004. [cytowanie na str. 65, 80]

[72] P. Saint-Marc and G. Medioni. B-spline contour representation and symmetry detection. In ECCV 90: Proceedings of the first european conference on Computer vision, pages 604–606, New York, NY, USA, - 1990. Springer-Verlag New York, Inc. [cytowanie na str. 38]

[73] K.R. Scherer and P. Ekman. The New Handbook of Methods in Nonverbal Behavior Research. Cambridge, UK: Cambridge University Press, 1982. [cytowanie na str. 16]

[74] K. Schwerdt and J.L. Crowley. Robust face tracking using color. In AFGR00, pages 90–95, 2000. [cytowanie na str. 48]

[75] N. Sebe, M.S. Lew, I. Cohen, A. Garg, and T.S. Huang. Emotion recognition using a cauchy naive bayes classifier. In ICPR02, volume I, pages 17–20, 2002. [cytowanie na str. 76]

[76] C. Sheng and Y. Xin. Shape-based image retrieval using shape matrix. In International Journal Of Signal Processing, volume 1, 2004. [cytowanie na str. 38]

[77] W. Skarbek. Segmentation of Colour Images. PAN Warszawa, 1994. [cytowanie na str. 49]

[78] M. Soriano, B. Martinkauppi, S. Huovinen, and M. Laaksonen. Skin detection in video under changing illumination conditions. In ICPR00, volume I, pages 839–842, 2000. [cytowanie na str. 49]

[79] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In CVPR99, volume II, pages 246–252, - 1999. [cytowanie na str. 69]

[80] M. Stegmann. Active appearance models: Theory and cases, 2000. [cytowanie na str. 65]

[81] R. Tadeusiewicz. Wykorzystanie sieci neuronowych do analizy sygnałów akustycznych w powiązaniu z głosowym sterowaniem robota dydaktycznego. Raport Instytutu Automatyki, 15, 1991. [cytowanie na str. 6]

[82] R. Tadeusiewicz. Sieci neuronowe. Problemy Wspolczesnej Nauki i Techniki. Informatyka. Akademicka Oficyna Wydaw. RM, 1993. [cytowanie na str. 39, 76]

[83] R. Tadeusiewicz and M. Flasiński. Rozpoznawanie obrazów. Problemy Wspolczesnej Nauki i Techniki. Informatyka. Akademicka Oficyna Wydaw. RM, 1993. [cytowanie na str. 75]

[84] R. Tadeusiewicz and P. Korohoda. Komputerowa analiza i przetwarzanie obrazów. Wydawnictwo Fundacji Postępu Telekomunikacji, 1997. [cytowanie na str. 44]

[85] Y-L. Tian, L. Brown, A. Hampapur, S. Pankanti, and R.M. Bolle. Real world real-time automatic recognition of facial expressions. In IEEE workshop on performance evaluation of tracking and surveillance, Graz, Austria, March 31 2003. [cytowanie na str. 39]

[86] Y.L. Tian, T. Kanade, and J.F. Cohn. Dual-state parametric eye tracking. Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), pages 110 – 115, 2000. [cytowanie na str. 16]

[87] Y.L. Tian, T. Kanade, and J.F. Cohn. Robust lip tracking by combining shape, color and motion. Proceedings of the 4th Asian Conference on Computer Vision (ACCV'00), pages –, January 2000. [cytowanie na str. 16]

[88] Y.L. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, (23(2)):97–116, 2001. [cytowanie na str. 16]

[89] Y.L. Tian, T. Kanade, and J.F. Cohn. Evaluation of gaborwavelet -based facial action unit recognition in image sequences of increasing complexity. In Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition (FG'02), DC, May 2002. [cytowanie na str. 39, 68]

[90] F.de la Torre, J. Campoy, Z. Ambadar, and J.F. Cohn. Temporal segmentation of facial behavior. In International Conference on Computer Vision, October 2007. [cytowanie na str. 76]

[91] M. Turk and A. Pentland. Eigenfaces for recognition. J. Cognitive Neuroscience, 3(1):71–86, 1991. [cytowanie na str. 39, 56]

[92] M. Valstar, P. Ioannis, and M. Pantic. Facial action unit recognition using temporal templates. Proceedings of IEEE Int'l Workshop on Human-Robot Interaction, pages 253–258, September 2004. [cytowanie na str. 17, 40, 75]

[93] V. Vezhnevets, V. Sazonov, and A. Andreeva. A survey on pixel-based skin color detection techniques. in Proc. Graphicon-2003, 2003. [cytowanie na str. 50, 53]

[94] M. Wertheimer. Laws of Organization in Perceptual Forms - A Source Book of Gestalt Psychology. W.D. Ellis ed. Routledge and Kegan Paul Ltd., -, 1955. [cytowanie na str. 47]

[95] M. Yang and N. Ahuja. Detecting human faces in color images. In ICIP-A 98, pages 127–130, 1998. [cytowanie na str. 50]

[96] Ming-Hsuan Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: a survey. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(1):34–58, 2002. [cytowanie na str. 55]

[97] P. Yao, G. Evans, and A.D. Calway. Face tracking and pose estimation using affine motion parameters. In SCIA01, pages O–Th2, 2001. [cytowanie na str. 90]

[98] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. n 3rd International Conference on Automatic Face and Gesture Recognition, I:336–341, 1998. [cytowanie na str. 76]

[99] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. ACM Comput. Surv., 35(4):399–458, 2003. [cytowanie na str. 44, 56]

# Dodatki

# Dodatek A

# Opis metod i algorytmów

W dodatku przedstawiono opis algorytmów oraz rozszerzono wybrane zagadnienia poruszane w rozprawie.

## A.1 Przestrzenie barw

Znormalizowana przestrzeń RGB posiada korzystną właściwość — niezmienniczość względem zmian orientacji powierzchni w stosunku do położenia źródła światła (przy założeniu powierzchni matowej oraz jednolitego światła otaczającego scenę (ang. ambient light). Jest to szczególnie istotne w przypadku detekcji twarzy, gdzie często występuje sytuacja zmian oświetlenia sceny. Znormalizowany model RGB powstaje poprzez normalizację składowych RGB (A.1). Trzecia składowa $I_b$ nie niesie żadnych dodatkowych informacji ponieważ można ją wyznaczyć na podstawie pozostałych ($I_r + I_g + I_b = 1$). Ponadto do mianownika zazwyczaj dodaje się stałą 1 aby uniknąć dzielenia przez zero dla pikseli o wartości RGB=0.

$$I_r = \frac{I_R}{I_R + I_G + I_B}, \;\; I_g = \frac{I_R}{I_R + I_G + I_B} \tag{A.1}$$

gdzie:

$I_r, I_g$ – Składowe znormalizowanego obrazu RGB
$I_R, I_G, I_B$ – Składowe obrazu RGB

Model HSV nawiązuje do sposobu, w jakim widzi ludzki narząd wzroku, gdzie poszczególne składowe nie są skojarzone bezpośrednio z przedziałami widma obrazu, lecz korespondują z subiektywną percepcją promieniowania widzialnego przez ludzkie oko — odcień, nasycenie oraz ton. Numerycznie odpowiadają

im wartości: barwa (ang. hue), nasycenie (ang. saturation), odcień (ang. value). Transformację RGB do HSV opisuje równanie (A.2):

$$
\begin{aligned}
I_H &= \arccos\left(\frac{0.5\cdot((I_R-I_G)+(I_R-I_B))}{\sqrt{\left((I_R-I_G)^2+(I_R-I_B)\cdot(I_G-I_B)\right)}}\right) \\
I_S &= 1-3\frac{\min(I_R,I_G,I_B)}{I_R+I_G+I_B} \\
I_V &= \frac{1}{3}(I_R+I_G+I_B)
\end{aligned}
\tag{A.2}
$$

gdzie:

$I_H, I_S, I_V$ – Składowe obrazu HSV
$I_R, I_G, I_B$ – Składowe obrazu RGB

Do głównych zalet przestrzeni HSV należy m.in. odseparowanie chrominancji (składowe H i S) od luminancji (V). Ponadto składowa H jest w dużym stopniu niezmiennicza względem zmian orientacji powierzchni w stosunku do położenia źródła światła (jak dla normalizowanej RGB) oraz oświetlenie światłem białym. Ograniczeniem analizy w przestrzeni HSV jest duża złożoność obliczeniowa transformacji RGB do HSV. Ma to szczególne znaczenie ze względu na wymóg szybkości przetwarzania ramek obrazu oraz udostępnienia określonej mocy obliczeniowej procesora dla znacznie bardziej wymagających algorytmów rozpoznawania i analizy twarzy.

Znacznie prostsza obliczeniowo jest transformacja RGB do przestrzeni YCbCr (A.3). Przestrzeń ta wykorzystywana jest w standardach przesyłania europejskiej telewizji i charakteryzuje się odseparowaniem chrominancji i luminancji.

$$
\begin{aligned}
I_Y &= 0.299\cdot I_R + 0.587\cdot I_G + 0.114\cdot I_B \\
I_{Cr} &= I_R - I_Y \\
I_{Cb} &= I_B - I_Y
\end{aligned}
\tag{A.3}
$$

gdzie:

$I_Y, I_Cr, I_Cb$ – Składowe obrazu YCbCr
$I_R, I_G, I_B$ – Składowe obrazu RGB

## A.2 Metody segmentacji twarzy

Segmentacja poprzez bezpośrednie określenie zestawu reguł, realizowana jest najczęściej z użyciem binaryzacji wybranych składowych przestrzeni kolorów. Jej

zaletą jest prostota i szybkość działania, okupiona koniecznością precyzyjnego określenia progów na podstawie analizy rozkładu kolorów pikseli obiektu. Przyjęty zestaw reguł, pozwala na segmentację twarzy w różnych przestrzeniach barw, przedstawiają równania: (A.4), (A.5), (A.6). W przypadku modeli YCbCr oraz znormalizowanego RGB wykorzystano dwie składowe, natomiast w przestrzeni HSV wystarczający okazał się wybór jednej składowej H.

$$BW_{r,g} = \begin{cases} 1; & (r_{low} < I_r < r_{high}) \bigcap (g_{low} < I_g < g_{high}) \\ 0; & w\, przeciwnym\, przypadku \end{cases} \tag{A.4}$$

gdzie:

$BW_{r,g}$ – binarna mapa przynależności pikseli do twarzy dla przestrzeni r-g
$I_r, I_g$ – Składowe znormalizowanego obrazu RGB
$r_{low}, r_{high}, g_{low}, g_{high}$ – progi binaryzacji składowych r-g

$$BW_{hue} = \begin{cases} 1; & (H_{low} < I_H < H_{high}) \\ 0; & w\, przeciwnym\, przypadku \end{cases} \tag{A.5}$$

gdzie:

$BW_{Hue}$ – binarna mapa przynależności pikseli do twarzy dla przestrzeni HSV
$I_H$ – Składowea Hue przestrzeni HSV
$H_{low}, H_{high}$ – progi binaryzacji składowej Hue

$$BW_{Cb,Cr} = \begin{cases} 1; & (Cb_{low} < I_{Cb} < Cb_{high}) \bigcap (Cr_{low} < I_{Cr} < Cr_{high}) \\ 0; & w\, przeciwnym\, przypadku \end{cases} \tag{A.6}$$

gdzie:

$BW_{Cb,Cr}$ – binarna mapa przynależności pikseli do twarzy dla przestrzeni YCbCr
$I_{Cb}, I_{Cr}$ – Składowe obrazu YCbCr
$Cb_{low}, Cb_{high}$ – progi binaryzacji składowej Cb
$Cr_{low}, Cr_{high}$ – progi binaryzacji składowej Cr

Do metod nieparametrycznych zaliczyć można: znormalizowane tablice przekodowań (LUT – ang. look-up-table), klasyfikatory bayesowskie, samoorganizujące się sieci neuronowe SOM (ang. self organizing map). Ich zaletą jest szybkość

uczenia modelu i działania oraz (raportowana w doniesieniach) niewrażliwość na kształt rozkładu statystycznego kolorów pikseli. Ograniczeniem są duże wymagania pamięciowe — szczególnie w przypadku wielowymiarowych tablic LUT.

Tablice LUT wyznaczane są następująco. Dla wybranych (ręcznie lub metodami automatycznymi) fragmentów obrazu zawierających piksele skóry, wyznaczany jest histogram. Histogram liczony jest dla jednej, dwóch lub trzech składowych przyjętej przestrzeni barw. Tablicę LUT otrzymuje się poprzez odpowiednią normalizację otrzymanego histogramu, formując w ten sposób dyskretną funkcję prawdopodobieństwa (A.7). W kategoriach statystycznych $P_{skin}(c)$ może być traktowane jako warunkowe prawdopodobieństwo obserwacji danego koloru, przy założeniu iż piksel należy do obszaru skóry $P_{skin}(c) \equiv P(c|skin)$.

$$P_{skin}(c) = \frac{h_{skin}(c)}{\sum\limits_{i} h_{skin}(c)}, \quad lub \;\; P_{skin}(c_1, c_2) = \frac{h_{skin}(c_1, c_2)}{\sum\limits_{i,j} h_{skin}(c_1, c_2)} \tag{A.7}$$

gdzie:

$P_{skin}(c), P_{skin}(c_1, c_2)$ – mapa prawdopodobieństwa skóry
$h_{skin}(c), h_{skin}(c_1, c_2)$ – histogramy wybranych składowych barw
$c, c_1, c_2$ – zakresy wartości (słupków histogramu) danej składowej

Wśród metod parametrycznych wyróżnić należy modelowanie rozkładu kolorów przy pomocy eliptycznej gaussowskiej funkcji rozkładu prawdopodobieństwa (A.8). Wykorzystuje się fakt, iż piksele należące do skóry tworzą w przestrzeni barw zwarty, niewielki obszar (rys. 5.8a).

$$P(c|skin) = \frac{1}{2 \cdot \pi \cdot |\Sigma_S|^{0.5}} \cdot e^{-\frac{1}{2}(\xi - \mu_S)^T \cdot \Sigma_S^{-1} \cdot (\xi - \mu_S)} \tag{A.8}$$

gdzie:

$\xi$ – wektor wybranych składowych koloru
$\Sigma_S, \mu_S$ – kowariancja oraz wartość oczekiwana obszaru należącego do skóry

Wartości $P(c|skin)$mogą być użyte do bezpośredniego określenia „jak bardzo podobny do skóry jest dany piksel" bądź też można wyznaczyć miarę błędu — odległość Mahalanobisa (A.9). Odległość Mahalanobisa jest odległością między dwoma punktami w n-wymiarowej przestrzeni, która różnicuje wkład poszczególnych składowych oraz wykorzystuje korelacje między nimi. Więcej informacji na ten temat można znaleźć w podręcznikach do matematyki i statystyki.

$$\lambda_S(\xi) = (\xi - \mu_S)^T \cdot \Sigma_S^{-1} \cdot (\xi - \mu_S) \tag{A.9}$$

gdzie:

$\lambda_S(\xi)$ – odległość Mahalanobisa wektora wybranych składowych koloru od modelu skóry

$\xi$ – wektor wybranych składowych koloru

$\Sigma_S, \mu_S$ – kowariancja oraz wartość oczekiwana obszaru należącego do skóry

W wielu przypadkach podstawowe założenie powyższej metody, czyli rozkład normalny pikseli należących do skóry, nie jest spełniony (rys. 5.8c). W takim przypadku użyteczne jest modelowanie barwy skóry przy pomocy mieszanki rozkładów prawdopodobieństwa (ang. multiple gaussian) i wykorzystanie algorytmu maksymalizacji wartości oczekiwanej EM (ang. expectation-maximization algorithm), do estymacji parametrów modelu opisanego równaniem (A.5).

## A.3 Estymacja i usuwanie szumów z sekwencji video

Pomiar szumów zrealizowano dla kilku sekwencji video pobranych w różnych warunkach i przy pomocy różnych kamer. Opis sekwencji znajduje się w rozdziale 5.2.2.

Ilościowy pomiar szumów zrealizowano następująco. Z sekwencji wideo wybrano kilka kolejnych obrazów, dla których nie występują ruchy głowy, gesty mimiczne lub zmiany oświetlenia sceny. Następnie wyznaczono znormalizowaną sumę obrazu różnicowego z kolejnych ramek (A.10). Jako miarę stopnia zaszumienia sekwencji przyjęto średnią z kolejnych wartości znormalizowanej sumy dla kilku obrazów (A.11).

$$Noise(k) = \frac{\sum_{m,n} |I(x,y,k) - I(x,y,k-1)|}{M_{IM} \cdot N_{IM}} \tag{A.10}$$

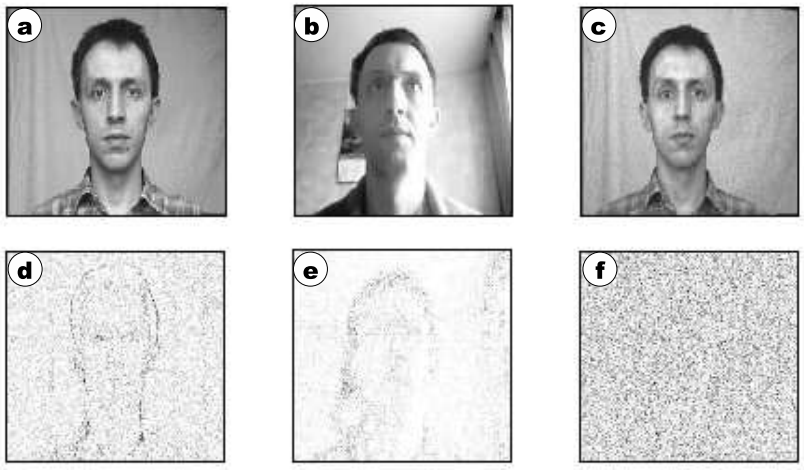
gdzie:

$M_{IM}, N_{IM}$ – wymiary obrazu (wiersz, kolumna)

$I(x,y,k)$ – Jasność piksela obrazu o współrzędnych $x, y$ w chwili $k$

$$Noise = mean(Noise(k)) \tag{A.11}$$

Rysunek A.1 przedstawia obrazy różnicowe dla wybranych sekwencji, natomiast tabela A.1 średni poziom szumów. Można zauważyć występowanie szumów — szczególnie widoczne w przypadku złych warunków akwizycji obrazów. Jest to



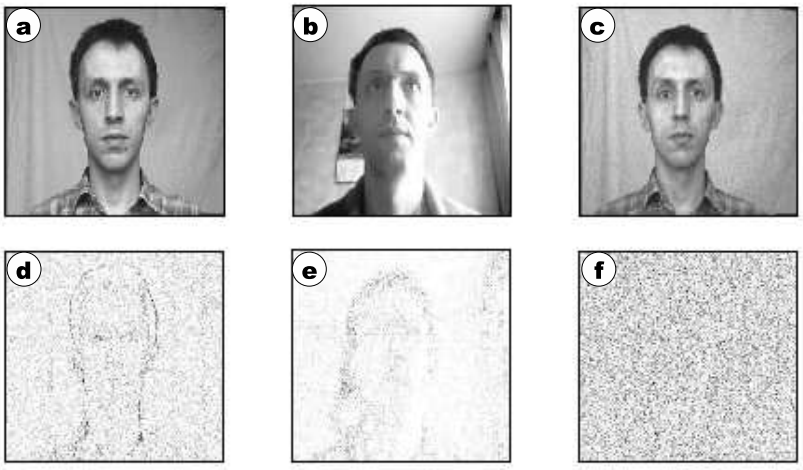

Rysunek A.1: Przykładowe obrazy z różnych sekwencji (a) sekwencja nr 1 (b) sekwencja nr 3 (c) sekwencja nr 2, oraz odpowiadające im moduły różnicy dwóch kolejnych obrazów (d)(e)(f).

Tabela A.1: Średni poziom szumów dla sekwencji z rys. A.1

| | $\mu_{noise}$ |
|---|---|
| **Sekwencja nr 1** | 2.0086 |
| **Sekwencja nr 3** | 1.4347 |
| **Sekwencja nr 2** | 2.9932 |

spodziewany fakt, ponieważ metody różniczkowania numerycznego (do których można zaliczyć obliczanie obrazu różnicowego) są wrażliwe na szum.

Otrzymane wyniki uzasadniają konieczność poprawy jakości obrazów i zmniejszenia zakłóceń. W tym celu można próbować odpowiednio dobrać parametry akwizycji obrazów (o ile kamera pozwala na regulacje swoich parametrów) lub dostosować warunki akwizycji (oświetlenie). Narzuca to dodatkowe wymagania pracy całego systemu, trudne do zagwarantowania gdy celem jest rozpoznawanie mimiki w nieznanych warunkach. Innym sposobem, wykorzystanym w niniejszej pracy, jest programowa redukcja szumów. W tym celu wykorzystano filtrację medianową, która skutecznie zmniejsza lokalne szumy nie powodując ich „rozmywania" na większym obszarze (w odróżnieniu od filtrów uśredniających). Dzięki temu usunięta została znaczna część zakłóceń z sekwencji (rys. A.2), tabela A.2), jednocześnie bez wprowadzania większych zmian na obrazach które mogłyby utrudnić np. detekcję ruchu.




Rysunek A.2: Wyniki redukcji szumów przy pomocy filtracji medianowej (maska o rozmiarze 5x5) dla sekwencji nr 3 — (a) wybrany obraz z sekwencji, (b) moduł różnicy dwóch kolejnych obrazów przed filtracją, (c) moduł różnicy dwóch kolejnych obrazów po filtracji

Tabela A.2: Średni poziom szumów dla sekwencji nr 3, przed i po filtracji medianowej

| | $\mu_{noise}$ |
|---|---|
| **przed filtracją** | 1.4347 |
| **po filtracji** | 0.97253 |

## A.4 Metoda PCA

Metoda składowych głównych PCA (ang. Principal Component Analysis) należy do grupy metod statystycznych i służy do badania związków zachodzących pomiędzy dwoma wielowymiarowymi zestawami zmiennych. Może być traktowana jako liniowa transformacja punktów z przestrzeni D-wymiarowej (przykładowo — D może być równe ilości pikseli wzorca obrazu) do nowej przestrzeni o mniejszej liczbie wymiarów. Nowa baza przestrzeni definiuje kierunki, które opisują największą zmienność z oryginalnego zestawu danych.

Kolejne etapy analizy PCA są następujące. Każdy D-wymiarowy wektor cech jest normalizowany poprzez odjęcie średniej ze wszystkich wektorów (A.12).

$$y_i = x_i - \bar{x} \tag{A.12}$$

gdzie:

$y_i$ – D-wymiarowy wektor cech po normalizacji
$x_i$ – D-wymiarowy wektor cech
$\bar{x}$ – średni wektor cech

Ze znormalizowanych wektorów cech tworzona jest macierz, w której każda kolumna odpowiada jednemu wektorowi (A.13).

$$\Phi = \begin{bmatrix} \varphi_1^1 & \cdots & \varphi_{N_T}^1 \\ \vdots & & \vdots \\ \varphi_1^D & \cdots & \varphi_{N_T}^D \end{bmatrix} \tag{A.13}$$

gdzie:

$\Phi$ – macierz złożona z wektorów cech
$N_T$ – ilość wektorów cech
$D$ – wymiar wektora cech

Funkcje bazowe nowej przestrzeni są obliczane poprzez wyznaczenie wektorów własnych następującej macierzy kowariancji (A.14), (A.15). W przypadku gdy wektory cech reprezentują kształty lub wzorce twarzy, efektywną metodą obliczania jest algorytm dekompozycji macierzy na wartości osobliwe (SVD ang. singular value decomposition). Wynika to z faktu, iż macierz kowariancji jest osobliwa (ilość wektorów cech jest mniejsza niż wymiar przestrzeni). Funkcje bazowe są określane również jako składowe główne (ang. principal components) lub „eigen-obiekty" (ang. eigenobjects).

$$\begin{aligned} C_\Phi &= cov(\Phi) \\ [U, S, V] &= svd\{C_\Phi\} \end{aligned} \tag{A.14}$$

gdzie:

$C_\Phi$ – macierz kowariancji dla $\Phi$
$U$ – macierz zawierająca w każdej kolumnie składowe główne
$S$ – diagonalna macierz zawierająca wartości własne odpowiadające składowym głównym
$V$ – macierz ortonormalna do $U$

$$S = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_D \end{bmatrix} \tag{A.15}$$

Z uszeregowanych w kolejności od największej do najmniejszej wartości własnych oraz odpowiadających im składowych głównych wybierane jest $K$ pierwszych elementów. Wartość $K$ obliczana jest zazwyczaj wg kryterium (A.16).

$$\sum_{i=1}^{K} \lambda_i \geqslant \frac{p}{100} \cdot \sum_{i=1}^{D} \lambda_i \tag{A.16}$$

gdzie:

$i$ – indeks kolejnego wektora cech

$\lambda_i$ – i-ta wartość własna macierzy kowariancji

$D$ – wymiar wektora cech

$p$ – zadana, procentowa wartość określająca proporcję wariancji danych

Składowe główne stanowią nową bazę przestrzeni, do której rzutowane są wektory cech. W procesie rzutowania otrzymuje się wektor wag $\Omega$ reprezentujący dany wektor cech (A.17).

$$\begin{aligned} &\Omega_i = [w_1^i \cdots w_K^i]^T \\ &w_j^i = \left(U^j\right)^T \cdot x_i \\ &i = 1 \cdots N_T,\ j = 1 \cdots K \end{aligned} \tag{A.17}$$

gdzie:

$\Omega_i$ – wektor wag dla i-go wektora cech

$U^j$ – j-ty element macierzy składowych głównych

$x_i$ – i-ty wektor cech

$N_T$ – ilość wektorów cech

$K$ – ilość uwzględnionych wektorów głównych

Każdy wektor cech możne zostać odtworzony poprzez kombinację liniową składowych głównych (A.18) oraz dodanie średniego wektora cech.

$$\tilde{y}_i = \bar{x} + \sum_{i=1}^{K} w_j^i \cdot U^j \tag{A.18}$$

gdzie:

$\bar{x}$ – średni wektor cech
$U^j$ – j-ty element macierzy składowych głównych
$x_i$ – i-ty wektor cech
$K$ – ilość uwzględnionych wektorów głównych

## A.5 Metoda przestrzeni skali (ang. scale-space)

Reprezentacja skali (ang. scale-space) jest specjalnym typem reprezentacji wieloskalowej zaproponowanym przez Tonny'ego Lindeberga [54]. Niech $f : \Re^N \to \Re$ reprezentuje dowolny ciągły sygnał. Reprezentacja skali $L : \Re^N \times \Re_+ \to \Re$ jest zdefiniowana przez (A.19):

$$\begin{aligned} L(\cdot;0) &= f \\ L(\cdot;\sigma) &= g(\cdot;\sigma) * f \end{aligned} \tag{A.19}$$

gdzie:

$\sigma \in \Re_+$ – parametr skali
$g : \Re^N \times \Re_+ \backslash \{0\} \to \Re$ – N-wymiarowy rozkład Gaussa
$f$ – dowolny ciągły sygnał

W przypadku obrazu rozkład Gaussa definiowany jest następująco (A.20):

$$g(x, y, \sigma) = \frac{1}{2 \cdot \pi \cdot \sigma^2} \cdot e^{\left(-\frac{x^2+y^2}{2 \cdot \sigma^2}\right)} \tag{A.20}$$

gdzie:

$x, y$ – współrzędne obrazu

Reprezentacja ta może zostać wykorzystana do detekcji cech w różnych skalach (w zależności od parametru $\sigma$) — rys. A.3. Detekcja cech w pojedynczej skali, odbywa się poprzez wyszukiwanie lokalnych przestrzennych maksimów lub miejsc zerowych różniczkowego deskryptora cech. Mogą to być plamy (ang. blobs), grzbiety (ang. ridges), krawędzie (ang. edges), itp. W pracy wykorzystano detekcję plam. Plama to obszar ciemny otoczony elementami jasnymi, lub odwrotnie. Matematycznie plama to ekstremum lokalne laplasjanu(A.21):

$$\begin{aligned} \nabla^2 L &= L_{xx} + L_{yy} \\ \nabla(\nabla^2 L) &= 0 \end{aligned} \tag{A.21}$$


Rysunek A.3: Reprezentacja skali obrazu: (a) obraz, (b) reprezentacja dla $\sigma = 2$, (c) reprezentacja dla $\sigma = 8$

Znak laplasjanu określa czy jest to plama jasna ($\nabla(\nabla^2 L) < 0$) czy ciemna ( $\nabla(\nabla^2 L) > 0$). W wyniku zastosowania powyższego operatora, otrzymywany jest obraz z wyraźnymi maksimami w miejscach występowania plam o rozmiarze odpowiadającym wybranej skali $\sigma$. Obraz ten podlega normalizacji względem skali i może zostać wykorzystany do detekcji cech. W przypadku uwzględnienia wszystkich skal, detekcja odbywa się w przestrzeni trójwymiarowej (tzn. nie tylko w płaszczyźnie obrazu ale również na osi skali).

## A.6 Filtry Gabora

Filtr Gabora jest funkcją Gaussa modulowana przez falę sinusoidalną oraz kosinusoidalną. Dla sygnału 2D definiowany jest następująco (A.22):

$$\begin{aligned}
&\Psi_G\left(x, y; f_{0G}, \theta_G\right) = \frac{f_{0G}^2}{\pi \cdot \gamma_G \cdot \eta_G} \cdot e^A \cdot e^B \\
&A = -\left(\frac{f_{0G}^2}{\gamma_G^2} \cdot x'^2 + \frac{f_{0G}^2}{\gamma_G^2} \cdot y'^2\right) \\
&B = \left(j2\pi \cdot f_{0G} \cdot x'\right) \\
&x' = x \cdot \cos\theta_G + y \cdot \sin\theta_G \\
&y' = -x \cdot \sin\theta_G + y \cdot \cos\theta_G
\end{aligned} \tag{A.22}$$

gdzie:

$\gamma_G$ – ostrość filtru wzdłuż dłuższej osi
$\eta_G$ – ostrość filtru wzdłuż krótszej osi
$f_{0G}$ – częstotliwość centralna filtru
$\theta_G$ – kąt obrotu głównej osi filtru (orientacja)

Wyszukiwane obiektów na obrazach odbywa się za pomocą bibliotek filtrów Gabora, uwzględniających różne częstotliwości, orientacje i szerokości.

## A.7 Wyznaczanie mapy prawdopodobieństwa oczu oraz ust

Obliczanie mapy prawdopodobieństwa oczu składa się z dwóch etapów. W pierwszym wykorzystywana jest informacja o kolorze (A.23), w wyniku czego otrzymywana jest mapa nr 1 (rys. A.4a):

$$EyeMap1 = \frac{1}{3} \cdot \left( I_{Cb}^2 + \tilde{I}_{Cr}^2 + I_{Cb}/I_{Cr} \right) \tag{A.23}$$

gdzie:

$\tilde{I}_{Cr}$ – Negacja składowej Cr
$I_{Cr}, I_{Cb}$ – Składowe obrazu YCbCr

Obszar oczu zazwyczaj odróżnia się wyraźnie od reszty twarzy tworząc ciemniejsze plamy. Dlatego tworzona jest druga mapa prawdopodobieństwa oczu (A.24) — rys. A.4b. Rejony oczu wzmacniane są w niej poprzez zastosowanie algorytmu „detekcji plam" w reprezentacji skali (por. A.5).

$$EyeMap2 = \sigma \cdot (L_{xx} + L_{yy}) \tag{A.24}$$

gdzie:

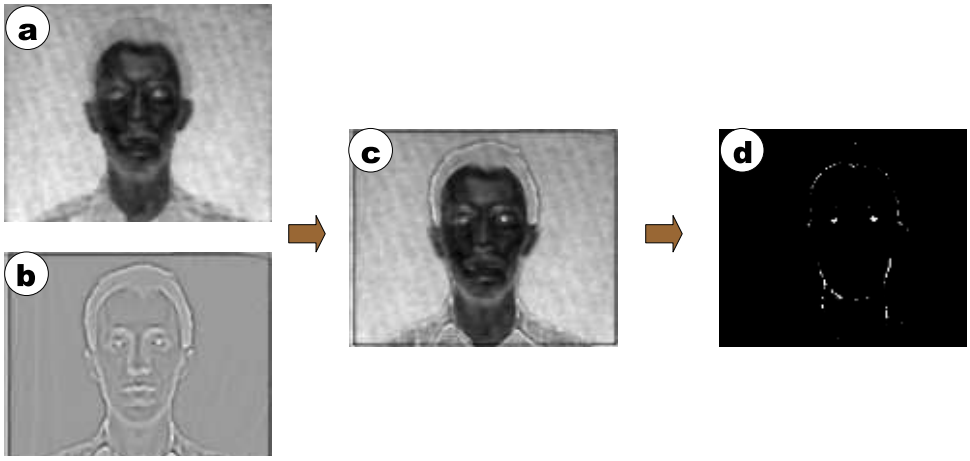
$L_{xx}, L_{yy}$ – drugie pochodne cząstkowe obrazu
$\sigma$ – parametr skali

Obydwie mapy podlegają normalizacji do zakresu [0-1]. Na podstawie złożenia obu map (rys. A.4c) oraz otrzymanej w poprzednich etapach maski twarzy (por. 5.3.1), tworzona jest binarna maska (rys. A.4d) zawierająca obiekty będące potencjalnymi oczami (A.25).

$$BW_{eye} = \begin{cases} 1; & [(EyeMap1 + EyeMap2) > eye_{thres}] \cap BW_{face} \\ 0; & w\, przeciwnym\, przypadku \end{cases} \tag{A.25}$$

gdzie:

$EyeMap1, EyeMap1$ – mapy prawdopodobieństwa oczu
$BW_{face}$ – binarna maska twarzy
$eye_{thres}$ – próg binaryzacji mapy prawdopodobieństwa oczu




Rysunek A.4: (a) mapa prawdopodobieństwa oczu nr 1, (b) mapa prawdopodobieństwa oczu nr 2, (c) połączenie obydwu map, (d) binarna maska oczu

Próg binaryzacji $eye_{thres}$ dobierany jest automatycznie według następującego kryterium (A.26):

$$eye_{thres} = \mu_{eye} + const \cdot \sigma_{eye} \tag{A.26}$$

gdzie:

$\mu_{eye}, \sigma_{eye}$ – odchylenie standardowe oraz wartość oczekiwana intensywności pikseli mapy oczu w obszarze należącym do twarzy

Mapa prawdopodobieństwa ust (rys. A.5a) wyznaczana jest następująco (A.27):

$$MouthMap = I_{Cb} \cdot \left( I_{Cr}^2 - \eta_m \cdot \frac{I_{Cr}}{I_{Cb}} \right)^2 \tag{A.27}$$

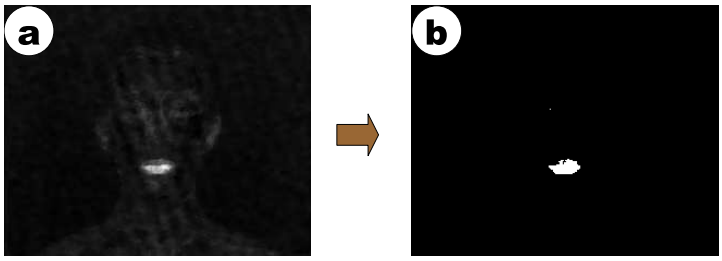
$$\eta_m = 0.95 \cdot \frac{\frac{1}{n} \cdot \sum\limits_{(x,y) \in BW_{face}} I_{Cr}^2}{\frac{1}{n} \cdot \sum\limits_{(x,y) \in BW_{face}} \frac{I_{Cr}}{I_{Cb}}}$$

gdzie:

$I_{Cr}, I_{Cb}$ – Składowe obrazu YCbCr

$BW_{face}$ – binarna maska twarzy

$n$ – ilość pikseli należących do obszaru twarzy




Rysunek A.5: (a) mapa prawdopodobieństwa ust, (b) połączenie obydwu map, (d) binarna maska ust

Parametr $\eta_m$ obliczany jest na podstawie zliczania ilości pikseli należących do obszaru twarzy (maska) dla odpowiednich składowych przestrzeni kolorów chrominancji-luminancji. W wyniku otrzymywany jest obraz na którym widoczne są jaśniejsze plamy w miejscu występowania ust. Podobnie jak w przypadku oczu, z mapy ust tworzona jest następnie binarna maska (rys. A.5b), zawierająca obiekty będące potencjalnymi ustami (A.28)(A.29).

$$BW_{mouth} = \begin{cases} 1; & [MouthMap > mouth_{thres}] \cap BW_{face} \\ 0; & w\,przeciwnym\,przypadku \end{cases} \tag{A.28}$$

gdzie:

$MouthMap$ – mapa prawdopodobieństwa ust
$BW_{face}$ – binarna maska twarzy
$mouth_{thres}$ – próg binaryzacji mapy prawdopodobieństwa ust

$$mouth_{thres} = \mu_{mouth} + const \cdot \sigma_{mouth} \tag{A.29}$$

gdzie:

$\mu_{mouth}, \sigma_{mouth}$ – odchylenie standardowe oraz wartość oczekiwana intensywności pikseli mapy ust w obszarze należącym do twarzy

## Dodatek B

# Opis jednostek czynnościowych mimiki

Tabela B.1: Wybrane jednostki czynnościowe dolnej części twarzy — cz.1

| **Jednostka czynnościowa** | **Opis** | **Uwagi** |
|---|---|---|
| AU12 | Unoszenie kącików ust (ang. lip corner puller)<br>Grupa mięśni: *zygomaticus major* | Ruch ten jest podstawą gestu uśmiechu i może być wykonywany niesymetrycznie. Często występuje równocześnie z AU6 (unoszenie policzka) oraz AU25 (otwarcie ust). Może być mylony z AU14 – w zależności od osobniczych cech człowieka. |
| AU14 | „Przyciąganie ust” (ang. dimpler)<br>Grupa mięśni: *buccinator* | Ruch ten powoduje powstawanie dołeczków w kącikach ust i może być wykonywany niesymetrycznie. |
| AU15 | Opuszczanie kącików ust (ang. lip corner depressor)<br>Grupa mięśni: *depressor anguli oris (a.k.a. triangularis)* | |
| AU17 | Unoszenie podbródka (ang. chin raiser)<br>Grupa mięśni: *mentalis* | |

Tabela B.2: Wybrane jednostki czynnościowe dolnej części twarzy — cz.2

| | | |
|---|---|---|
| AU19 | Wysunięcie języka (ang. tongue out) | Istnieje wiele możliwości wykonania tego gestu. |
| AU25, AU26, AU27 | AU25 – Otwarcie ust (ang. lip part)<br>Grupa mięśni: *depressor labii inferioris or relaxation of mentalis, or Oorbicularis oris*<br><br>AU26 – opuszczenie szczęki (ang. jaw drop)<br>Grupa mięśni: *masseter, relaxed temporalis and internal pterygoid*<br><br>AU27 – szerokie otwarcie ust (ang. mouth strech)<br>Grupa mięśni: *pterygoids, digastric* | Zazwyczaj stosuje się tylko jednostkę AU25 przy czym określa się intensywność gestu (stopień otwarcia ust). |
| AU51, AU52 | Obrót głowy lewo/prawo (ang. head turn left/right) | Często niemożliwy do wykonania dla osób z dużym upośledzeniem możliwości ruchu. |
| AU53, AU54 | Przechylenie głowy góra/dół (ang. head up/down) | j.w. |
| AU55, AU56 | Przechylenie głowy lewo/prawo (ang. head titl left/right) | j.w. |

Tabela B.3: Wybrane jednostki czynnościowe górnej części twarzy.

| **Jednostka czynnościowa** | **Opis** | **Uwagi** |
|---|---|---|
| AU1+2 | Ruch unoszący brwi (ang. inner and outer brow raiser):<br>AU1 – część przyśrodkowa,<br>AU2 – część boczna.<br>Grupa mięśni: *m. occipitofrontalis, venter frontalis* | U większości osób jednak niezależne wywołanie ruchu AU1 lub AU2 jest niemożliwe bez długotrwałego treningu, a unoszenie brwi odpowiada obu jednostkom czynnościowym działającym równocześnie, co oznaczane jest jako AU1+2. |
| AU4 | Ruch marszczenia brwi (ang. brow lowerer)<br>Grupa mięśni: *corrugator supercilii, depressor supercilii* | |
| AU6 | Unoszenie policzka (ang. cheek raiser)<br>Grupa mięśni: *orbicularis oculi, pars orbitalis* | Gest trudny do wykonania samodzielnie. Zazwyczaj występuje w połączeniu z AU7 (mrużenie oczu) lub AU12 (unoszenie kącików ust) |
| AU45, AU46 | AU45 – mruganie (ang. blink)<br>AU46 – zmrużenie oka (ang. wink)<br>Grupa mięśni: *relaxation of levator palpebrae superioris; orbicularis oculi, pars palpebralis* | Gesty podobne ale zmrużenie oka jest zazwyczaj wolniejsze od mrugania oraz wykonywane jednym okiem. |

Tabela B.4: Opis wybranych elementów mimiki — AU1, AU2, AU1+2

| **Element mimiki** | **Opis** |
|---|---|
| AU1,AU2, AU1+2 | AU1 – Unoszenie wewnętrznej części brwi:<br>• Wewnętrzna część brwi podciągana jest w górę.<br>• U wielu ludzi powoduje powstanie ukośnego kształtu brwi.<br>• Powoduje powstawanie poziomych zmarszczek skóry na czole, rozciągających się raczej w środkowej części czoła. Mogą one być zakrzywione, wyższe w środkowej części. U dzieci zmarszczki mogą nie występować. W przypadku osób z permanentnymi zmarszczkami pogłębiają się one.<br>• Może powodować lekki ruch zewnętrznej części brwi do środka<br>AU2 – Unoszenie zewnętrznej części brwi:<br>• Zewnętrzna część brwi podciągana jest w górę.<br>• Powoduje powstawanie łukowatego kształtu brwi.<br>• Powoduje rozciąganie w górę bocznej części fałdu powieki górnej.<br>• U niektórych ludzi powoduje powstanie krótkich poziomych zmarszczek powyżej bocznej części brwi. Mogą również pojawić się zmarszczki w środkowej części czoła ale dużo słabsze od bocznych.<br>• Może powodować lekki ruch wewnętrznej części brwi.<br>AU1+2 – Unoszenie brwi:<br>• Całość brwi podciągana jest w górę.<br>• Powoduje powstawanie łukowatego i zakrzywionego kształtu brwi.<br>• Marszczy skórę czoła tak że powstają poziome zmarszczki. U dzieci zmarszczki mogą nie występować.<br>• Rozciąga fałd powieki górnej tak że jest bardziej widoczne.<br>• U niektórych ludzi (z głęboko osadzonymi oczyma) rozciąganie fałdu powieki górnej odkrywa górną powiekę która zazwyczaj jest zakryta przez fałd. |

Tabela B.5: Opis wybranych elementów mimiki — AU4

| **Element mimiki** | **Opis** |
|---|---|
| AU4 | Marszczenie brwi:<br>• Opuszczenie brwi — różne przypadki (całość brwi, tylko zewnętrzna część lub tylko wewnętrzna część).<br>• Popycha fałd powieki górnej w dół oraz może zwęzić aperturę oka.<br>• Przyciąga brwi ku sobie.<br>• Wywołuje pionowe zmarszczki pomiędzy brwiami (mogą one być głębokie). U niektórych ludzi zmarszczki mogą być pod kątem 45° lub zarówno pionowe jaki i pod kątem. Mogą również powstawać poziome zmarszczki u nasady nosa. W przypadku permanentnych zmarszczek pogłębiają się one.<br>• Może powodować powstawanie ukośnych zmarszczek lub wypukłości mięśni rozciągających się od środka czoła ponad środkową częścią brwi, aż do wewnętrznego kącika brwi.<br>• Mogą również wystąpić inne rodzaje zmarszczek. |

Tabela B.6: Opis wybranych elementów mimiki — AU6

| **Element mimiki** | **Opis** |
|---|---|
| AU6 | Unoszenie policzka:<br>• Skurcz mięśni wokół oka podciąga skórę od skroni i policzka w kierunku oka.<br>• Podniesienie trójkąta podoczodołowego, uniesienie policzka w górę.<br>• Wypycha skórę otaczającą oko w kierunku oczodołu zwężając aperturę oczu, powodując marszczenie skóry pod okiem oraz wypychanie fałdu powieki górnej i zmianę jego kształtu.<br>• Może powodować powstawanie „kurzych łapek" lub zmarszczek, które rozszerzają się radialnie od zewnętrznych kącików oka.<br>• Pogłębia fałd powieki dolnej.<br>• Może obniżyć w niewielkim stopniu boczne części brwi.<br>• Bardziej intensywny gest:<br>– uwidacznia lub pogłębia bruzdę nosowo-wargową,<br>– unosi w niewielkim stopniu zewnętrzną część górnej wargi,<br>– uwidacznia lub pogłębia bruzdę podoczołową, tak że od góry trójkąta podoczodołowy pojawiają się zmarszczki (proste lub półksiężycowate). |

Tabela B.7: Opis wybranych elementów mimiki — AU12

| **Element mimiki** | **Opis** |
|---|---|
| AU12 | Unoszenie kącików ust:<br>• Pociągnięcie kącików ust w tył oraz ukośnie do góry, tak że kształt ust staje się zakrzywiony.<br>• Pogłębienie bruzda nosowo-wargowej wraz z pociągnięciem jej w bok i do góry, przylegająca część skóry również przesuwa się do góry oraz jest unoszona na bok.<br>• Słabsza lub umiarkowana jednostka AU12 — podniesienie trójkąta podoczodołowego oraz prawdopodobne pogłębienie bruzdy podoczodołowej<br>• Bardziej zdecydowanie wykonanie akcji powoduje ujawnienie następujących cech:<br>– większy skurcz mięśnia uwydatnia bardziej trójkąt podoczodołowy,<br>– pogłębienie bruzdy podoczodołowej jest bardziej widoczne,<br>– „worki" pod dolnymi powiekami,<br>– skóra policzka naciskając na dolną powiekę powoduje zmniejszenie otworu oczu,<br>– czasami połączone z AU6. |

Tabela B.8: Opis wybranych elementów mimiki — AU14, AU15

| **Element mimiki** | **Opis** |
|---|---|
| AU14, AU15 | AU14 – „Przyciąganie ust”:<br>• Zmarszczki w kącikach ust.<br>• Powstawanie „dołeczków” w policzku.<br>• Ruch kącików ust w górę, możliwy również ruch okrężny.<br>• Wargi rozciągnięte na boki, spłaszczone i naprężone.<br>• W niektórych przypadkach pogłębienie bruzdy nosowo-wargowej.<br>• Skóra brody podciągana jest w górę oraz wygładza się i rozciąga.<br>• Może występować symetrycznie lub niesymetrycznie.<br>AU15 – Opuszczanie kącików ust:<br>• Opuszczenie kącików ust w dół.<br>• Zmiana kształtu warg (jak na zdjęciu) z lekkim poziomym rozciągnięciem.<br>• Dla intensywnego gestu — tworzenie się zmarszczek i innych zmian wyglądu skóry (wypukłości, worki, kieszonki) w okolicy kącików ust.<br>• Może powodować pojawianie się lub wygładzanie wybrzuszeń na bródce oraz wklęsłości pod dolną wargą.<br>• Jeśli występują stałe dla danej osoby bruzdy nosowo-wargowe mogą się one pogłębić i wydłużyć. |

Tabela B.9: Opis wybranych elementów mimiki — AU17, AU19

| **Element mimiki** | **Opis** |
|---|---|
| AU17, AU19 | AU17 – Unoszenie podbródka:<br>• Koniec podbródka unoszony do góry.<br>• Dolna warga unoszona do góry.<br>• Możliwe powstawanie zmarszczek na podbródku oraz zagłębienia poniżej środkowej części dolnej wargi.<br>• Część ust formuje się w kształt odwróconego U.<br>• Lekkie wysunięcie dolnej wargi.<br>AU19 – Wysunięcie języka:<br>• Przynajmniej koniuszek języka widoczny.<br>• Może wystąpić przysłonięcie jednej z warg lub kącika ust.<br>• Usta mogą być otwarte lub mogą wystąpić zmiany ich kształtu. |

Tabela B.10: Opis wybranych elementów mimiki — AU25, AU26, AU27

| **Element mimiki** | **Opis** |
|---|---|
| AU25, AU26, AU27 | AU25 – otwarcie ust:<br>• Wargi oddzielone od siebie, wewnętrzna część warg bardziej widoczna.<br>• W niektórych przypadkach widoczne zęby i dziąsła.<br>• Widoczna jama ustna (w zależności od jednostki AU26/27).<br>AU26 – opuszczenie szczęki:<br>• Rozwarcie żuchwy/szczęki dolnej.<br>• Przy otwarciu ust (AU25+AU26) widoczna przestrzeń między górnymi i dolnymi zębami.<br>• Dolna szczęka rozluźniona.<br>• Może wystąpić przypadkowe otwarcie ust.<br>• Stopniowe napinanie mięśni dolnej szczęki.<br>• Możliwe rozwarcie żuchwy/szczęki dolnej z jednoczesnym zaciśnięciem warg (AU17/24).<br>AU27 – szerokie otwarcie ust:<br>• Owalny kształt ust.<br>• Wargi z dala od siebie.<br>• Poprzez rozciąganie policzki wydłużają się i spłaszczają.<br>• Zmienia się wygląd podbródka i strefy poniżej.<br>• Jednostka połączona z AU25. |

Tabela B.11: Opis wybranych elementów mimiki — AU51-56

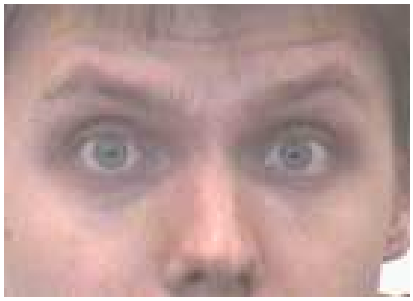
| **Element mimiki** | **Opis** |
|---|---|
| AU51-56 | Ruchy głowy:<br>• AU51,AU52 – obrót głowy lewo/prawo.<br>• AU53,AU54 – przechylenie głowy góra/dół.<br>• AU55,AU56 – przechylenie głowy lewo/prawo.<br>Uwagi:<br>• Opis w metodyce FACS mniej dokładny.<br>• Zmiany wyglądu twarzy wywołane zniekształceniami perspektywicznymi.<br>• Duże ruchy mogą powodować przysłonięcie niektórych elementów morfologicznych twarzy (np. oko, ucho) lub pojawienie się dodatkowych elementów (np. szyja). |

Tabela B.12: Opis wybranych elementów mimiki — AU43, AU45, AU46

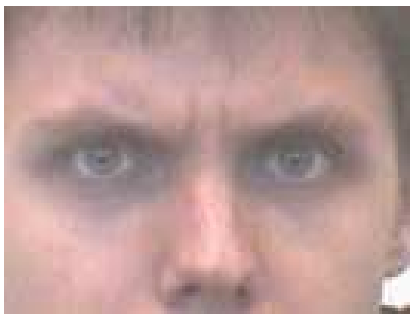
| **Element mimiki** | **Opis** |
|---|---|
| AU43, AU45, AU46 | AU43 – zamknięcie oczu:<br>• Powieki opuszczane są w dół zmniejszając aperturę oka.<br>• Widoczna jest większa część powiek górnych.<br>• Możliwe różne stopnie intensywności gestu – np. przymknięcie oczu lub zaciśnięcie oczu.<br>AU45 – Mruganie:<br>• Bardzo szybkie zamknięcie i otwarcie oczu (bez pauzy w momencie gdy oczy są zamknięte).<br>• Możliwy gest jednostronny lub obustronny.<br>AU46 – Przymrużenie oka:<br>• Intencjonalne zmrużenie jednego z oczu.<br>• Oko zostaje zamknięte na krótką chwilę (z pauzą w momencie gdy jest w pozycji zamkniętej)<br>• Maksymalny czas zmrużenia oka – 2 sec. |

Tabela B.13: Opis wybranych elementów mimiki — jednostki MPEG4-FAP

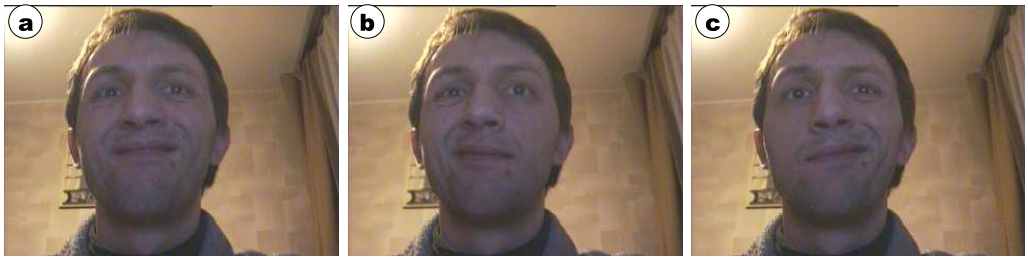
| **Element mimiki** | **Opis** |
|---|---|
| FAP 31-36 | Wertykalne przemieszczenie punktów brwi:<br>• FAP31 – przemieszczenie wewnętrznej części lewej brwi (punkt 4.1)<br>• FAP32 – przemieszczenie wewnętrznej części prawej brwi (punkt 4.2)<br>• FAP33 – przemieszczenie środkowej części lewej brwi (punkt 4.3)<br>• FAP34 – przemieszczenie środkowej części prawej brwi (punkt 4.4)<br>• FAP35 – przemieszczenie zewnętrznej części lewej brwi (punkt 4.5)<br>• FAP36 – przemieszczenie zewnętrznej części prawej brwi (punkt 4.6) |



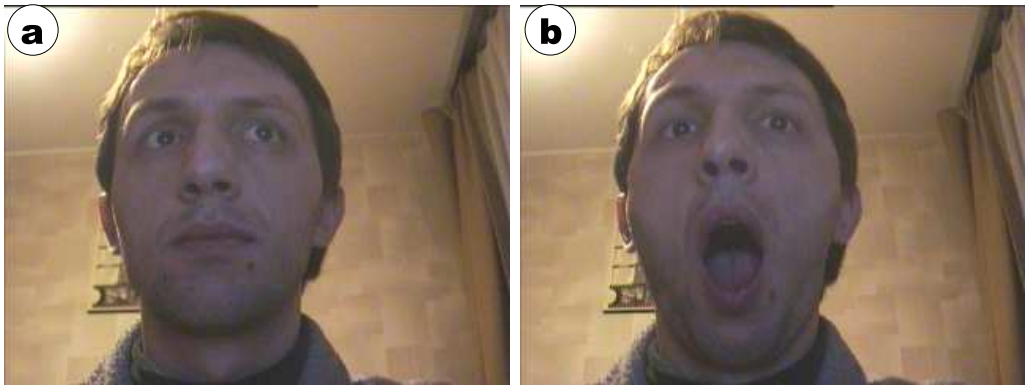
Rysunek B.1: Element mimiki AU1+2 — unoszenie brwi



Rysunek B.2: Element mimiki AU4 — marszczenie brwi




Rysunek B.3: Element mimiki AU12 — unoszenie kącików ust: (a) obustronne, (b)(c) jednostronne




Rysunek B.4: Elementy mimiki: (a) AU0 — położenie frontalne, (b) AU25+26 — otwarcie ust

# Dodatek C

# Rezultaty rozpoznawania elementów mimiki — wykresy i tabele

## C.1 Informacje ogólne

Liczebność elementów w zbiorach danych dla poszczególnych metod jest następująca:

Tabela C.1: Liczebność dla histogramów orientacji

| | **całość** | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU 25+26** |
|---|---|---|---|---|---|---|
| **zbiór uczący** | 280 | 141 | 32 | 33 | 35 | 39 |
| **zbiór testowy 1** | 514 | 200 | 63 | 81 | 83 | 87 |
| **zbiór testowy 2** | 339 | 131 | 47 | 52 | 54 | 55 |
| **zbiór testowy 3** | 319 | 116 | 54 | 47 | 52 | 51 |
| **zbiór testowy 4** | 322 | 121 | 58 | 49 | 48 | 47 |

Ze względu na małą ilość danych (dla rozpoznawania kształtów), cały zbiór danych podzielono losowo na zbiór uczący i testowy w stosunku 70% do 30% (tzw. stratified sampling).

Tabela C.2: Liczebność dla modeli kształtów — zbiór uczący

| | **całość** | **AU0** | **AU1+2** | **AU4** |
|---|---|---|---|---|
| **test 1** | 75 | 48 | 27 | - |
| **test 2** | 114 | 73 | 41 | - |
| **test 3** | 86 | 48 | 27 | 11 |
| **test 4** | 131 | 73 | 41 | 17 |

Tabela C.3: Liczebność dla modeli kształtów — zbiór testowy

| | **całość** | **AU0** | **AU1+2** | **AU4** |
|---|---|---|---|---|
| **test 1** | 32 | 21 | 11 | - |
| **test 2** | 49 | 31 | 18 | - |
| **test 3** | 36 | 21 | 11 | 4 |
| **test 4** | 56 | 31 | 18 | 7 |

Parametry jakości klasyfikacji (C.1)—(C.4).

$$PPV = \frac{TP}{TP + FP} \cdot 100\% \tag{C.1}$$

$$NPV = \frac{TN}{FN + TN} \cdot 100\% \tag{C.2}$$

$$Sens = \frac{TP}{TP + FN} \cdot 100\% \tag{C.3}$$

$$Spec = \frac{TN}{FP + TN} \cdot 100\% \tag{C.4}$$

gdzie:

$PPV$ – wartość predykcyjna dodatnia klasyfikacji (ang. positive predictive value)

$NPV$ – wartość predykcyjna ujemna klasyfikacji (ang. negative predictive value)

$Sens$ – czułość (ang. sensitivity)

$Spec$ – swoistość (ang. specificity)

$TP$ – wyniki prawdziwie dodatnie (ang. true positives)

$FN$ – wyniki fałszywie ujemne (ang. false negatives)

$FP$ – wyniki fałszywie dodatnie (ang. false positives)

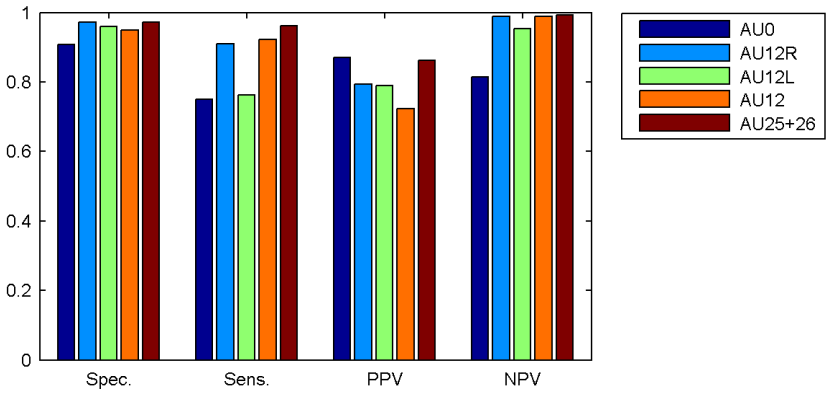
$TN$ – wyniki prawdziwie ujemne (ang. true negatives)

## C.2 Histogramy orientacji — konfiguracja nr 1

Obszar zainteresowań — wybrane ROI. Histogramy orientacji — gradienty w przestrzeni skali.

Tabela C.4: Konfiguracja 1, klasyfikator kNN

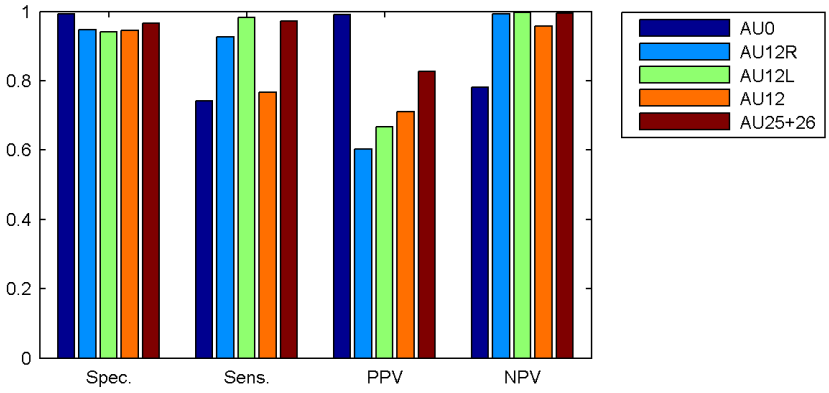
| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 90.78% | 97.17% | 96.05% | 94.88% | 97.25% |
| **Sens.** | 75.00% | 90.91% | 76.19% | 92.31% | 96.15% |
| **PPV** | 87.00% | 79.37% | 79.01% | 72.29% | 86.21% |
| **NPV** | 81.53% | 98.89% | 95.38% | 98.84% | 99.30% |




Rysunek C.1: Konfiguracja 1, klasyfikator kNN

Tabela C.5: Konfiguracja 1, klasyfikator QDA

| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 99.19% | 94.71% | 94.12% | 94.51% | 96.59% |
| **Sens.** | 74.16% | 92.68% | 98.18% | 76.62% | 97.30% |
| **PPV** | 99.00% | 60.32% | 66.67% | 71.08% | 82.76% |
| **NPV** | 78.03% | 99.33% | 99.77% | 95.82% | 99.53% |



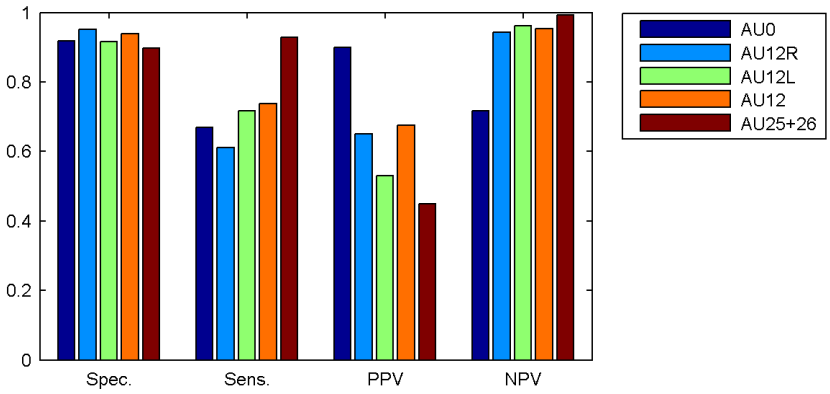

Rysunek C.2: Konfiguracja 1, klasyfikator QDA

## C.3 Histogramy orientacji — konfiguracja nr 2

Obszar zainteresowań — wybrane ROI. Histogramy orientacji — filtry Gabora.

Tabela C.6: Konfiguracja 2, klasyfikator kNN

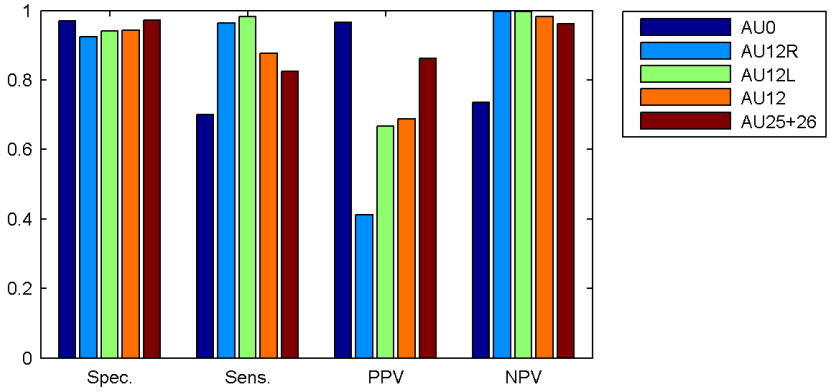
| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 91.84% | 95.08% | 91.63% | 93.84% | 89.83% |
| **Sens.** | 66.91% | 61.19% | 71.67% | 73.68% | 92.86% |
| **PPV** | 90.00% | 65.08% | 53.09% | 67.47% | 44.83% |
| **NPV** | 71.66% | 94.24% | 96.07% | 95.36% | 99.30% |




Rysunek C.3: Konfiguracja 2, klasyfikator kNN

Tabela C.7: Konfiguracja 2, klasyfikator QDA

| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 97.06% | 92.40% | 94.12% | 94.21% | 97.16% |
| **Sens.** | 69.93% | 96.30% | 98.18% | 87.69% | 82.42% |
| **PPV** | 96.50% | 41.27% | 66.67% | 68.67% | 86.21% |
| **NPV** | 73.57% | 99.78% | 99.77% | 98.14% | 96.25% |



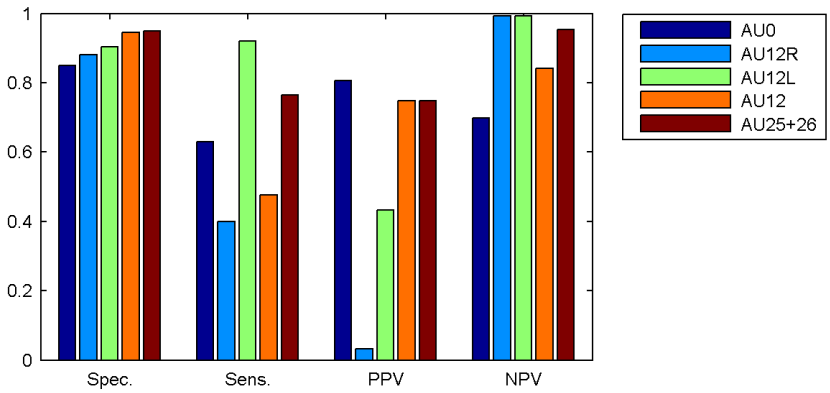

Rysunek C.4: Konfiguracja 2, klasyfikator QDA

## C.4 Histogramy orientacji — konfiguracja nr 3

Obszar zainteresowań — cała twarz. Histogramy orientacji — gradienty w przestrzeni skali.

Tabela C.8: Konfiguracja 3, klasyfikator kNN

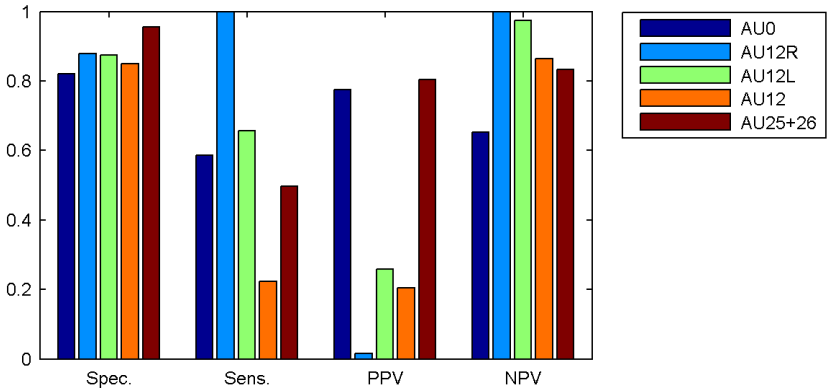
| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 84.88% | 88.02% | 90.34% | 94.53% | 94.87% |
| **Sens.** | 62.89% | 40.00% | 92.11% | 47.69% | 76.47% |
| **PPV** | 80.50% | 3.17% | 43.21% | 74.70% | 74.71% |
| **NPV** | 69.75% | 99.33% | 99.31% | 84.22% | 95.32% |




Rysunek C.5: Konfiguracja 3, klasyfikator kNN

Tabela C.9: Konfiguracja 3, klasyfikator QDA

| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 82.00% | 87.91% | 87.55% | 84.93% | 95.44% |
| **Sens.** | 58.71% | 100.00% | 65.63% | 22.37% | 49.65% |
| **PPV** | 77.50% | 1.59% | 25.93% | 20.48% | 80.46% |
| **NPV** | 65.29% | 100.00% | 97.46% | 86.31% | 83.37% |



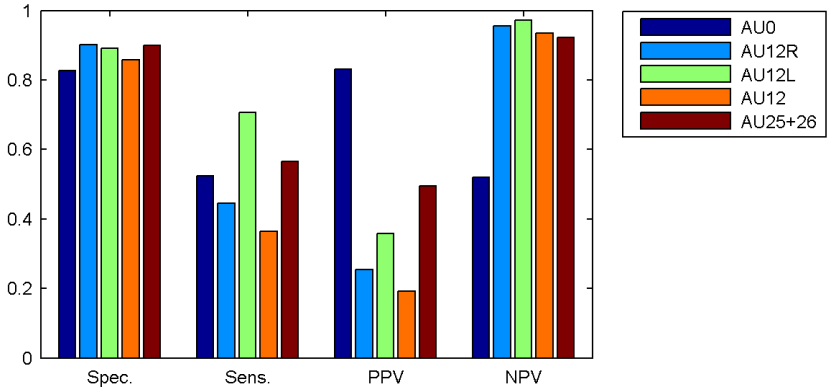

Rysunek C.6: Konfiguracja 3, klasyfikator QDA

## C.5 Histogramy orientacji — konfiguracja nr 4

Obszar zainteresowań — cała twarz. Histogramy orientacji — filtry Gabora.

Tabela C.10: Konfiguracja 4, klasyfikator kNN

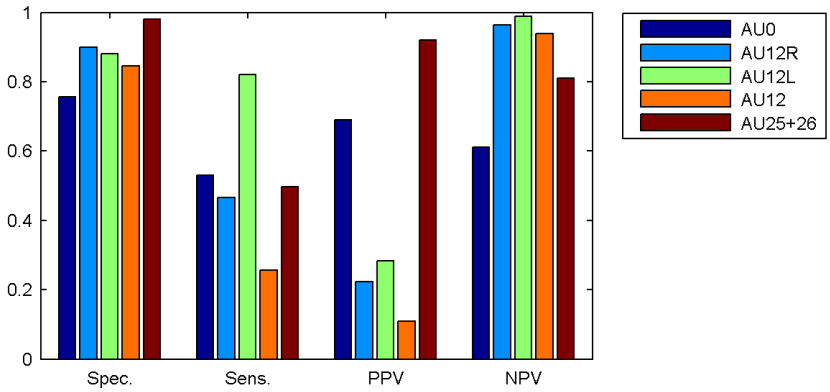
| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 82.74% | 90.17% | 89.01% | 85.74% | 89.95% |
| **Sens.** | 52.37% | 44.44% | 70.73% | 36.36% | 56.58% |
| **PPV** | 83.00% | 25.40% | 35.80% | 19.28% | 49.43% |
| **NPV** | 51.91% | 95.57% | 97.23% | 93.50% | 92.27% |




Rysunek C.7: Konfiguracja 4, klasyfikator kNN

Tabela C.11: Konfiguracja 4, klasyfikator QDA

| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 75.59% | 89.88% | 88.07% | 84.55% | 98.02% |
| **Sens.** | 53.08% | 46.67% | 82.14% | 25.71% | 49.69% |
| **PPV** | 69.00% | 22.22% | 28.40% | 10.84% | 91.95% |
| **NPV** | 61.15% | 96.45% | 98.85% | 93.97% | 81.03% |



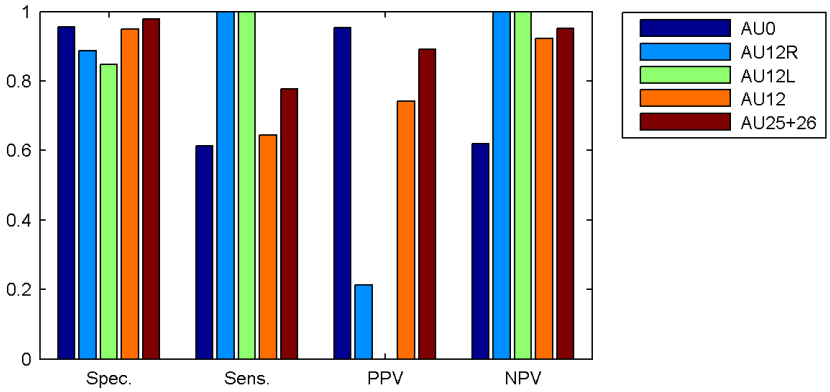

Rysunek C.8: Konfiguracja 4, klasyfikator QDA

## C.6 Histogramy orientacji — badanie wpływu oświetlenia

Histogramy orientacji — ocena wpływu bocznego oświetlenia sztucznego, sekwencja 2.

Tabela C.12: Wpływ oświetlenia, klasyfikator kNN

| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 95.56% | 88.75% | 84.66% | 94.95% | 97.83% |
| **Sens.** | 61.27% | 100.00% | 100.00% | 64.52% | 77.78% |
| **PPV** | 95.42% | 21.28% | 0.00% | 74.07% | 89.09% |
| **NPV** | 62.02% | 100.00% | 100.00% | 92.28% | 95.07% |



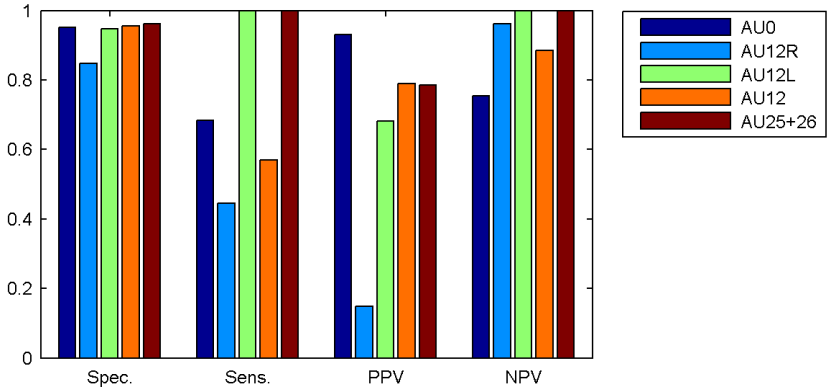

Rysunek C.9: Wpływ oświetlenia, klasyfikator kNN

## C.7 Histogramy orientacji — badanie wpływu skali

Histogramy orientacji — ocena wpływu zmian skali (odsunięcie głowy od kamery), sekwencja 3.

Tabela C.13: Wpływ skali, klasyfikator kNN

| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 95.06% | 84.77% | 94.79% | 95.56% | 96.07% |
| **Sens.** | 68.35% | 44.44% | 100.00% | 56.94% | 100.00% |
| **PPV** | 93.10% | 14.81% | 68.09% | 78.85% | 78.43% |
| **NPV** | 75.49% | 96.24% | 100.00% | 88.43% | 100.00% |



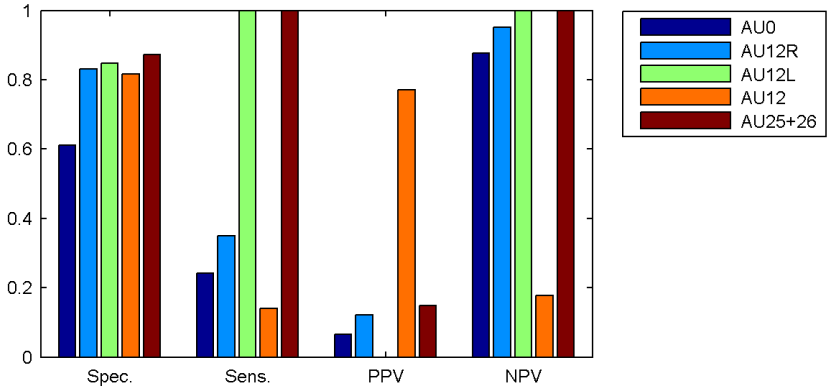

Rysunek C.10: Wpływ skali, klasyfikator kNN

## C.8 Histogramy orientacji — badanie wpływu skali i rotacji

Histogramy orientacji — ocena wpływu zmian skali i rotacji (odsunięcie głowy połączone z jej przechyleniem), sekwencja 4.

Tabela C.14: Wpływ skali i rotacji, klasyfikator kNN

| | **AU0** | **AU12R** | **AU12L** | **AU12** | **AU25+26** |
|---|---|---|---|---|---|
| **Spec.** | 61.03% | 83.17% | 84.83% | 81.67% | 87.34% |
| **Sens.** | 24.24% | 35.00% | 100.00% | 14.07% | 100.00% |
| **PPV** | 6.61% | 12.07% | 0.00% | 77.08% | 14.89% |
| **NPV** | 87.62% | 95.09% | 100.00% | 17.82% | 100.00% |



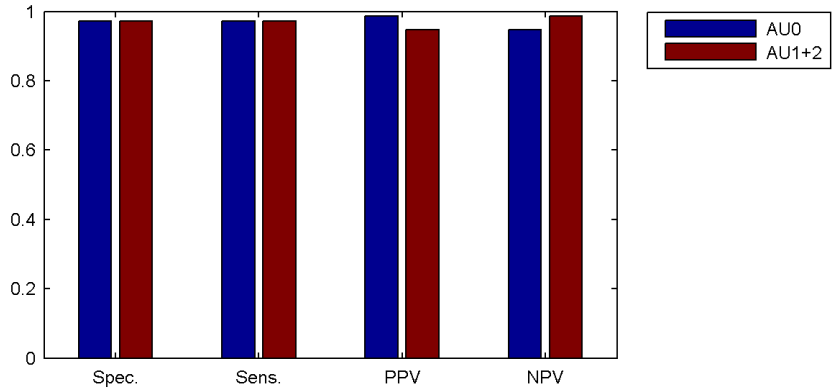

Rysunek C.11: Wpływ skali i rotacji, klasyfikator kNN

## C.9 Statystyczne modele kształtu — test 1

Sprawdzenie skuteczności rozróżniania jednostki AU0 od AU1+2 dla danych jednej osoby (test1).

Tabela C.15: Rozpoznawanie kształtów — test 1, klasyfikator kNN

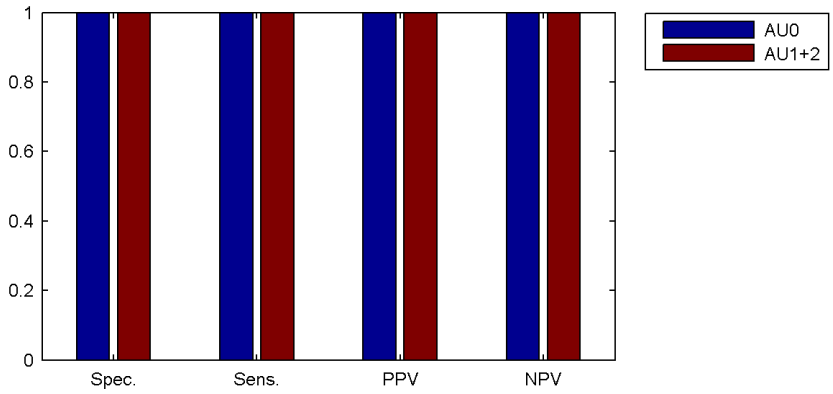
| | **AU0** | **AU1+2** |
|---|---|---|
| **Spec.** | 97.30% | 97.14% |
| **Sens.** | 97.14% | 97.30% |
| **PPV** | 98.55% | 94.74% |
| **NPV** | 94.74% | 98.55% |




Rysunek C.12: Rozpoznawanie kształtów — test 1, klasyfikator kNN

Tabela C.16: Rozpoznawanie kształtów — test 1, klasyfikator LDA

| | **AU0** | **AU1+2** |
|---|---|---|
| **Spec.** | 100.00% | 100.00% |
| **Sens.** | 100.00% | 100.00% |
| **PPV** | 100.00% | 100.00% |
| **NPV** | 100.00% | 100.00% |



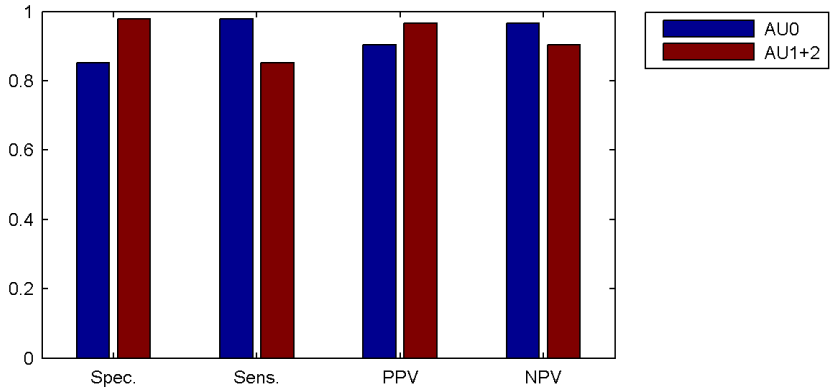

Rysunek C.13: Rozpoznawanie kształtów — test 1, klasyfikator LDA

## C.10 Statystyczne modele kształtu — test 2

Sprawdzenie skuteczności rozróżniania jednostki AU0 od AU1+2 dla danych kilku osób (test2).

Tabela C.17: Rozpoznawanie kształtów — test 2, klasyfikator kNN

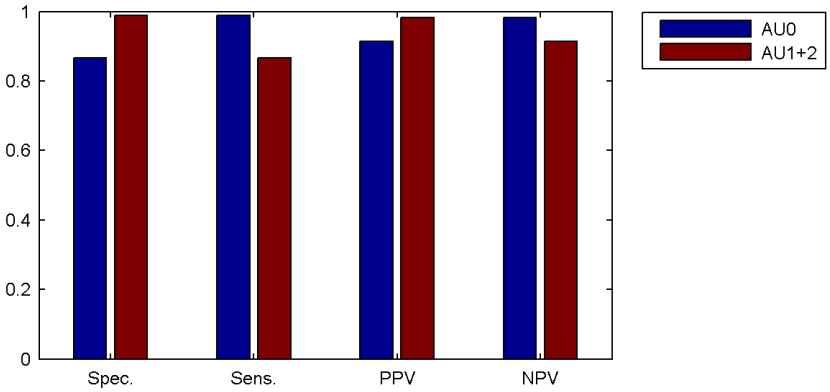
| | **AU0** | **AU1+2** |
|---|---|---|
| **Spec.** | 85.07% | 97.92% |
| **Sens.** | 97.92% | 85.07% |
| **PPV** | 90.38% | 96.61% |
| **NPV** | 96.61% | 90.38% |




Rysunek C.14: Rozpoznawanie kształtów — test 2, klasyfikator kNN

Tabela C.18: Rozpoznawanie kształtów — test 2, klasyfikator LDA

| | **AU0** | **AU1+2** |
|---|---|---|
| **Spec.** | 86.57% | 98.96% |
| **Sens.** | 98.96% | 86.57% |
| **PPV** | 91.35% | 98.31% |
| **NPV** | 98.31% | 91.35% |



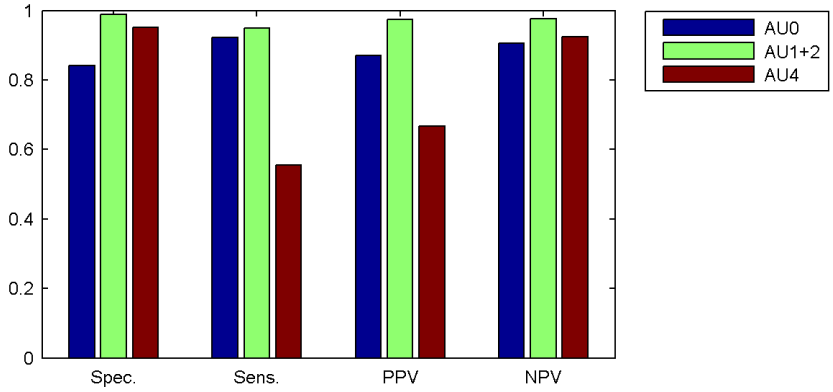

Rysunek C.15: Rozpoznawanie kształtów — test 2, klasyfikator LDA

## C.11 Statystyczne modele kształtu — test 3

Sprawdzenie skuteczności rozróżniania jednostek AU0, AU1+2 oraz AU4 dla danych jednej osoby (test3).

Tabela C.19: Rozpoznawanie kształtów — test 3, klasyfikator kNN

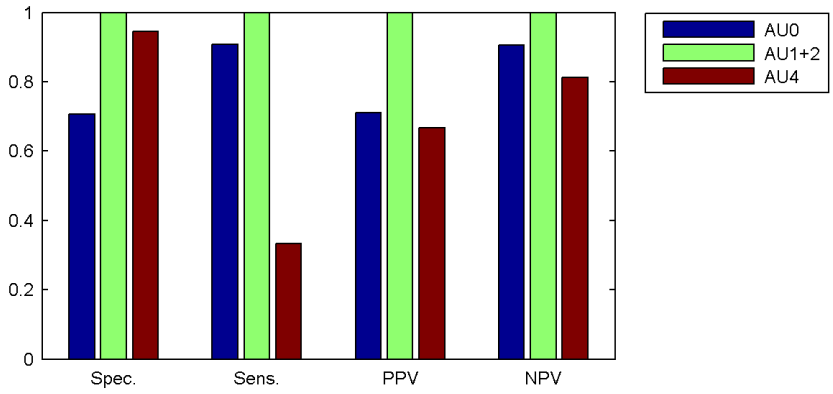
| | **AU0** | **AU1+2** | **AU4** |
|---|---|---|---|
| **Spec.** | 84.21% | 98.80% | 95.19% |
| **Sens.** | 92.31% | 94.87% | 55.56% |
| **PPV** | 86.96% | 97.37% | 66.67% |
| **NPV** | 90.57% | 97.62% | 92.52% |




Rysunek C.16: Rozpoznawanie kształtów — test 3, klasyfikator kNN

Tabela C.20: Rozpoznawanie kształtów — test 3, klasyfikator LDA

| | **AU0** | **AU1+2** | **AU4** |
|---|---|---|---|
| **Spec.** | 70.59% | 100.00% | 94.57% |
| **Sens.** | 90.74% | 100.00% | 33.33% |
| **PPV** | 71.01% | 100.00% | 66.67% |
| **NPV** | 90.57% | 100.00% | 81.31% |



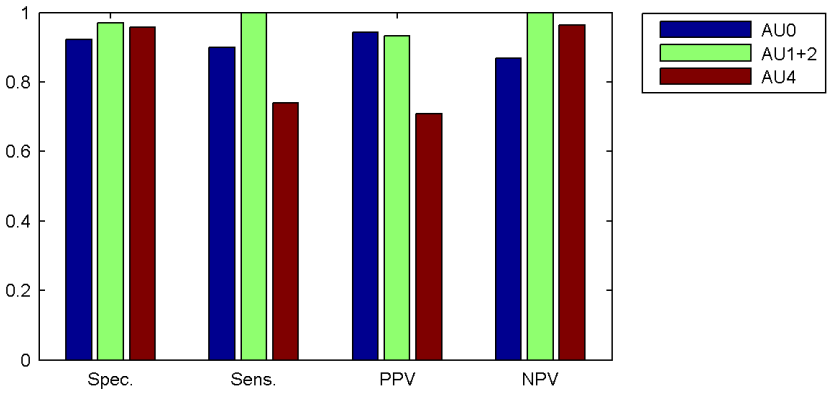

Rysunek C.17: Rozpoznawanie kształtów — test 3, klasyfikator LDA

## C.12 Statystyczne modele kształtu — test 4

Sprawdzenie skuteczności rozróżniania jednostek AU0, AU1+2 oraz AU4 dla danych wielu osób (test4).

Tabela C.21: Rozpoznawanie kształtów — test 4, klasyfikator kNN

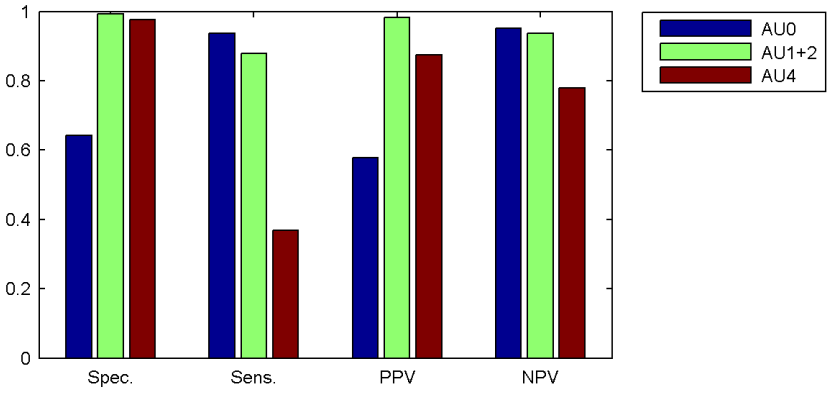
| | **AU0** | **AU1+2** | **AU4** |
|---|---|---|---|
| **Spec.** | 92.31% | 96.97% | 95.73% |
| **Sens.** | 89.91% | 100.00% | 73.91% |
| **PPV** | 94.23% | 93.22% | 70.83% |
| **NPV** | 86.75% | 100.00% | 96.32% |




Rysunek C.18: Rozpoznawanie kształtów — test 4, klasyfikator kNN

Tabela C.22: Rozpoznawanie kształtów — test 4, klasyfikator LDA

| | **AU0** | **AU1+2** | **AU4** |
|---|---|---|---|
| **Spec.** | 64.23% | 99.17% | 97.69% |
| **Sens.** | 93.75% | 87.88% | 36.84% |
| **PPV** | 57.69% | 98.31% | 87.50% |
| **NPV** | 95.18% | 93.75% | 77.91% |




Rysunek C.19: Rozpoznawanie kształtów — test 4, klasyfikator LDA

# Dodatek D

# Rezultaty segmentacji twarzy

W dodatku przedstawiono rezultaty badań, pozwalające na dobór przestrzeni kolorów oraz metody segmentacji twarzy. Badania przeprowadzona na kilku sekwencjach video — opis sekwencji znajduje się w rozdziale (5.2.2). Porównano następujące przestrzenie kolorów:

- znormalizowaną RGB (ozn. r-g),
- HSV (ang. hue-saturation-value) — wybrana składowa barwa (Hue),
- YCbCr — wybrane składowe Cb-Cr.

Spośród metod segmentacji wybrano do porównania następujące:

- bezpośrednie określenie zestawu reguł,
- znormalizowane tablice przekodowań (ang. LUT – look-up-table)
- eliptyczny model gaussa barwy skóry.

Opis dla rysunków:

- (a) zestaw reguł (progi), znormalizowana przestrzeń RGB,
- (b) zestaw reguł (progi), przestrzeń Cb-Cr,
- (c) zestaw reguł (progi), przestrzeń barwy (Hue),
- (d) tablice LUT, znormalizowana przestrzeń RGB,

- (e) tablice LUT, przestrzeń Cb-Cr,
- (f) tablice LUT, przestrzeń Hue,
- (g) metoda parametryczna (eliptyczny model gaussa), znormalizowana przestrzeń RGB,
- (h) metoda parametryczna (eliptyczny model gaussa), przestrzeń Cb-Cr,
- (i) metoda parametryczna (eliptyczny model gaussa), przestrzeń Hue

Tabela D.1: Wartości funkcji celu dla wybranych metod i przestrzeni barw — sekwencja nr 1

| | **r-g** | **Cb-Cr** | **Hue** |
|---|---|---|---|
| **Metoda regułowa** | -0.86395 | -0.87263 | -0.81603 |
| **Tablice LUT** | -0.85251 | -0.86907 | -0.84857 |
| **Model gaussowski** | -0.73357 | -0.78978 | -0.83666 |

Tabela D.2: Wartości funkcji celu dla wybranych metod i przestrzeni barw — sekwencja nr 2

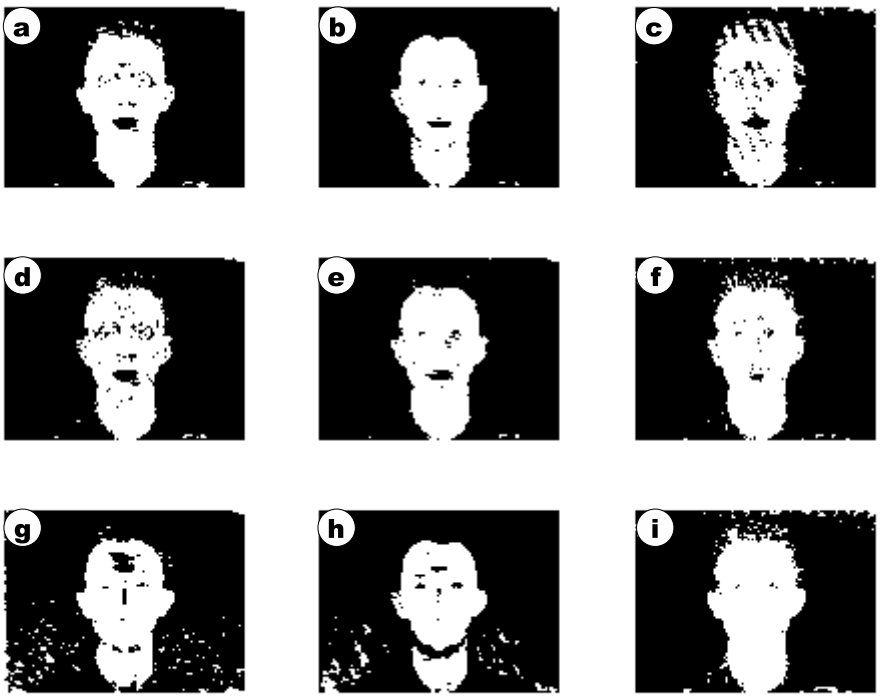
| | **r-g** | **Cb-Cr** | **Hue** |
|---|---|---|---|
| **Metoda regułowa** | -0.7293 | -0.70078 | -0.3179 |
| **Tablice LUT** | -0.55755 | -0.32397 | -0.32244 |
| **Model gaussowski** | -0.58985 | -0.56684 | -0.33678 |

Tabela D.3: Wartości funkcji celu dla wybranych metod i przestrzeni barw — sekwencja nr 3

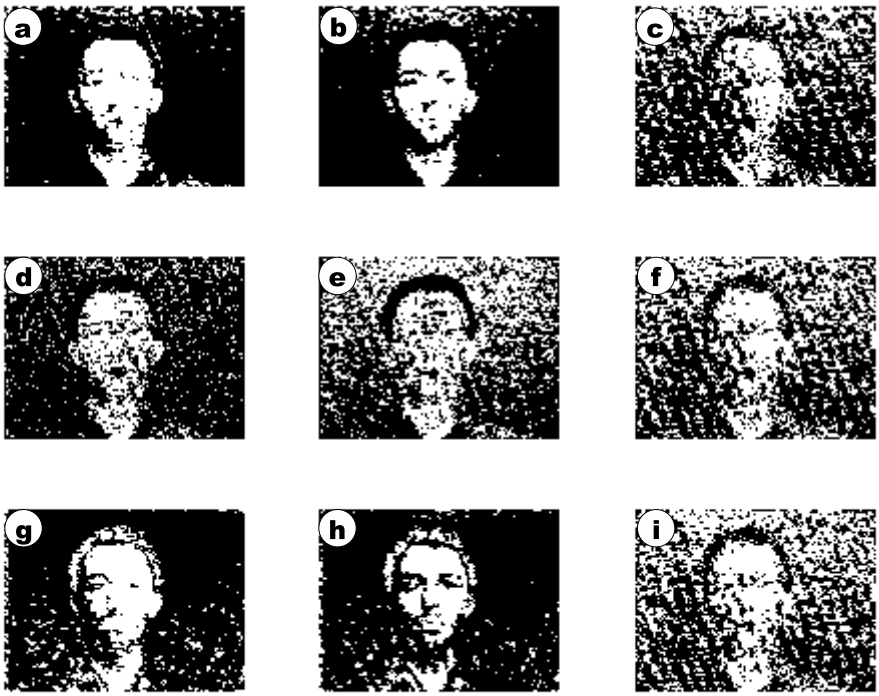
| | **r-g** | **Cb-Cr** | **Hue** |
|---|---|---|---|
| **Metoda regułowa** | -0.70611 | -0.51838 | -0.56166 |
| **Tablice LUT** | -0.73582 | -0.6128 | -0.60464 |
| **Model gaussowski** | -0.3341 | -0.25709 | -0.63027 |

Tabela D.4: Wartości funkcji celu dla wybranych metod i przestrzeni barw — sekwencja nr 4

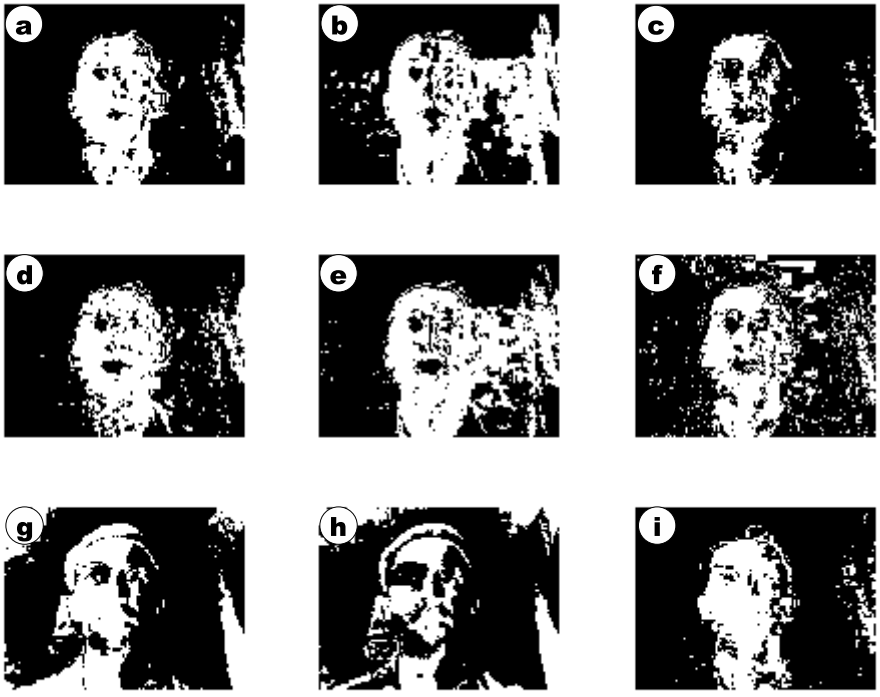
| | **r-g** | **Cb-Cr** | **Hue** |
|---|---|---|---|
| **Metoda regułowa** | -0.34594 | -0.2755 | -0.41012 |
| **Tablice LUT** | -0.48391 | -0.30291 | -0.44862 |
| **Model gaussowski** | -0.17463 | -0.27584 | -0.50289 |



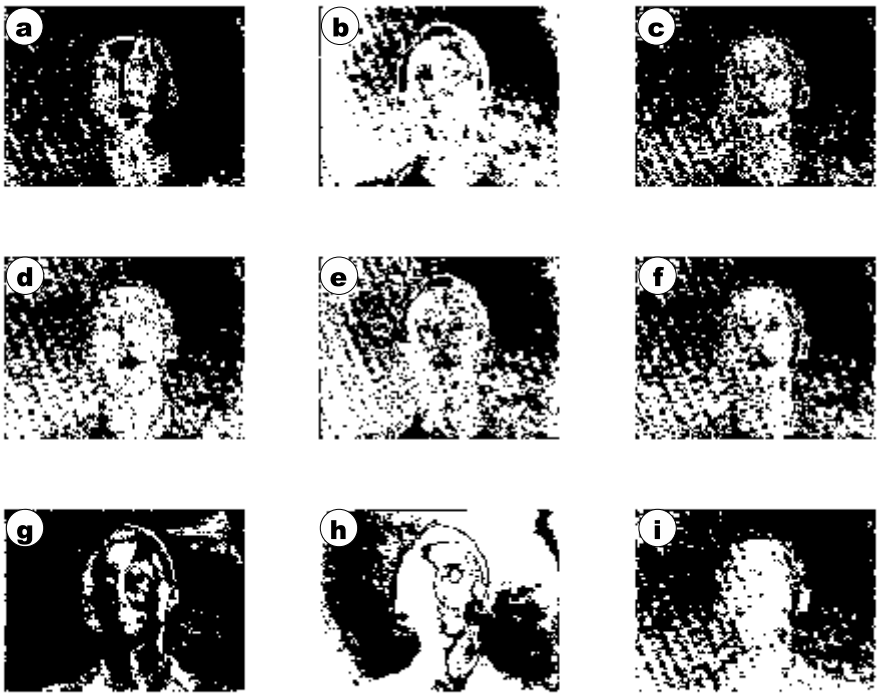

Rysunek D.1: Rezultaty segmentacji twarzy dla sekwencji nr 1.




Rysunek D.2: Rezultaty segmentacji twarzy dla sekwencji nr 2.




Rysunek D.3: Rezultaty segmentacji twarzy dla sekwencji nr 3.




Rysunek D.4: Rezultaty segmentacji twarzy dla sekwencji nr 4.

# Dodatek E

# Rezultaty detekcji i lokalizacji twarzy

Parametry jakości detekcji twarzy (E.1)—(E.4).

$$PPV = \frac{TP}{TP + FP} \cdot 100\% \tag{E.1}$$

$$NPV = \frac{TN}{FN + TN} \cdot 100\% \tag{E.2}$$

$$Sens = \frac{TP}{TP + FN} \cdot 100\% \tag{E.3}$$

$$Spec = \frac{TN}{FP + TN} \cdot 100\% \tag{E.4}$$

gdzie:

$PPV$ – wartość predykcyjna dodatnia klasyfikacji (ang. positive predictive value)

$NPV$ – wartość predykcyjna ujemna klasyfikacji (ang. negative predictive value)

$Sens$ – czułość (ang. sensitivity)

$Spec$ – swoistość (ang. specificity)

$TP$ – wyniki prawdziwie dodatnie (ang. true positives)

$FN$ – wyniki fałszywie ujemne (ang. false negatives)

$FP$ – wyniki fałszywie dodatnie (ang. false positives)

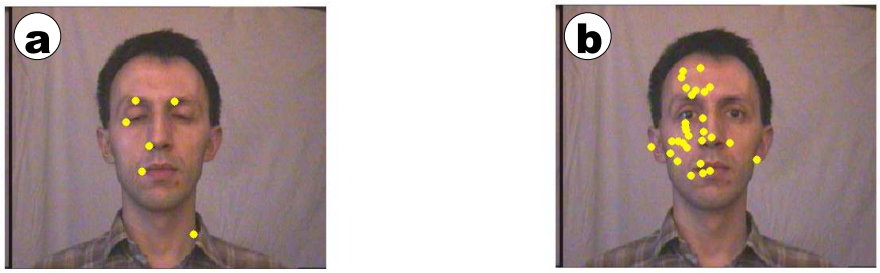
$TN$ – wyniki prawdziwie ujemne (ang. true negatives)

Tabela E.1: Liczebność poszczególnych grup obrazów dla sekwencji 1-3

| | **Sekwencja 1** | **Sekwencja 2** | **Sekwencja 3** |
|---|---|---|---|
| **Ilość obrazów całej sekwencji** | 532 | 424 | 170 |
| **Ilość obrazów dla położenia frontalnego** | 278 | 226 | 105 |

Tabela E.2: Detekcja i lokalizacja twarzy — rezultaty dla położenia frontalnego

| | **Sekwencja 1** | **Sekwencja 2** | **Sekwencja 3** |
|---|---|---|---|
| **Spec.** | 99.35% | 77.98% | 65.51% |
| **Sens.** | 73.09% | 73.83% | 75.89% |
| **PPV** | 99.64% | 83.63% | 80.95% |
| **NPV** | 59.84% | 66.16% | 58.46% |




Rysunek E.1: Rezultaty detekcji (dla wybranego obrazu) potencjalnych obiektów: (a) oczu, (b) ust

# Dodatek F

# Rezultaty estymacji położenia głowy — wykresy i tabele

## F.1 Informacje ogólne

Parametry oceny jakości rozpoznawania położenia głowy (F.1)—(F.4).

$$PPV = \frac{TP}{TP + FP} \cdot 100\% \tag{F.1}$$

$$NPV = \frac{TN}{FN + TN} \cdot 100\% \tag{F.2}$$

$$Sens = \frac{TP}{TP + FN} \cdot 100\% \tag{F.3}$$

$$Spec = \frac{TN}{FP + TN} \cdot 100\% \tag{F.4}$$

gdzie:

$PPV$ – wartość predykcyjna dodatnia klasyfikacji (ang. positive predictive value)

$NPV$ – wartość predykcyjna ujemna klasyfikacji (ang. negative predictive value)

$Sens$ – czułość (ang. sensitivity)

$Spec$ – swoistość (ang. specificity)

$TP$ – wyniki prawdziwie dodatnie (ang. true positives)

$FN$ – wyniki fałszywie ujemne (ang. false negatives)

$FP$ – wyniki fałszywie dodatnie (ang. false positives)

$TN$ – wyniki prawdziwie ujemne (ang. true negatives)

Liczebność obrazów sekwencji poszczególnych ruchów głowy dla poszczególnych sekwencji (opis sekwencji w rozdziale 8.2.2) jest następująca:

Tabela F.1: Liczebność obrazów sekwencji

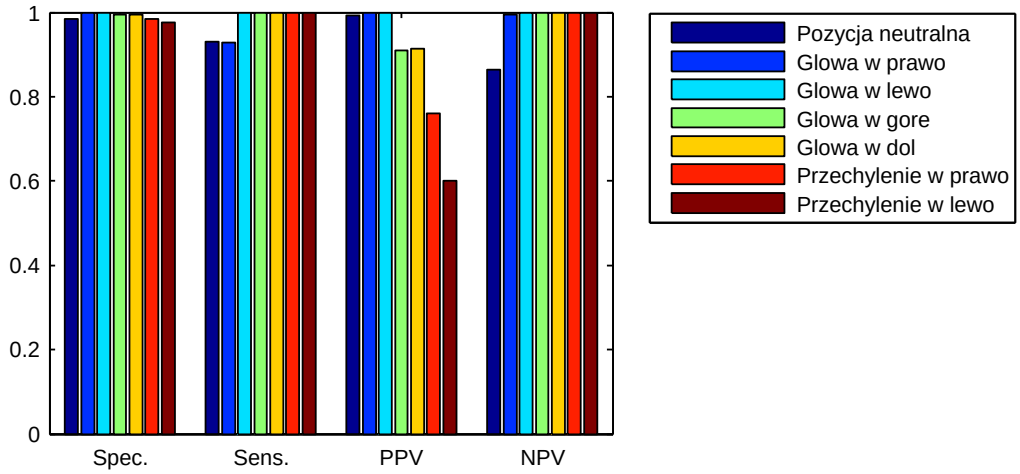
| | **Sekw. wzorcowa** | **Sekw. nr 1** | **Sekw. nr 2** | **Sekw. nr 3** |
|---|---|---|---|---|
| **Pozycja neutralna** | 351 | 274 | 617 | 268 |
| **Głowa w prawo** | 39 | 26 | 70 | 31 |
| **Głowa w lewo** | 37 | 27 | 58 | 28 |
| **Głowa w górę** | 17 | 22 | 57 | 21 |
| **Głowa w dół** | 22 | 23 | 37 | 19 |
| **Przechylenie w prawo** | 27 | 25 | 40 | 19 |
| **Przechylenie w lewo** | 37 | 25 | 51 | 23 |

## F.2 Sekwencja nr 1 — rezultaty

Rezultaty estymacji położenia głowy dla sekwencji testowej nr 1

Tabela F.2: Rezultaty dla sekwencji nr 1

| | **Spec.** | **Sens.** | **PPV.** | **NPV** |
|---|---|---|---|---|
| **Pozycja neutralna** | 98.46% | 93.15% | 99.27% | 86.49% |
| **Głowa w prawo** | 100.00% | 92.86% | 100.00% | 99.49% |
| **Głowa w lewo** | 100.00% | 100.00% | 100.00% | 100.00% |
| **Głowa w górę** | 99.50% | 100.00% | 90.91% | 100.00% |
| **Głowa w dół** | 99.50% | 100.00% | 91.30% | 100.00% |
| **Przechylenie w prawo** | 98.51% | 100.00% | 76.00% | 100.00% |
| **Przechylenie w lewo** | 97.54% | 100.00% | 60.00% | 100.00% |



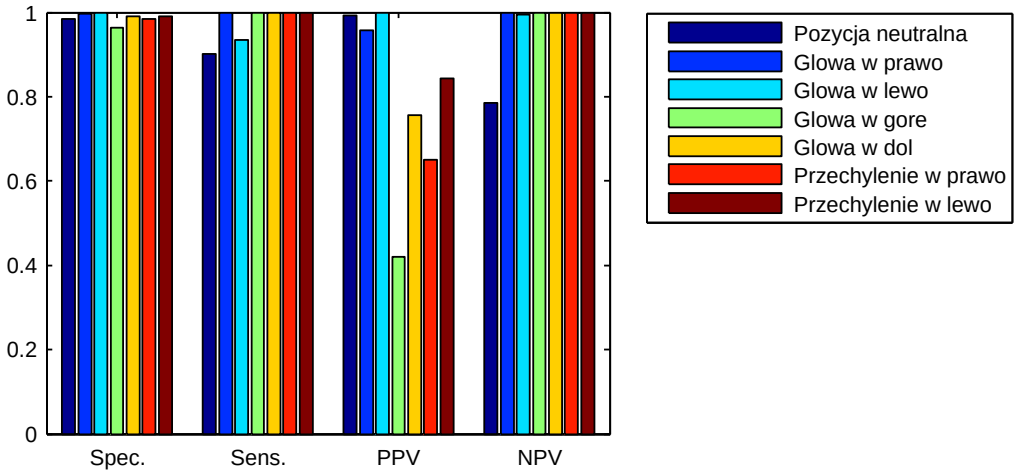

Rysunek F.1: Rezultaty dla sekwencji nr 1

## F.3 Sekwencja nr 2 — rezultaty

Rezultaty estymacji położenia głowy dla sekwencji testowej nr 2

Tabela F.3: Rezultaty dla sekwencji nr 2

| | **Spec.** | **Sens.** | **PPV.** | **NPV** |
|---|---|---|---|---|
| **Pozycja neutralna** | 98.40% | 90.15% | 99.35% | 78.59% |
| **Głowa w prawo** | 99.65% | 100.00% | 95.71% | 100.00% |
| **Głowa w lewo** | 100.00% | 93.55% | 100.00% | 99.54% |
| **Głowa w górę** | 96.36% | 100.00% | 42.11% | 100.00% |
| **Głowa w dół** | 99.00% | 100.00% | 75.68% | 100.00% |
| **Przechylenie w prawo** | 98.45% | 100.00% | 65.00% | 100.00% |
| **Przechylenie w lewo** | 99.10% | 100.00% | 84.31% | 100.00% |



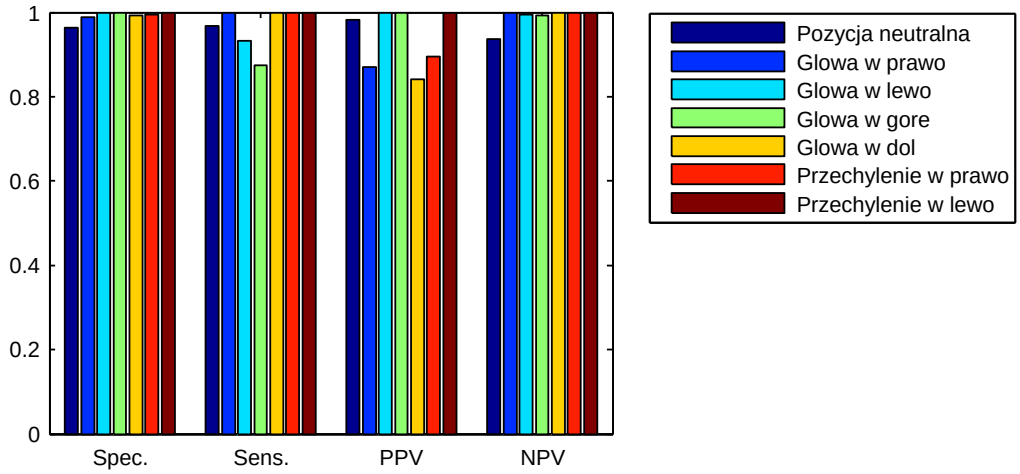

Rysunek F.2: Rezultaty dla sekwencji nr 2

## F.4 Sekwencja nr 3 — rezultaty

Rezultaty estymacji położenia głowy dla sekwencji testowej nr 3

Tabela F.4: Rezultaty dla sekwencji nr 3

| | **Spec.** | **Sens.** | **PPV.** | **NPV** |
|---|---|---|---|---|
| **Pozycja neutralna** | 96.35% | 96.69% | 98.13% | 93.62% |
| **Głowa w prawo** | 98.95% | 100.00% | 87.10% | 100.00% |
| **Głowa w lewo** | 100.00% | 93.33% | 100.00% | 99.48% |
| **Głowa w górę** | 100.00% | 87.50% | 100.00% | 99.23% |
| **Głowa w dół** | 99.24% | 100.00% | 84.21% | 100.00% |
| **Przechylenie w prawo** | 99.49% | 100.00% | 89.47% | 100.00% |
| **Przechylenie w lewo** | 100.00% | 100.00% | 100.00% | 100.00% |




Rysunek F.3: Rezultaty dla sekwencji nr 3

# Dodatek G

# Rezultaty detekcji mrugnięć

## G.1 Informacje ogólne

Parametry oceny skuteczności detekcji mrugnięć (G.1)—(G.4).

$$PPV = \frac{TP}{TP + FP} \cdot 100\% \tag{G.1}$$

$$NPV = \frac{TN}{FN + TN} \cdot 100\% \tag{G.2}$$

$$Sens = \frac{TP}{TP + FN} \cdot 100\% \tag{G.3}$$

$$Spec = \frac{TN}{FP + TN} \cdot 100\% \tag{G.4}$$

gdzie:

$PPV$ – wartość predykcyjna dodatnia klasyfikacji (ang. positive predictive value)

$NPV$ – wartość predykcyjna ujemna klasyfikacji (ang. negative predictive value)

$Sens$ – czułość (ang. sensitivity)

$Spec$ – swoistość (ang. specificity)

$TP$ – wyniki prawdziwie dodatnie (ang. true positives)

$FN$ – wyniki fałszywie ujemne (ang. false negatives)

$FP$ – wyniki fałszywie dodatnie (ang. false positives)

$TN$ – wyniki prawdziwie ujemne (ang. true negatives)

Testy wykonano dla następujących sekwencji, z których wybrano obrazy dla położenia twarzy w pozycji w przybliżeniu frontalnej:

- sekwencja 1 — kamera EVI, jednolite tło, optymalne warunki oświetlenia (natężenie oświetlenia około 320lux),(rys. G.1a)
- sekwencja 2 — kamera EVI, jednolite tło, słabe oświetlenie (natężenie oświetlenia około 40lux), (rys. G.1b),
- sekwencja 3 — kamera EVI, jednolite tło, optymalne warunki oświetlenia lekko różniąca się od sekwencji nr 1 (natężenie oświetlenia około 320lux), (rys. G.1c),
- sekwencja 4 — kamera EVI, jednolite tło, optymalne warunki oświetlenia (natężenie oświetlenia około 320lux), inna mimika twarzy (rys. G.1d),


Rysunek G.1: Przykładowe obrazy z sekwencji video: (a) sekwencja nr 1, (b) sekwencja nr 2, (c) sekwencja nr 3, (d) sekwencja nr 4

Liczebność wybranych obrazów z poszczególnych sekwencji jest następująca:

Tabela G.1: Liczebność obrazów sekwencji

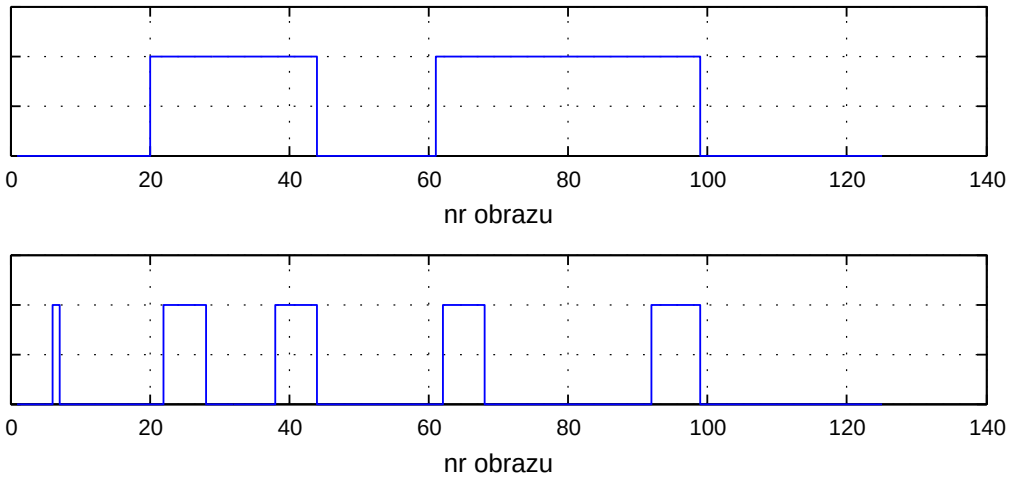
| | **Sekwencja nr 1** | **Sekwencja nr 2** | **Sekwencja nr 3** | **Sekwencja nr 4** |
|---|---|---|---|---|
| **Pozycja neutralna** | 125 | 105 | 83 | 93 |
| **Gest mrugania** | 62 | 50 | 50 | 46 |

## G.2 Skuteczność detekcji

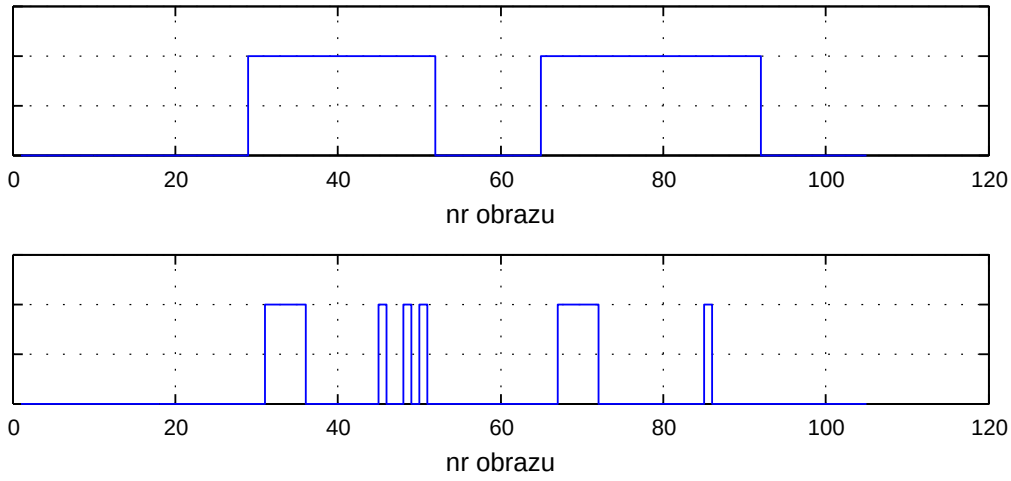
Rezultaty detekcji mrugnięć dla poszczególnych sekwencji.

Tabela G.2: Rezultaty detekcji mrugnięć

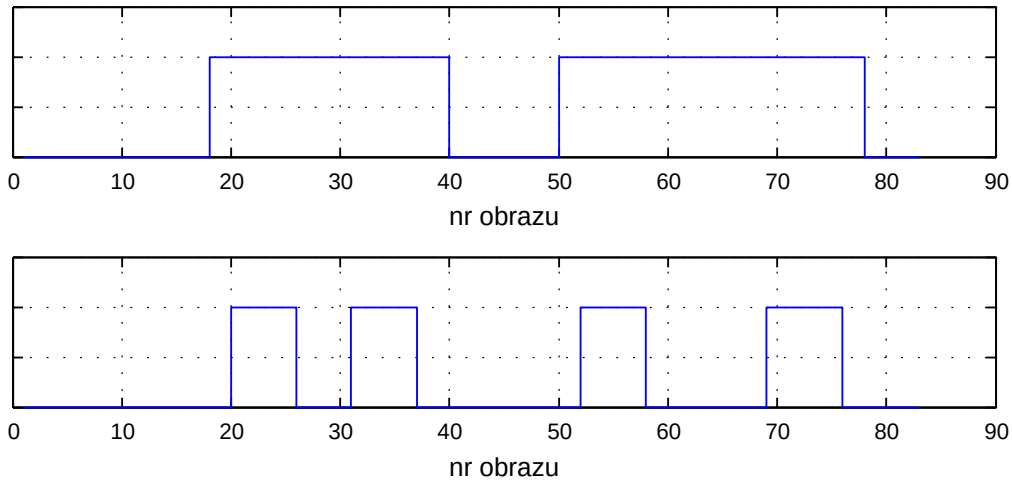
| | **Sekwencja nr 1** | **Sekwencja nr 2** | **Sekwencja nr 3** | **Sekwencja nr 4** |
|---|---|---|---|---|
| **TP** | 25 | 14 | 25 | 30 |
| **FP** | 1 | 0 | 0 | 10 |
| **TN** | 62 | 55 | 33 | 37 |
| **FN** | 37 | 36 | 25 | 16 |
| **Spec.** | 98.41% | 100% | 100% | 78.72% |
| **Sens.** | 40.32% | 28.00% | 50.00% | 65.21% |
| **PPV** | 96.15% | 100% | 100% | 75.00% |
| **NPV** | 62.62% | 60.44% | 56.90% | 69.81% |



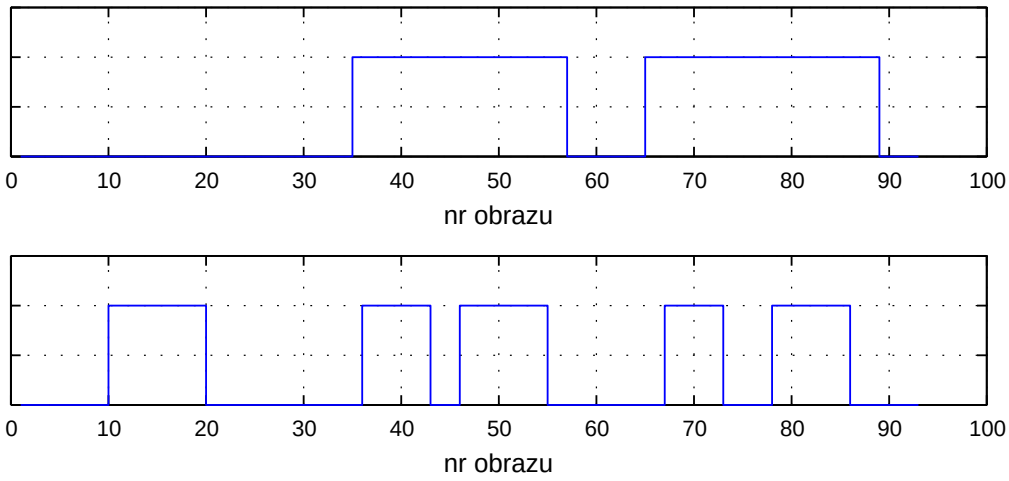

Rysunek G.2: Ręczna adnotacja obrazów gestu mrugania (górny wykres) oraz rezultaty detekcji mrugnięcia (dolny wykres) dla sekwencji nr 1




Rysunek G.3: Ręczna adnotacja obrazów gestu mrugania (górny wykres) oraz rezultaty detekcji mrugnięcia (dolny wykres) dla sekwencji nr 2




Rysunek G.4: Ręczna adnotacja obrazów gestu mrugania (górny wykres) oraz rezultaty detekcji mrugnięcia (dolny wykres) dla sekwencji nr 3




Rysunek G.5: Ręczna adnotacja obrazów gestu mrugania (górny wykres) oraz rezultaty detekcji mrugnięcia (dolny wykres) dla sekwencji nr 4

## G.3 Średni błąd lokalizacji oczu

Średni błąd lokalizacji oczu dla poszczególnych sekwencji.

Tabela G.3: Średni błąd (w pikselach) lokalizacji oczu

| | **Sekwencja nr 1** | **Sekwencja nr 2** | **Sekwencja nr 3** | **Sekwencja nr 4** |
|---|---|---|---|---|
| $\varepsilon_{eye}$ | 8.143 | 6.331 | 6.9525 | 5.6961 |

# Dodatek H

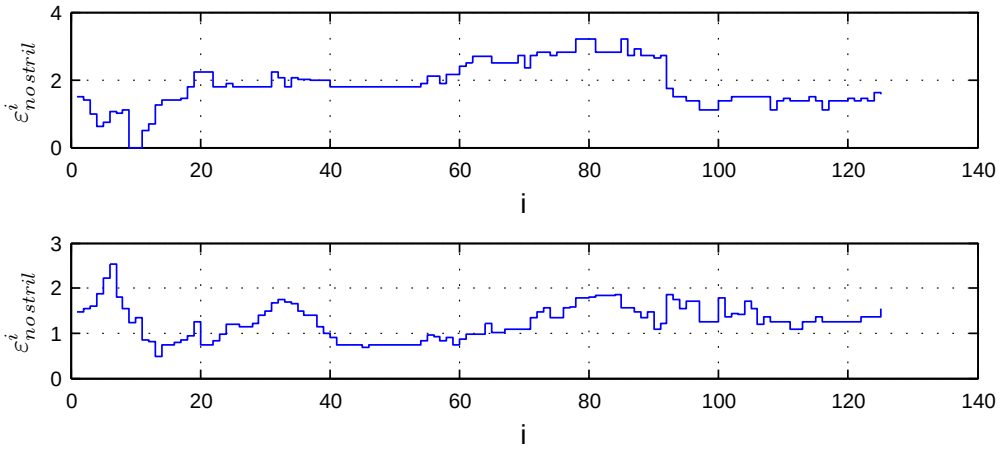
# Rezultaty innych algorytmów

Dodatek zawiera rezultaty działania następujących algorytmów:

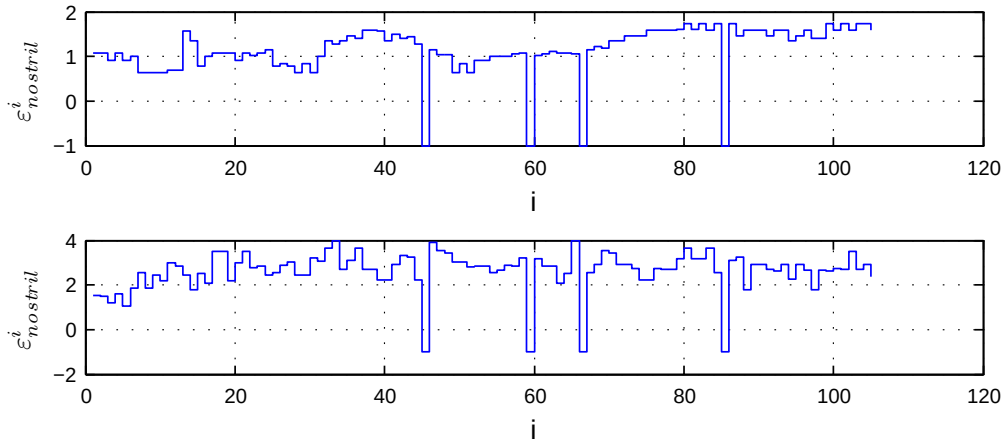
- Lokalizacja nozdrzy.

## H.1 Lokalizacja nozdrzy

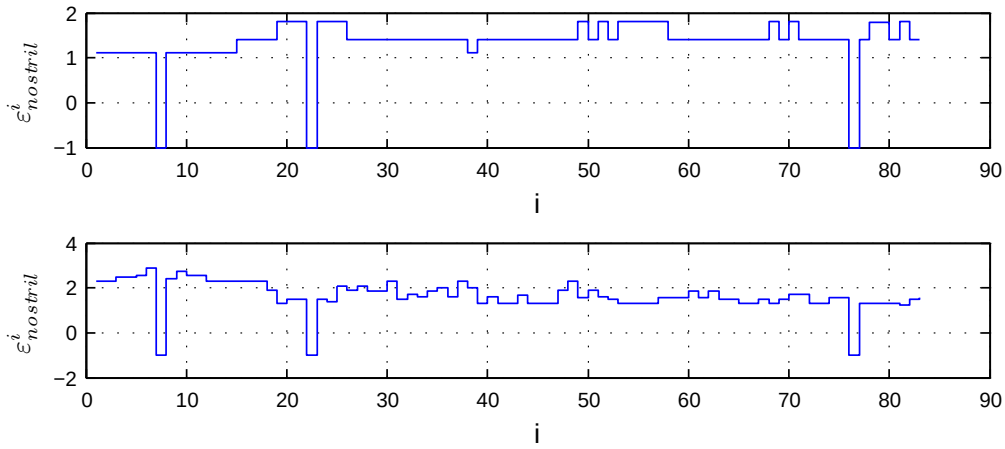
Błąd lokalizacji (w pikselach) został wyznaczony dla sekwencji video opisanych w dodatku G.1. Wartość błędu poniżej zera oznacza, że dla danego obrazu nozdrza nie zostały poprawnie znalezione.



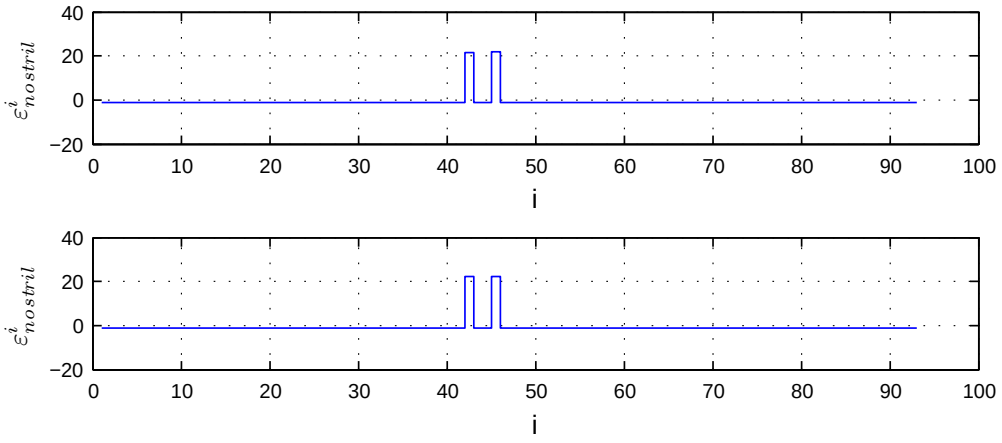

Rysunek H.1: Błąd lokalizacji nozdrzy dla sekwencji nr 1: nozdrze prawe (górny wykres), nozdrze lewe (dolny wykres)




Rysunek H.2: Błąd lokalizacji nozdrzy dla sekwencji nr 2: nozdrze prawe (górny wykres), nozdrze lewe (dolny wykres)




Rysunek H.3: Błąd lokalizacji nozdrzy dla sekwencji nr 3: nozdrze prawe (górny wykres), nozdrze lewe (dolny wykres)




Rysunek H.4: Błąd lokalizacji nozdrzy dla sekwencji nr 4: nozdrze prawe (górny wykres), nozdrze lewe (dolny wykres)