AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA
STASZICA W KRAKOWIE

Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki
Kierunek: Informatyka Stosowana

Praca magisterska

Tomasz Kotliński

Automatyczne śledzenie twarzy i rozpoznawanie gestów na twarzy użytkownika komputera siedzącego przed kamerą.

Promotor: dr Adrian Horzyk

Kraków, 2009

Oświadczam, świadomy odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonałem osobiście i samodzielnie i że nie korzystałem ze źródeł innych niż wymienione w pracy.

Pragnę złożyć podziękowania
Panu dr Adrianowi Horzykowi
za wszystkie cenne wskazówki i rady
udzielone podczas pisania pracy.

Spis treści

Rozdział 1

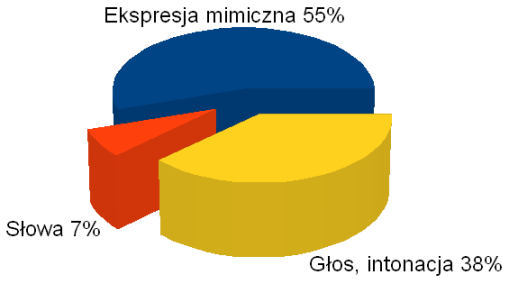
Wstęp

Gdy tworzono pierwsze komputery (właściwszym określeniem byłoby „maszyny liczące"), nikt nie przypuszczał takiego rozwoju, który doprowadzi do ich wszechobecności w naszym codziennym życiu. Komputer już dawno przestał być tylko narzędziem wykorzystywanym przez naukowców i centra badawcze; technologia mikroprocesorowa pod różnymi postaciami wkracza w coraz to nowe dziedziny życia. Przez wiele lat postęp ograniczał się jednak głównie do wzrostu mocy obliczeniowej i ciągłej miniaturyzacji. Wciąż to człowiek musiał dostosowywać się do ograniczeń maszyny, nie odwrotnie. Użytkowanie komputera nie jest rzeczą intuicyjną, nawet pomimo powstawania coraz bardziej „przyjaznego" oprogramowania. Nadal wymaga nauki i pewnego szkolenia. Widać to najdobitniej na przykładzie osób starszych, rozpoczynających obsługę komputera.

Dlatego też w ostatnich latach intensywnie prowadzone są badania z zakresu ułatwienia interakcji człowiek-komputer (*HCI, Human-Computer Interaction*). Obejmują one zarówno opracowanie nowych, bardziej intuicyjnych metod wprowadzania danych (np. dyktowanie tekstu), ich prezentacji (np. wirtualna rzeczywistość), jak również wykorzystywanie nowych, do tej pory nieużywanych kanałów komunikacji. Do tych ostatnich można zaliczyć na przykład przekazywanie bodźców dotykowych (interfejsy haptyczne, *force feedback*). Wszystkie te zabiegi mają na celu uczynienie bardziej naturalną komunikacji człowieka z komputerem, poprzez obdarzenie tego drugiego takimi zmysłami i środkami wyrazu, aby praca z nim jak najbardziej przypominała pracę z człowiekiem.

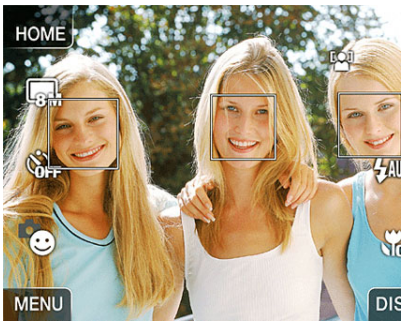
Jedną z fascynujących umiejętności, jaką posiadają ludzie, jest zdolność do komunikacji niewerbalnej. Okazuje się, iż umiemy doskonale porozumiewać się bez użycia słów. Przykładowo, podczas rozmowy znaczenie naszej wypowiedzi dla odbiorcy jest określane głównie przez intonację i gesty [5]. Rysunek 1.1 prezentuje szacowany rozkład znaczenia komunikatu.

Dlatego też coraz częściej zauważa się potrzebę konstruowania systemów ana-



Rysunek 1.1: Rozkład znaczenia komunikatu (źródło: [5])

lizujących mimikę użytkownika. Uzyskiwane rezultaty są obiecujące, aczkolwiek daleko jeszcze maszynom do osiągnięcia ludzkiej sprawności w tej dziedzinie. Mimo to, znaleziono już komercyjne zastosowanie takich systemów – przykładem niech będzie oprogramowanie wbudowane w aparat fotograficzny Sony DSC-T200, które śledzi twarz portretowanej osoby i wyzwala migawkę w chwili, gdy osoba ta się uśmiechnie [11] (funkcja *Smile Shutter*).



Rysunek 1.2: System *Smile Shutter* w działaniu (źródło: [12])

Innym potencjalnym zastosowaniem takich systemów jest ułatwienie rozmów prowadzonych poprzez komunikatory internetowe. Z jednej strony, gdy stosowanie transmisji audio-video jest ograniczone przepustowością łącza, użytkownicy mieliby możliwość poznania nawzajem swoich emocji i reakcji na rozmowę dzięki zastosowaniu komunikatora analizującego mimikę jednego z rozmówców i przekazywaniu informacji o niej drugiemu (i vice versa). Z drugiej strony, nie jest powiedziane, iż obydwie strony takiej konwersacji muszą być ludźmi... Twórcy tzw. *chatbotów* usiłują sprawić, aby ich programy potrafiły prowadzić rozmowę w jak najbardziej „ludzki” sposób. Reakcja takiego sztucznego rozmówcy na emocje wyrażane przez człowieka przyniosłaby znaczący postęp w ich wysiłkach.

W niniejszej pracy starano się bliżej przedstawić aspekty techniczne tego zagadnienia. Powołano się przy tym na przykłady istniejących rozwiązań a także zbudowano własny system, realizujący wspomniane cele.

Rozdział 2

Wprowadzenie

2.1 Analiza zagadnienia

2.1.1 Aspekty fizjologiczne i behawioralne

Jak już wspomniano, podczas komunikacji międzyludzkiej mamy do czynienia z wieloma rodzajami bodźców. Jednym z nich jest widok twarzy naszego rozmówcy. Często jest to wręcz jedyna informacja, jaką otrzymujemy (spojrzenie na twarz przypadkowego przechodnia dużo mówi nam o jego stanie emocjonalnym). Okazuje się, iż umiejętność komunikacji za pomocą gestów mimicznych jest uniwersalna - wrodzona, niezależna od kultury, wychowania czy środowiska. Tezę taką zaproponował już Darwin [7], [6], natomiast potwierdził swoimi badaniami Paul Ekman. Stworzył on również listę podstawowych emocji, wyrażanych mimicznie [8]:

- smutek,
- złość (gniew),
- zaskoczenie,
- strach,
- radość,
- niesmak (wstręt).

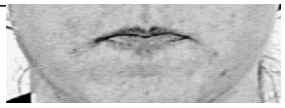
Lista ta została później rozszerzona o 9 dodatkowych, głównie pozytywnych emocji.

Badania te doprowadziły również do utworzenia specjalnej taksonomii możliwych jednostek ekspresji mimicznej. System ten nazwano FACS (*Facial Action*

Coding System) [1],[14]. Przyjęto go jako obowiązujący standard w m.in. psychologii. Definiuje on 32 tzw. AU (*Action Units*) odpowiadające ruchom poszczególnych mięśni twarzy. Wystąpienie określonego gestu mimicznego polega zazwyczaj na jednoczesnym wykonaniu kilku AU. Dodatkowo, dla innych AU, dla których nie zdefiniowano odpowiadającym im mięśni, przyjęto określenie *Action Descriptor*. Tabela 2.1 przedstawia wykaz opisanych AU.

Tabela 2.1: Zestawienie opisanych AU (źródło: [10])

AU	**Nazwa oryginalna**	**Polska nazwa**	**Przykład**
1	Inner Brow Raiser	Uniesienie wewnętrznej części brwi	
2	Outer Brow Raiser	Uniesienie zewnętrznej części brwi	
4	Brow Lowerer	Opuszczenie brwi	
5	Upper Lid Raiser	Uniesienie górnych powiek	
6	Cheek Raiser	Uniesienie policzków	
7	Lid Tightener	Napięcie powiek	
9	Nose Wrinkle	Marszczenie nosa	
10	Upper Lip Raiser	Uniesienie górnej wargi	
11	Nasolabial Deepener	Pogłębienie części nosowo-wargowej	
12	Lip Corner Puller	Ściągnięcie kącików ust	

AU	Nazwa oryginalna	Polska nazwa	Przykład
13	Cheek Puffer	Wydęcie policzków	
14	Dimpler	Dołki w policzkach	
15	Lip Corner Depressor	Obniżenie kącików ust	
16	Lower Lip Depressor	Obniżenie dolnej wargi	
17	Chin Raiser	Uniesienie podbródka	
18	Lip Puckerer	Ściągnięcie warg	
20	Lip Stretcher	Rozciągnięcie warg	
22	Lip Funneler	Wargi ułożone w lejek	
23	Lip Tightener	Napięcie warg	
24	Lip Pressor	Ściśnięcie warg	
25	Lips part	Wargi rozchylone	

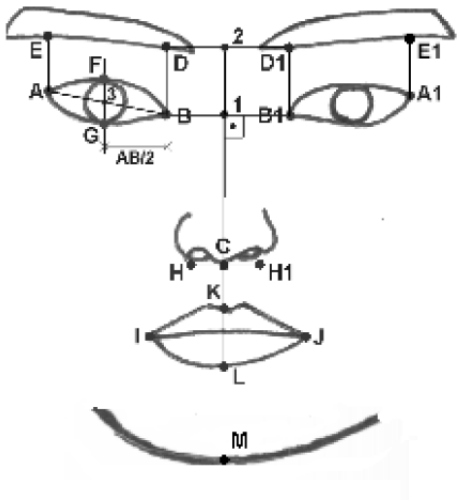
AU	Nazwa oryginalna	Polska nazwa	Przykład
26	Jaw Drop	Szczęka opadnięta	
27	Mouth Stretch	Rozciągnięcie ust	
28	Lip Suck	Wargi wessane	
41	Lid droop	Powieki opadnięte	
42	Slit	Oczy w szparki	
43	Eyes Closed	Oczy zamknięte	
44	Squint	Zezowanie	
45	Blink	Zmrużenie oczu	
46	Wink	Mrugnięcie oczami	

Zgodnie z założeniami systemu FACS, stan emocjonalny wyrażany poprzez ekspresję mimiczną można odczytać, analizując obecność określonych AU. Każde z nich charakteryzuje się pewnym określonym napięciem mięśni poszczególnych części twarzy. To z kolei przekłada się na specyficzne dla każdego AU ułożenie ust, policzków, brwi, powiek itp. Co więcej, jeżeli wziąć pod uwagę anatomię twarzy, można zauważyć, iż aby określić kształt danej części, wystarczy znać wzajemne położenie pewnych punktów węzłowych. Takie punkty to na przykład: kąciki oczu i ust, czubek brody czy końce brwi. Obrazuje to rysunek 2.1.

2.1.2 Aspekty techniczne

System rozpoznający emocje wyrażane poprzez mimikę twarzy można zaliczyć do szerszej grupy systemów automatycznego rozpoznawania obrazów. Łączy go z nimi także pewien konieczny do przyjęcia schemat funkcjonowania i kolejność operacji:

1. Akwizycja obrazu (*image acquisition*).



Rysunek 2.1: Punkty węzłowe twarzy wykorzystywane w analizie mimiki (źródło: [18])

2. Wstępne przetwarzanie obrazu (*preprocessing*).

3. Ekstrakcja parametrów i cech właściwych dla zadania stawianego systemowi (*feature extraction*).

4. Klasyfikacja obrazu do danej grupy na podstawie danych uzyskanych w punkcie poprzednim (*classification, recognition*).

Pokrótce omówione zostaną poszczególne etapy, ze wskazaniem stosowanych i opisanych rozwiązań.

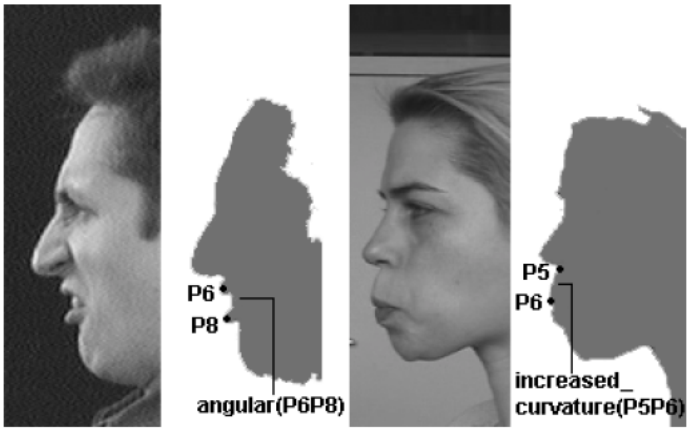
Akwizycja obrazu

Aby wprowadzić informację wizyjną do systemu, należy zarejestrować obraz (sekwencję obrazów). Jest to etap zarazem najprostszy, jak i wymagający dużo uwagi, gdyż determinuje rodzaj danych, podlegających dalszemu przetwarzaniu. Ponadto w tym miejscu zostaje zdeterminowane przeznaczenie systemu: analiza statycznych obrazów (zdjęć) bądź ich sekwencji (sygnał wideo). Z racji różnic technicznych w sposobie pozyskiwania takich obrazów, mogą one być różnej jakości. Z zasady zdjęcia dostarczają dużo lepszej rozdzielczości i są pozbawione zakłóceń, co znacznie ułatwia ich analizę. Z kolei sekwencja obrazów niesie dodatkową informację czasową o analizowanych gestach. Okazuje się, że w rozpoznawaniu gestów może być przydatna obserwacja ich dynamiki. Badacze wyróżniają fazy gestów: narastania, trwania i opadania [27]. Informacja o czasach ich trwania jest przydatna m.in. w analizie intencji gestu (szczery bądź udawany). Zważywszy

na tematykę niniejszej pracy starano się skupić na pozyskaniu danych z obrazu wideo.

Zazwyczaj wykorzystuje się do tego celu zwykłą kamerę barwną wyposażoną w przetwornik CCD. Kamery monochromatyczne są również stosowane, lecz już na wstępie pozbawiają nas cennej informacji o barwie obiektów, co przy dalszych etapach obróbki (segmentacja) nie pozwala na korzystanie z kilku efektywnych metod.

Powszechność i dostępność kamer cyfrowych powoduje, iż bardzo łatwo zarejestrować potrzebny obraz. Możliwe są natomiast różne warianty takiej rejestracji. Zazwyczaj wykonuje się ujęcie twarzy od przodu, co jest naturalne również dla naszych przyzwyczajeń w kontaktach międzyludzkich. Okazuje się jednak, iż ujecie twarzy z profilu również niesie dużą ilość informacji i znajduje zastosowanie w analizie mimiki. Praca [17] przedstawia udaną próbę określania gestów twarzy na podstawie właśnie takiego obrazu (rysunek 2.2).



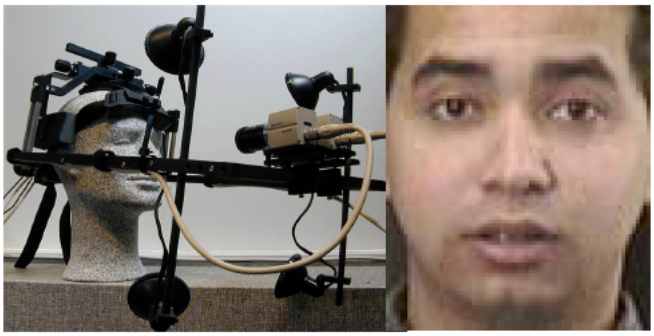
Rysunek 2.2: Analiza profilu twarzy (źródło: [17])

Do zagadnień akwizycji obrazu należy też zaliczyć rolę odpowiedniego oświetlenia. Zazwyczaj najlepsze efekty daje oświetlanie twarzy równomiernie, z różnych kierunków, rozproszonym światłem. Uniemożliwia to powstawanie cieni, mogących być mylnie interpretowanymi przez system, oraz zapewnia odpowiednie dostosowanie jasności sceny do czułości kamery. Rysunek 2.3 przedstawia efekty oświetlania twarzy pod różnymi kątami:

Z kolei rysunek 2.4 przedstawia specjalistyczne stanowisko do akwizycji obrazów twarzy, wykorzystywane przez autorów pracy [6]. Oprócz odpowiedniego oświetlenia zapewniono właściwe i powtarzalne wykadrowanie twarzy w scenie poprzez montaż całego oprzyrządowania na głowie użytkownika.

W warunkach laboratoryjnych niekiedy stosuje się kamery termowizyjne, operujące w paśmie promieniowania podczerwonego. Umożliwia to uzyskanie dodatkowej informacji o obiektach (temperatura), co w przypadku rejestracji i analizy

Rysunek 2.3: Wpływ oświetlenia na obraz twarzy (źródło: [16])



Rysunek 2.4: Stanowisko do akwizycji obrazów twarzy (źródło: [20])

twarzy dostarcza kolejnych istotnych informacji. Sprzęt taki jest jednak trudno dostępny, zwłaszcza w warunkach domowo-biurowych.

Wstępne przetwarzanie

Etap ten polega na wykonaniu takich operacji na obrazie, które pozwolą na łatwe uzyskanie potrzebnych danych w dalszym etapie przetwarzania. W przypadku analizy gestów mimicznych należy przede wszystkim zmniejszyć ilość przetwarzanej informacji wizyjnej (rozmiar obrazu), wyznaczając w obrazie tzw. ROI (*Region Of Interest*), czyli obszar naszego zainteresowania i dalszej analizy. Oczywiście będzie to twarz użytkownika, która niesie wszystkie potrzebne nam informacje. Pozostałe zarejestrowane elementy sceny (tło, tors, ręce) mogą być pomijane w dalszej obróbce.

Metod pozwalających na określenie obszaru twarzy jest wiele. Każda z nich może sprawdzać się w innych warunkach, zatem celem twórcy systemu jest dobranie takiej, która będzie optymalna dla obranych założeń (zarówno pod względem jakości działania jak i złożoności obliczeniowej). Do metod takich można zaliczyć:

- Metody oparte na segmentacji obrazu pod kątem wyszukania obszarów o kolorze skóry [15], [16], [20].
- Wyszukiwanie na obrazie obszaru najlepiej skorelowanego z deformowalnym szablonem (*template*) odpowiadającym twarzy ludzkiej [13], [25].

- Metody probabilistyczne, wykorzystujące zestaw detektorów, uprzednio wytrenowanych pod kątem wykrywania obszaru twarzy. (np. często stosowany detektor Viola-Jones) [28].

- Analiza obrazu gradientowego (przedstawiającego krawędzie) - wyszukiwanie owalu twarzy [16], [24].

- Analiza ruchu w celu oddzielenia obiektu od tła (*background subtraction*) - zastosowanie tylko przy nieruchomej kamerze i w miarę niezmiennym tle [23], [24]

- Metody przepływu optycznego (*Optical flow*) i analiza pól wektorowych ruchu [24], [18].

- Metody korzystające z aktywnych modeli kształtu (*Active Shape Model*) - dopasowują zadany kształt twarzy pozyskany ze zbioru uczącego [13], [25].

- Dopasowywanie elipsy do obiektów na obrazie (zgrubna analiza kształtu twarzy) [16], [24]

- Uczenie sieci neuronowych.

Najczęściej lokalizacja twarzy prowadzona jest za pomocą połączonych kilku z wyżej wymienionych metod. Pozwala to na uzyskanie dokładniejszych wyników i zwiększa uniwersalność systemu. Takie hybrydowe podejście sprawdza się także w późniejszych etapach analizy, opisanych w dalszej części pracy.

Następnie tak wykadrowany obraz jest poddawany przetwarzaniu, którego przebieg jest określony przez obraną metodę działania analizatora. Mogą to być (przykładowo) operacje:

- zmiany przestrzeni barwnej (np. z RGB do HSV) [20],

- wyostrzanie obrazu [24],

- wyznaczania krawędzi (filtr Sobela, Canny'ego) [24],

- binaryzacja [16],

- zmiana rozdzielczości obrazu (niekiedy utworzenie tzw. piramidy rozdzielczości. (*Multi Resolution Pyramid*), czyli zestawu kopii obrazu w różnej skali) [19].

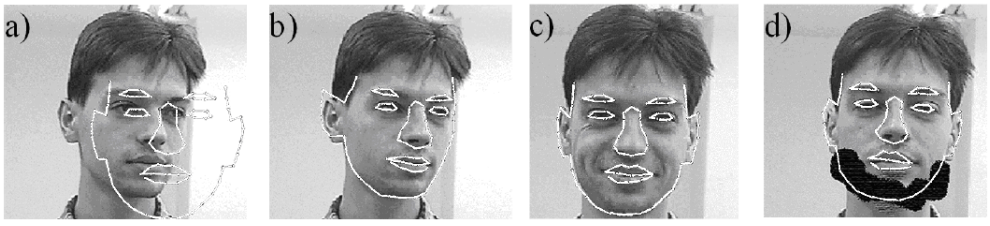
Ekstrakcja parametrów

Kolejnym etapem jest pozyskanie właściwej informacji o cechach analizowanej twarzy. Zależy nam głównie na zredukowaniu ilości danych niesionych przez obraz i wybraniu takich cech, które przydadzą się w analizie mimiki.

Istotne jest wybranie takich parametrów, które pozwolą na łatwe i jednoznaczne sklasyfikowanie badanego obrazu jako twarzy wyrażającej pewien stan emocjonalny. Dobór tych cech jest wręcz kluczowy dla sprawności działania klasyfikatora. Niektóre ze stosowanych podejść to:

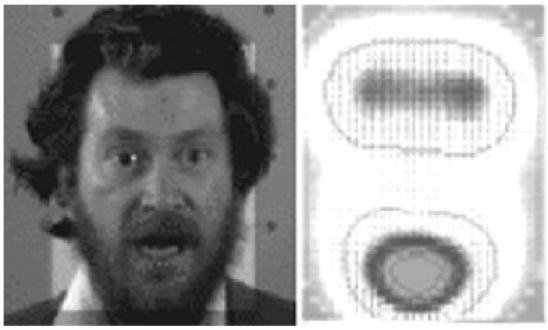
- Wyznaczenie wzajemnego położenia pewnych węzłowych punktów twarzy, tak jak ilustruje to rysunek 2.1.
- Analiza wybranych fragmentów twarzy pod kątem ich korelacji z zadanymi *a priori* wzorcami (np. oka) charakterystycznymi dla określonych AU.
- Korzystanie z aktywnego modelu kształtu lub wyglądu [13] (metoda holistyczna, obejmuje całą twarz, a nie tylko jej fragmenty) - przykład działania na rys. 2.5.
- Modelowanie 3D twarzy na podstawie obrazu 2D a następnie analiza kształtu takiego modelu.

Ponadto, każde z tych podejść może być stosowane zarówno do statycznego obrazu, jak i ich sekwencji. W tym drugim przypadku uzyskiwana jest dodatkowa informacja o ruchach poszczególnych części twarzy. Znajduje to zastosowanie w systemach automatycznie rejestrujących emocje użytkownika, gdyż dostarcza danych o czasowej dynamice obserwowanych gestów. Niekiedy do określenia danego gestu wystarczy obraz przedstawiający ruch na danych obszarach (oraz porównanie tego obrazu z wyznaczonym statystycznie szablonem). Przykład przedstawia rysunek 2.6



Rysunek 2.5: Przykład zastosowania ASM do opisu twarzy (Źródło:[25])

Znów, podobnie jak przy lokalizacji twarzy, można stosować jednocześnie kilka metod. Często właściwości wyszukiwanych obiektów są na tyle różne, że warto zastanowić się, jaka metoda będzie najlepsza właśnie dla tego jednego elementu (w konkretnym zastosowaniu). Przykładowo, autorzy pracy [21] do lokalizacji



Rysunek 2.6: Szablon obrazu ruchu dla gestu zaskoczenia (Źródło:[18])

ust zastosowali segmentację w przestrzeni HSV, natomiast dokładny model ich kształtu uzyskano, dopasowując doń zestaw krzywych (parabol).

Osobnym zagadnieniem jest dobór metod analizy obrazu, umożliwiających zastosowanie powyższych ścieżek postępowania. Jest to dziedzina, w której twórcy systemów mogą wykazać najwięcej inwencji, wciąż proponując nowe rozwiązania i doskonaląc istniejące. W przypadku systemów działających w czasie rzeczywistym bardzo istotna jest tu złożoność obliczeniowa. Gdy jest zbyt duża, z góry przekreśla szanse na stosowanie niekiedy bardzo dokładnych, lecz kosztownych obliczeniowo metod.

Klasyfikacja

Gdy wyznaczony jest już wektor cech obserwowanej w danym momencie twarzy, można przystąpić do jej klasyfikacji ze względu na wyrażane emocje. W tym miejscu istotne jest określenie cech, jakimi może charakteryzować się dany gest oraz jakie własności najlepiej go opisują. Na tej podstawie można wyznaczyć wagi poszczególnych cech w opisie stanu emocjonalnego wyrażanego mimicznie. Pomocne są w tym badania prowadzone przez psychologów i antropologów, jak chociażby wspomniane już prace P. Ekmana [14].

Punktem wyjścia może być założenie o istnieniu pewnego skończonego zbioru podstawowych gestów. Wówczas system stara się zakwalifikować aktualnie przetwarzany obraz twarzy do jednego z nich (bądź też w przypadku otwartego zbioru wyników stwierdzić niemożność zakwalifikowania go do żadnej grupy). Innym podejściem jest opracowanie tylko pewnych wzorców gestów i badanie prawdopodobieństwa, z jakim dana twarz może przedstawiać emocje z każdego wzorca. Jest to podejście bardziej otwarte, pozostawiające większą swobodę w interpretacji uzyskanych wyników.

W obydwu przedstawionych przypadkach konieczne jest posiadanie bazy wiedzy o gestach, które mają zostać rozpoznane. Konieczne jest zatem empiryczne wyznaczanie zakresów zmienności poszczególnych parametrów dla każdej grupy na podstawie zarejestrowanych uprzednio sekwencji uczących. Jakość klasyfikacji

w dużej mierze zależy od postaci takiej bazy, ilości osób w niej zarejestrowanych czy warunków akwizycji.

System regułowy, taki jak zaproponowany w [20] określa wypracowane dzięki długim obserwacjom zależności pomiędzy wzajemnym położeniem punktów twarzy a danym AU. Przykładowo (oznaczenia punktów z rysunku 2.1):

AU5 - otwarcie szeroko oczu :
$\Delta\overline{FG} > 0$

AU6 - uśmiech :
$AU12 \vee AU13$

AU12 - uniesienie kącików ust :
$[(\Delta\overline{IB} > 0) \wedge (\Delta\overline{CI} < 0)] \vee [(\Delta\overline{JB1} > 0) \wedge \ (\Delta\overline{CJ} < 0)]$

AU13 - uniesione policzki :
$[(\Delta\overline{IB} > 0) \wedge (\Delta\overline{CI} > 0)] \vee [(\Delta\overline{JB1} > 0) \wedge \ (\Delta\overline{CJ} > 0)]$

Dużym udoskonaleniem systemu może być możliwość jego adaptacji do konkretnych warunków pracy, co w szczególności oznacza konkretną osobę poddawaną badaniu. Pomimo tego, iż gesty mimiczne (podstawowe) można traktować jako kategorie uniwersalne, to wiadomym jest, że różni ludzie wyrażają je w inny sposób. To, co u jednej osoby będzie lekkim rozszerzeniem powiek, u innego może oznaczać wielkie zaskoczenie. Pożądane jest, aby system mógł klasyfikować obrazy twarzy, mając na względzie zakres zmienności parametrów dla danej osoby.

Na koniec należy jednakże uświadomić sobie jeden fakt: często nawet człowiek nie jest w stanie odczytać jednoznacznie emocji z twarzy rozmówcy. Badania wykazały, iż wyćwiczony w tym kierunku ludzki obserwator jest w stanie klasyfikować 6 podstawowych gestów mimicznych (str. 4) z ok. 87% trafnością [18]. Wynika to ze sporej odmienności w ekspresji tych samych gestów u różnych ludzi, w różnych warunkach. Trudno więc wymagać, aby automatyczny system uzyskał lepsze wyniki. Już samo zbliżenie się do tych rezultatów może być uznane za sukces klasyfikacji systemu sztucznej inteligencji.

2.2 Przykłady działających systemów

Omówione zostaną teraz pokrótce dwa istniejące rozwiązania problemu analizy gestów mimicznych. Metody te prezentują dwie zupełnie odmienne ścieżki analizy obrazu, wykorzystują inne zestawy informacji, jednakże prowadzą do tego samego celu. Pierwiastkiem wspólnym jest natomiast wykorzystanie wiedzy o charakterystycznym wyglądzie ludzkiej twarzy.

2.2.1 System hybrydowy ISFER[19]

System ISFER (*Integrated System for Facial Expression Recognition*), opisany w pracy [19], umożliwia odnajdywanie charakterystycznych punktów twarzy, badanie ich wzajemnego ułożenia, wychwytywanie obecności poszczególnych AU (tabela 2.1) oraz wnioskowania na ich podstawie o wyrażanych mimicznie emocjach.

Główną cechą wyróżniającą ten system jest zastosowanie wielu technik i algorytmów analizy obrazu, działających równolegle dla zwiększenia sprawności detekcji cech twarzy. Twórcy systemu wyszli ze słusznego założenia, iż pojedynczy detektor może działać poprawnie tylko w pewnych warunkach. Łącząc kilka technik, uzyskali większą pewność, iż dana cecha twarzy jest poprawnie zidentyfikowana.

Na wstępie należy wspomnieć również o przyjętym odnośnie akwizycji obrazów założeniu. System nie wykrywa twarzy na obrazie. Na wejście podawane są zdjęcia twarzy (jednocześnie *an face* i z profilu) umieszczone w ściśle określonym miejscu.

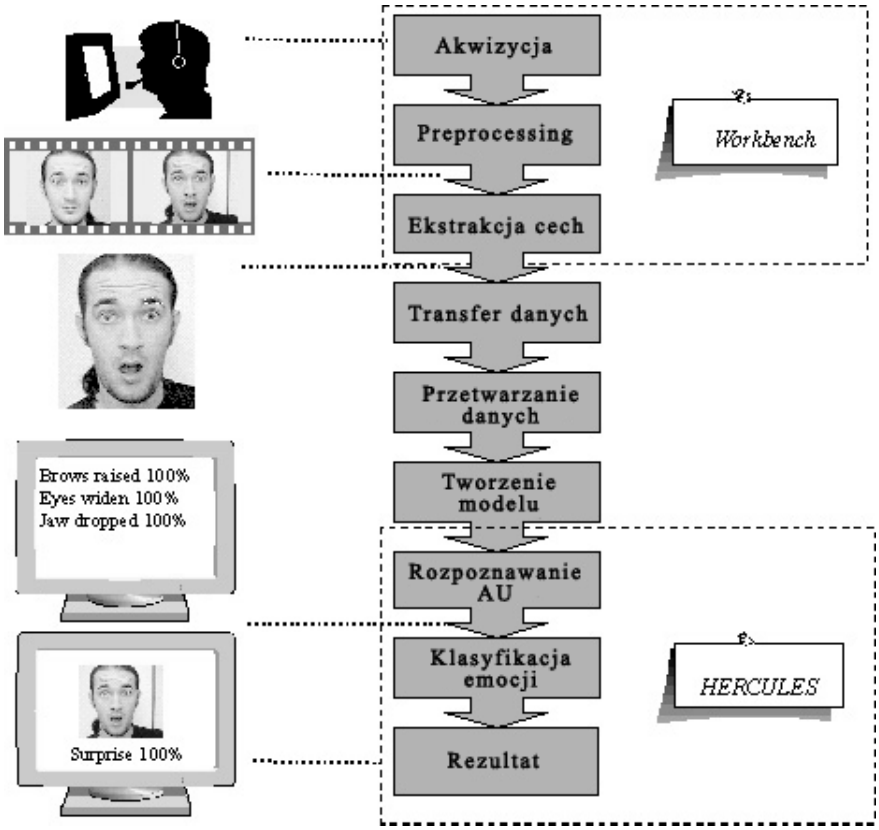
Architektura systemu przedstawiona jest na rysunku 2.7. Pierwsza część (*Workbench*) angażuje różne algorytmy przetwarzania obrazu i jego analizy w celu uzyskania (czasem redundantnych a czasem sprzecznych) danych o cechach twarzy (położeniu punktów charakterystycznych). Kolejny moduł dokonuje selekcji właściwych danych i przekazuje je dalej (do modułu*Hercules*), gdzie rozpoznawane są konkretne AU i emocje. System daje możliwość dowolnego wyboru detektorów w module *Workbench*, co umożliwiło kompleksowe zbadanie sprawności algorytmów (i ich kombinacji).

Punkty charakterystyczne są wyszukiwane w celu utworzeni dwóch modeli twarzy: frontalnego i bocznego (rys.2.8). Następnie analizowane są wzajemne położenia tych punktów, co daje informacje, jakie AU jest aktualnie prezentowane. Dodatkowo, badany jest kształt ust. Podsumowując:

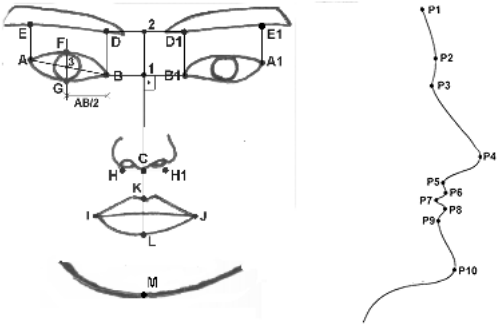
- Twarz opisywana jest 30 cechami.
- Cechy te są obliczane na podstawie położenia 29 punktów charakterystycznych (19 z przodu, 10 z profilu) i ułożenia warg.
- Zastosowany system regułowy pozwala na podstawie tych cech rozpoznać 29 różnych AU (z 44 opisanych w [14].)

Przedstawione teraz zostaną niektóre z algorytmów, użytych w systemie do ekstrakcji różnych cech twarzy.

a) Zgrubna lokalizacja części twarzy przy użyciu projekcji obrazu.



Rysunek 2.7: Przepływ danych w systemie ISFER



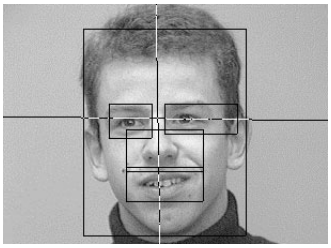
Rysunek 2.8: Modele twarzy i punkty charakterystyczne w systemie ISFER

W celu początkowego wyznaczenia regionów, gdzie znajdują się oczy, usta i nos, analizowane są gradienty projekcji obrazu: wertykalnej i horyzontalnej. Projekcję wertykalną otrzymuje się, sumując piksele w kolejnych rzędach obrazu. Efektem jest jednowymiarowy histogram. Badany jest jego gradient, czyli różnice pomiędzy sumami pikseli w kolejnych rzędach. Podobnie postępuje się przy projekcji horyzontalnej.

Wykorzystywane są informacje płynące z analizy projekcji:

- Maksima projekcji wertykalnej odpowiadają linii między włosami a czołem, oczom, dziurkom w nosie, ustom i granicy między podbródkiem a szyją.
- Linia oczu to drugie od góry maksimum tej projekcji.
- Współrzędna x pionowej linii biegnącej przez nos to minimum różnicy kontrastów pikseli leżących na poziomej linii oczu.
- Oczy, usta i nos są znajdowane przez dalszą analizę histogramów obszarów wyznaczonych przez znalezione uprzednio linie.

Należy zauważyć, iż metoda ta sprawdza się jedynie w przypadku twarzy ustawionych równo pionowo.



Rysunek 2.9: Efekt analizy projekcji obrazu

b) Detekcja oczu przy użyciu sieci neuronowych.

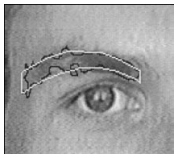
Do wyszukania punktów charakterystycznych oka wykorzystano sieć neuronową typu *Backpropagation*, zawierającą 81 neuronów w warstwie wejściowej, 4 w ukrytej i jeden wyjściowy. Sieć ta analizuje wycinki obrazu 9×9 pikseli i została nauczona do wykrywania źrenicy oka. Badane są kolejne wycinki obszaru oka, wyznaczonego przez wcześniej opisaną analizę histogramów.

Następnie ta sama sieć, inaczej nauczona, lokalizuje punkty węzłowe oka, zaznaczone na rysunku 2.10. Są to charakterystyczne miejsca, zawsze występujące w obszarze oka, niezależnie od bieżącej mimiki twarzy.

Rysunek 2.10: Punkty węzłowe oka

c) Wykrywanie kształtu brwi przy użyciu metody *curve fitting*.

Trójkątny region brwi jest lokalizowany na podstawie położenia punktów węzłowych oka. Obszar ten jest poddawany progowaniu i wybierany jest największy obiekt. Następnie jego kontur aproksymowany jest dwoma krzywymi drugiego stopnia.



Rysunek 2.11: Wykrywanie kształtu brwi

d) Określanie ekspresji ust.

Wewnątrz wstępnie wyznaczonego regionu, gdzie znajdują się usta, wyliczane jest pole wektorowe, reprezentujące lokalne gradienty intensywności obrazu. Następnie wektory te są uśredniane na 100 obszarach i przekazywane do sieci neuronowej. Badana jest oddzielnie lewa i prawa część obrazu ust, po czym wyniki są przekazywane do kolejnej sieci, której wyjścia odpowiadają stanom: neutralny, uśmiech, smutek.

Inne z zastosowanych algorytmów to między innymi metoda aktywnego konturu do określenia kształtu oczu i ust oraz segmentacja koloru skóry przydatna przy wyznaczaniu regionów zainteresowania.

Autorzy systemu przeprowadzili wyczerpujące testy, dowodzące wysokiej sprawności ich aplikacji. Rysunek 2.12 ukazuje wyniki testów na rozpoznawanie poszczególnych AU. Jak widać osiągnięto sprawność rzędu 70-90%, przy czym niektóre AU są rozpoznawane bezbłędnie (zob. tabela 2.1). Tabela 2.3 zawiera wyniki testów na określanie emocji. Uzyskana sprawność rzędu 90% może być uznana za doskonałą.

2.2.2 System wykorzystujący aktywne modele kształtu [13]

Praca [13] prezentuje działanie i zastosowanie algorytmów ASM (*Active Shape Model*) i AAM (*Active Apperance Model*) . Uwaga zostanie poświęcona pierwszemu z nich.

Zasadę działania aktywnych modeli kształtu można najprościej wyjaśnić jako dopasowanie do obrazu pewnego uniwersalnego konturu, uzyskanego wcześniej na podstawie danych statystycznych wyliczonych ze zbioru uczącego. Kontur taki (model kształtu) to właściwie zbiór punktów, leżących w pewnych charakterystycznych miejscach przedstawianego przedmiotu, przy czym określa się go

Tabela 2.2: Reguły wnioskowania zastosowane w systemie ISFER)

AU	Emocje	AU	Emocje	AU	Emocje
1+2	Zaskoczenie	1	Smutek	23+17	Złość
2	Złość	4	Złość	10+17	Niesmak
6	Radość	5	Zaskoczenie	23+26	Złość
1+4+5+7	Strach	7	Złość	10+(25/26)	Niesmak
1+4+5	Strach	24+17	Złość	23	Złość
1+4+7	Smutek	27	Zaskoczenie	10	Niesmak
1+5+7	Strach	20+(25/26)	Strach	24+17+26	Złość
1+4	Smutek	20	Strach	9+(25/26)	Niesmak
1+5	Strach	15+(25/26)	Smutek	9+17	Niesmak
1+7	Smutek	15	Smutek	24+26	Złość
5+7	Strach	23+17+26	Złość	9	Niesmak
24	Złość	12+(25/26)	Radość	10+16+(25/26)	Złość
12	Radość	10+17+(25/26)	Niesmak	16+(25/26)	Złość
9+17+(25/26)	Niesmak	17	Smutek	12+16+(25/26)	Radość
26	Zaskoczenie				

Tabela 2.3: Wyniki rozpoznawania emocji przez system ISFER (ilość poprawnych detekcji w procentach)

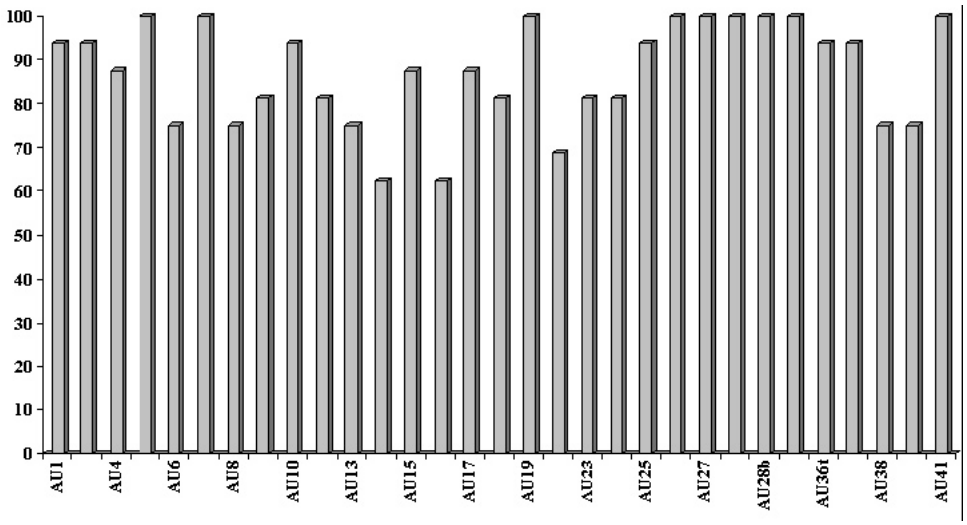
Emocje	Rozpoznane emocje						
	Zaskoczenie	Strach	Niesmak	Złość	Radość	Smutek	Mieszane
Zaskoczenie	**97**	1	0	0	0	0	2
Strach	0	**84**	0	0	0	9	7
Niesmak	0	0	**82**	14	0	0	3
Złość	0	1	12	**84**	0	0	2
Radość	1	0	0	0	**98**	0	1
Smutek	0	2	0	0	0	**96**	2
Mieszane	3	1	0	0	2	1	**93**
Średnio	**90.57**						

mianem aktywnego, gdyż uwzględnia on zmienność kształtu (położenia punktów) danego przedmiotu na różnych obrazach ze zbioru testowego. Rysunek 2.13 przedstawia taki model, umieszczony na obrazie, który opisuje.

W opisywanym systemie wspomniany wcześniej zbiór uczący to zestaw zdjęć różnych twarzy, zróżnicowany pod względem wyrażanych emocji (deformacje podstawowego kształtu). Im większy zbiór i większe jego zróżnicowanie, tym lepszy model kształtu.

Na każdym obrazie uczącym zostają oznaczone (zazwyczaj ręcznie) punkty węzłowe - charakterystyczne miejsca, takie jak kąciki oczu i ust oraz punkty na owalu twarzy (Rys. 2.13). Autorzy wykorzystali około 300 zdjęć, na każdym oznaczono 133 punkty. Niektóre z konturów przedstawia rysunek 2.14. Widać na nim, jak poszczególne kontury różnią się od siebie, jednak są to zmiany ograniczone do pewnego zakresu, wyznaczonego wariancją zbioru uczącego.

W celu uzyskania jednego, elastycznego konturu (czyli modelu) wszystkie kon-



Rysunek 2.12: Jakość rozpoznawania AU przez system ISFER

Rysunek 2.13: Obraz uczący z naniesionymi punktami węzłowymi

tury uczące są normalizowane. Usuwana jest informacja o skali, obrocie i przesunięciu, co pozwala rozpatrywać wszystkie kontury w tym samym układzie współrzędnych.

Mając dane s konturów uczących, każdy opisany zestawem n punktów w 2-wymiarowej przestrzeni, poszczególne kształty można opisać $2n$ elementowymi wektorami:

$$x = (x_1, \ldots, x_n, y_1, \ldots, y_n) \tag{2.1}$$

Zbiór tych wektorów tworzy pewien rozkład w $2n$-wymiarowej przestrzeni. Analizując ten rozkład, można wyznaczyć cechy elastycznego modelu kształtu.

Na wstępie dąży się do zmniejszenia wymiarowości przestrzeni zmienności danych z $2n$ do dogodniejszej wielkości. W tym celu stosuje się tzw. analizę głównych składowych (PCA, *Principal Component Analysis*). W uproszczeniu polega ona na wyznaczeniu w chmurze punktów głównych osi zmienności.

Jeżeli za $\bar{x}$ przyjąć środek takiej chmury (co odpowiada uśrednionemu kształ-

Rysunek 2.14: Przykłady kształtów twarzy ze zbioru uczącego

towi), wówczas każdy z wektorów x można aproksymować jako

$$x \approx \bar{x} + Pb \tag{2.2}$$

gdzie $P = (p_1|p_2|\ldots|p_t)$ zawiera t wektorów, będącymi osiami nowej, zredukowanej do t wymiarów przestrzeni zmienności danych, natomiast b jest t-elementowym wektorem, danym

$$b = P^T(x - \bar{x}) \tag{2.3}$$

(Wyliczanie macierzy P opisano dokładnie w [13] oraz [25]).

Dzięki temu otrzymuje się wektor b, będący zestawem t parametrów, opisujących deformację modelu, przy czym każdy jego element odpowiada zmianie jakiejś cechy całego kształtu, a nie tylko konkretnego punktu (na przykład otwarcie ust, uniesienie brwi, itp.).

Zmieniając parametry b_i można generować nowe kształty, korzystając z wzoru (2.2). Nakłada się przy tym ograniczenie na zmienność elementów b_i wektora b, aby uniknąć nienaturalnie zdeformowanych kształtów. Przyjmuje się zazwyczaj zmienność w zakresie $\pm 3\sqrt{\lambda_i}$, gdzie λ_i jest wariancją b_i w zbiorze uczącym. Rysunek 2.15 ukazuje skutki modyfikacji pewnych 3 parametrów.

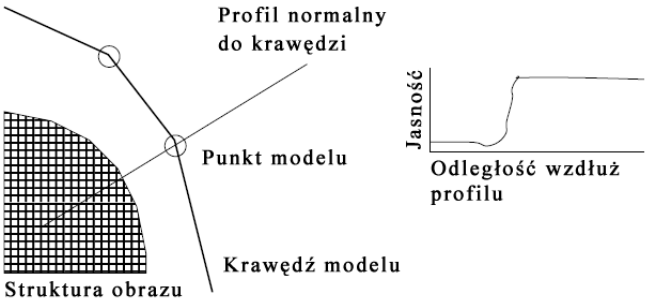
Mając tak skonstruowany model, można próbować dopasowywać go do analizowanego obrazu. Po dopasowaniu wystarczy odczytać jego parametry, zawarte w wektorze b, aby uzyskać informacje o danej twarzy (na przykład o wyrażanych emocjach).

Dopasowanie modelu do obrazu opiera się znów na analizie statystycznej. Tym razem rozpatrywany jest rozkład jasności obrazu w otoczeniu punktów węzłowych.

Początkowo umieszcza się podstawowy (uśredniony) model w pewnym miejscu analizowanego obrazu. Najlepiej gdy jest to mniej więcej jego środek. Następnie dla każdego punktu węzłowego analizowany jest profil jasności obrazu wzdłuż

Rysunek 2.15: Modyfikacja parametrów modelu kształtu. Poszczególne wiersze przedstawiają efekt zmiany wartości pojedynczych parametrów, podczas gdy pozostałe wartości są wyzerowane

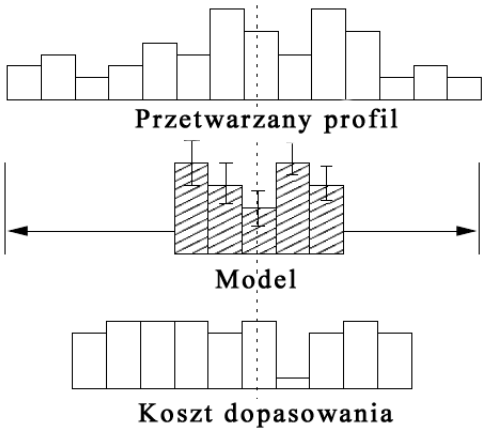
linii normalnej do krawędzi modelu. Na tym profilu wyszukiwane jest domniemane właściwe położenie danego punktu. W najprostszym przypadku szukana jest najbardziej wyraźna krawędź. Obrazuje to rysunek 2.16.



Rysunek 2.16: Analiza profilu normalnego do krawędzi modelu

Dla rzeczywistych obrazów lepiej jednak analizować otoczenie każdego punktu we wszystkich obrazach uczących i na tej podstawie budować statystyczny model oczekiwanego otoczenia danego punktu. Bada się przy tym raczej gradient obrazu niż jego bezwzględne wartości. Rysunek 2.17 przedstawia, jak analizowany jest profil pod kątem dopasowania do modelu. Widać też punkt, którego otoczenie najbardziej przypomina to modelowe. Dokładna metoda analizy otoczenia punktu opisana jest w [13].

Po znalezieniu wszystkich nowych domniemanych punktów węzłowych następuje próba dopasowania do nich modelu kształtu. Model kształtu umieszczony



Rysunek 2.17: Dopasowanie otoczenia punktu węzłowego do obrazu

na obrazie może być w ogólności opisany jako:

$$X = T_{X_t,Y_t,s,\theta}(\bar{x} + Pb) \tag{2.4}$$

gdzie funkcja $T_{X_t,Y_t,s,\theta}$ odpowiada za translację, rotację i skalowanie znormalizowanego modelu.

Dopasowanie modelu (wyznaczenie X_t, Y_t, s, θ, b) do nowego zestawu punktów, Y odbywa się poprzez minimalizację sumy kwadratów odległości:

$$|Y - T_{X_t,Y_t,s,\theta}(\bar{x} + Pb)|^2 \tag{2.5}$$

Dopasowywanie powtarzane jest iteracyjnie, aż do pewnej ustalonej dokładności. Rysunek 2.18 prezentuje działanie algorytmu.

W procesie dopasowywania ważne jest początkowe umiejscowienie modelu. Na rysunku 2.19 pokazano błędne dopasowanie modelu.

Pomimo wysokiej sprawności działania systemu opartego o ASM, nie należy zapominać o pewnym fakcie. Jego działanie i trafność analizy twarzy w dużym stopniu zależy od zawartości zbioru uczącego.

Jeżeli, przykładowo, znajdzie się w nim stosunkowo mało zdjęć, przedstawiających smutek na twarzy, wówczas odpowiadającej takiemu gestowi deformacji kształtu twarzy nie zostanie przyporządkowana żadna oś zmienności (wektor p_i w równaniu 2.2). W związku z tym, obrazy na których znajdzie się smutna twarz nie będą cechować się żadną specyficzną zawartością wektora b, w związku z czym system nie będzie w stanie stwierdzić, kiedy analizowana twarz jest smutna.

Rysunek 2.18: Poprawne dopasowanie modelu kształtu

Rysunek 2.19: Niepoprawne dopasowanie modelu kształtu

Rozdział 3

Opis systemu

3.1 Użyte oznaczenia

W dalszej części pracy pojawiają się wzory, które odnoszą się do operacji na obrazach. W celu ułatwienia ich interpretacji podano poniżej wyjaśnienie znaczenia użytych oznaczeń i operatorów:

I – Obraz, ramka sekwencji wideo.

I_i – Obraz o danym indeksie (np. kolejny w dziedzinie czasu).

I_R – Składowa R (czerwona) obrazu barwnego.

Analogicznie dla kanału zielonego (G) oraz niebieskiego (B).

$I_i + I_j$ – Suma obrazów; każdy jego piksel to suma wartości odpowiednich pikseli dwóch podanych obrazów (podobnie dla operacji $-, *, /$).

Zakłada się, iż jest to obraz o jednym kanale (a więc nie kolorowy) .

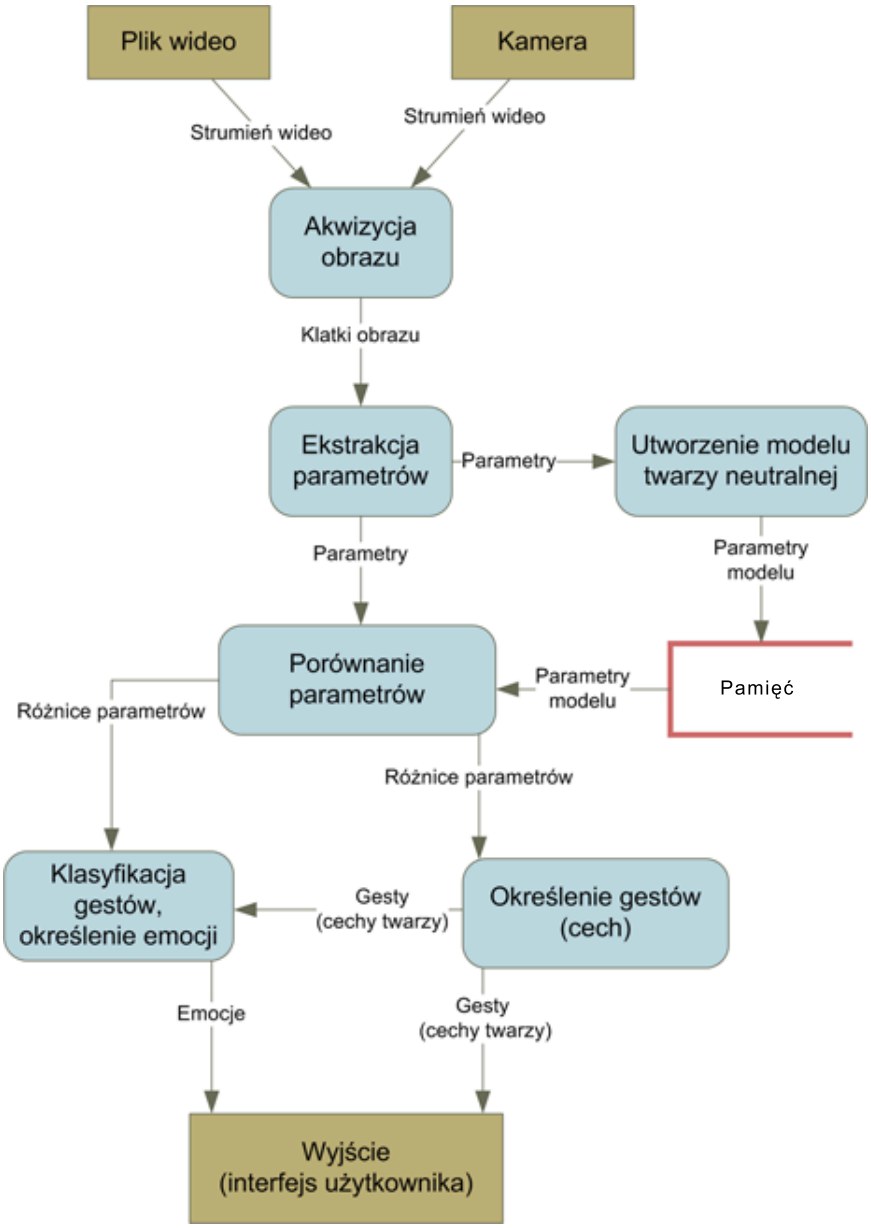
$I(i, j)$ – Wartość piksela o współrzędnych x, y.

$I(i, j, R)$ – Wartość piksela kanału R (czerwonego) o współrzędnych x, y.

Analogicznie dla kanału zielonego (G) oraz niebieskiego (B).

3.2 Koncepcja systemu

Zbudowany system automatycznego wykrywania i klasyfikacji gestów mimicznych składa się z szeregu komponentów. Jego konstrukcję najlepiej przedstawia diagram DFD (przepływu danych) - rys. 3.1. Na schemacie zaznaczono poszczególnymi etapy przetwarzania danych oraz ich przepływ systemie.



Rysunek 3.1: Schemat przepływu danych w systemie

Użytkownik, siedząc na wprost kamery podłączonej do komputera, uruchamia program, który odczytuje obraz z urządzenia, przetwarza go przy pomocy opisanych dalej algorytmów, a następnie zwraca wynik w postaci informacji o rozpoznanych gestach i emocjach. Prezentowana jest także wizualizacja wyekstrahowanych cech obrazu, nałożonych na obraz z kamery, co ma na celu kontrolę poprawności działania algorytmów.

Alternatywnie, zamiast pobierać dane z kamery, można załadować je z uprzednio nagranego pliku wideo.

Kamera dostarcza danych w postaci strumienia wideo o szybkości ok. 15-30 ramek/sek. Taka duża rozdzielczość czasowa nie jest konieczna przy omawianym zastosowaniu. Dlatego też system analizuje pojedyncze klatki obrazu, pobierane 4 razy na sekundę. Wystarcza to w zupełności do wychwycenia wszelkich gestów wyrażanych za pomocą twarzy a zarazem pozwala systemowi działać w czasie rzeczywistym. Przy większym *framerate* jakość rozpoznawania nie uległaby znacznej poprawie, poważnie zwiększyłoby się za to zapotrzebowanie na moc obliczeniową.

Rejestrowany jest obraz barwny. Domyślna rozdzielczość to 320x240 pikseli. Większość nowoczesnych kamer umożliwia akwizycję obrazów o dużo większej rozdzielczości, jednak nakład obliczeniowy wielu algorytmów systemu rośnie wraz z kwadratem rozdzielczości obrazów. Jednocześnie doświadczenia pozwoliły stwierdzić, że przy omawianej rozdzielczości i dobrych warunkach oświetleniowych system uzyskuje wystarczająco dobre rezultaty. W przypadku załadowania z dysku pliku wideo o większej rozdzielczości, zostanie on przeskalowany do rozmiarów 320x240 pikseli.

Operacja "Ekstrakcja parametrów" ze schematu 3.1 zawiera szereg operacji kluczowych dla działania systemu. Każdy etap ekstrakcji parametrów wykorzystuje pewien wyspecjalizowany algorytm. Poszczególne kroki opisane są w dalszych rozdziałach.

Na wyjściu system prezentuje zarówno wykryte gesty twarzy (np. uniesienie brwi, kącików ust itp.) jak i rozpoznane emocje, które wyraża taka twarz.

3.3 Lokalizacja twarzy

Zastosowana metoda lokalizacji twarzy bazuje na przyjętych założeniach odnośnie warunków stosowania systemu:

- Kamera jest nieruchoma.
- Kolejne klatki obrazu pobierane są z określoną częstotliwością.
- Przed kamerą (stosunkowo blisko) siedzi użytkownik.
- Za użytkownikiem znajduje się niezmienne tło (nie przechodzą tamtędy inne osoby itp.).

Mając na uwadze takie warunki akwizycji, można wykorzystać pewne informacje, płynące z analizy ruchu w sekwencji wideo. W szczególności to, iż obszar ruchu na obrazie odpowiada ruchom użytkownika (jego głowy oraz górnej części tułowia) Ponadto sylwetka taka jest na tyle charakterystyczna, iż można wyznaczyć na niej rejon zajmowany przez głowę. Podejście takie zaproponował dr Adrian Horzyk - promotor pracy. Opisane jest ono także w [24] pod nazwą "metoda hierarchiczna".

Aby przyspieszyć działanie algorytmu, klatki obrazu są na wstępie skalowane do połowy swej pierwotnej wielkości. Wydatnie zmniejsza to złożoność obliczeniową. Nie wpływa natomiast znacząco na jakość wykrywania ruchomych obiektów, gdyż analiza ruchu na tym etapie i tak jest stosunkowo zgrubna.

3.3.1 Różnicowy obraz ruchu

Wstępnym etapem algorytmu lokalizacji twarzy jest uzyskanie z sekwencji obrazów informacji o ruchu obiektów, znajdujących się przed kamerą.

Pierwszym, nasuwającym się na myśl rozwiązaniem jest wyliczenie modułu różnicy między bieżącą klatką I_i a obrazem samego tła B (tu, jak i później operacje wykonywane są na obrazach w skali szarości):

$$I_{diff} = |I_i - B| \tag{3.1}$$

Problemem jest jednak uzyskanie tego ostatniego. Zamiast niego można zastosować na przykład poprzednią klatkę:

$$I_{diff} = |I_i - I_{i-1}| \tag{3.2}$$

Uzyskany wówczas obraz przedstawia zmiany, które zaszły pomiędzy tymi dwoma klatkami, a więc i ruch obiektów.

Takie rozwiązanie niesie ze sobą jednak pewną niedogodność. Otóż taka metoda jest bardzo wrażliwa na drobne zmiany w obrazie, wynikłe chociażby z zakłóceń w przetworniku CCD kamery. Konieczne staje się otrzymanie uśrednionego obrazu tła z większej ilości (n) poprzednich klatek.

$$B_i = \frac{\sum_{j=1}^{n} I_{i-j}}{n} \tag{3.3}$$

Aby uniknąć jednak konieczności przechowywania w pamięci wielu obrazów i każdorazowego ich uśredniania, można zastosować metodę, opisaną w [23]. Jej ideę wyraża wzór (3.4).

$$\begin{aligned} I_{diff} &= |I_i - B_i| \\ B_{i+1} &= \alpha * I_i + (1-\alpha) * B_i, \quad \alpha \in (0;1) \end{aligned} \tag{3.4}$$

Obraz tła jest każdorazowo odświeżany i stanowi sumę ważoną bieżącej klatki i poprzedniego obrazu tła. Dzięki temu obraz tła wyliczany jest z większej ilości klatek, lecz do wyliczeń tych nie potrzeba przechowywać całej historii obrazu. Wartość parametru α ustalono doświadczalnie i wynosi on $0,85$.

W celu pozbycia się drobnych zakłóceń obrazu zastosowano filtrację medianową zarówno dla bieżącej klatki, jak i dla obrazu różnicowego.

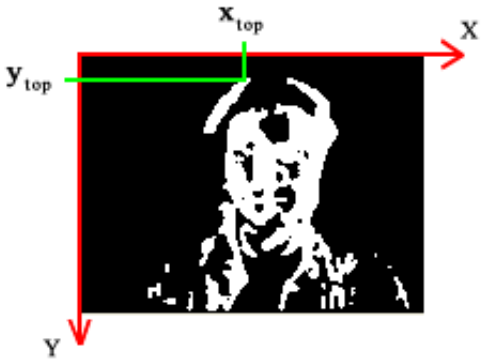
Tak otrzymany obraz jest poddawany binaryzacji, co pozwala na wyróżnienie obszarów ruchomych i tła. Uzyskany obraz różnicowy jest dość ciemny, tak więc wartość progu binaryzacji musi być mała. Wyznaczono ją doświadczalnie na poziomie = 3 (w 256 stopniowej skali szarości).

Rysunek 3.2: Klatka obrazu wejściowego, obraz różnicowy ruchu oraz ten sam obraz po binaryzacji.

3.3.2 Analiza obrazu ruchu

Mając do dyspozycji binarny obraz ruchu użytkownika, można wyznaczyć na nim obszar głowy (a co za tym idzie - twarzy).

Na początku określane jest położenie czubka głowy: jest to pierwszy od góry niezerowy punkt na obrazie ruchu. Jak widać na rys. 3.3 punkt ten nie zawsze może być wyznaczony poprawnie, lecz wystarczy jego dobre przybliżenie. Współrzędne tego punktu oznaczono jako (x_{top}, y_{top}).

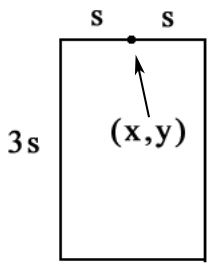


Rysunek 3.3: Wyznaczanie czubka głowy

Następnie szukany jest obszar o określonym kształcie, którego górna granica odpowiada y_{top}, natomiast współrzędna x środka wynosi mniej więcej x_{top}. Najkorzystniejszym kształtem byłaby tu elipsa, która dobrze przybliża zarys głowy, jednakże ze względu na ilość obliczeń konieczną przy wyznaczaniu takiego kształtu zastąpiono ją prostokątem o stosunku szerokości do wysokości $= \frac{2}{3}$.

Wyznaczenie obszaru głowy przebiega iteracyjnie. Sprawdzane jest dopasowanie prostokątów w różnej skali i o różnym położeniu.

Niech w, h oznaczają odpowiednio szerokość i wysokość całego obrazu. Jeśli przez (x, y) oznaczy się położenie środka górnej krawędzi takiego prostokąta, a przez s - połowę długości tej krawędzi (rys. 3.4), wówczas po kolei konstruowane są prostokąty, dla których:



Rysunek 3.4: Prostokątny kształt dopasowywany do zarysu głowy

$$\begin{array}{rcl} y & = & y_{top} \\ x & \in & \langle x_{top} - \frac{w}{10}, x_{top} + \frac{w}{10} \rangle \\ s & \in & \langle \frac{h}{8}, \frac{h}{3} \rangle \end{array} \tag{3.5}$$

Dla każdego wyznaczonego w ten sposób obszaru liczona jest wartość funkcji dopasowania, dana wzorem (3.6).

$$fit(I, x, y, s) = \frac{sum(I, x, y, s)}{s^p} \tag{3.6}$$

,gdzie

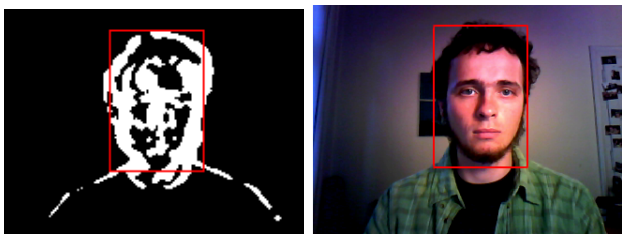
$$sum(I, x, y, s) = \sum_{i=x-s}^{x+s} \sum_{j=y}^{y+3s} I(i, j) \tag{3.7}$$

Funkcję dopasowania 3.6 można interpretować jako miarę wypełnienia wskazanego obszaru.

Zależność 3.7 opisuje sumę pikseli obrazu w aktualnie rozpatrywanym obszarze prostokątnym. Zastosowany w systemie sposób obliczania tej sumy przedstawiono na stronie 32.

Za obszar zajmowany przez głowę przyjmuje się ten, dla którego otrzymano największą wartość funkcji dopasowania. Faworyzuje ona przy tym obszary większe (docelowo - największy, stosunkowo dobrze wypełniony obszar). Za tę właściwość funkcji odpowiada wartość wykładnika p we wzorze (3.6), którą doświadczalnie ustalono na $1,2$.

Przedstawiona powyżej metoda działa poprawnie, lecz jedynie w przypadku, kiedy użytkownik wykonuje jakieś ruchy głową. W momencie, gdy znieruchomieje, z obrazu różnicowego znika charakterystyczna sylwetka (popiersie). W skrajnym przypadku obraz taki nie zawiera żadnego obiektu. Nie można wówczas określić położenia twarzy.



Rysunek 3.5: Efekt działania algorytmu wyszukiwania twarzy

Jednakże można przyjąć, iż brak ruchu, reprezentowanego na obrazie różnicowym (lub ruch na bardzo małym obszarze) oznacza, iż twarz dalej znajduje się w tym samym miejscu, które zostało zlokalizowane jako ostatnie.

Dlatego też zastosowano usprawnienie algorytmu, polegające na dodatkowym sprawdzaniu wartości funkcji dopasowania dla nowo wyznaczonego obszaru:

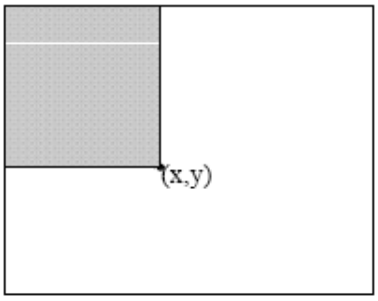
$$\begin{array}{lcl} fit(I,x,y,s) \leqslant 20 & \Rightarrow & \text{pozostaw obszar z poprzedniej ramki} \\ fit(I,x,y,s) > 20 & \Rightarrow & \text{użyj nowego obszaru} \end{array} \tag{3.8}$$

Omówiona zostanie teraz zastosowana metoda liczenia sum pikseli w obrazie.

Ponieważ dla każdej klatki należy policzyć kilkadziesiąt razy takie sumy na różnych obszarach, operacja ta prowadzona w tradycyjny sposób nie jest optymalna obliczeniowo. Można jednak usprawnić ten proces, wykorzystując tzw. obraz całkowy (*integral image*), opisany w [28].

Obraz taki ma wymiary obrazu wejściowego. Buduje się go, przypisując każdemu z jego pikseli sumę pikseli obrazu wejściowego, znajdujących się ponad i z lewej strony tego punktu (rys 3.6).

$$C(x,y) = \sum_{x' \leqslant x, y' \leqslant y} I(x',y') \tag{3.9}$$

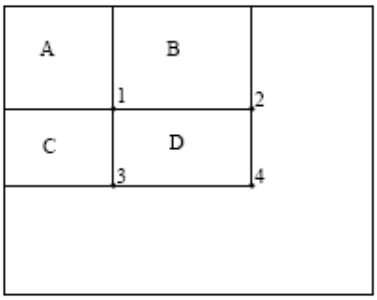


Rysunek 3.6: Tworzenie obrazu całkowego (źródło [28])

Obraz całkowy można wyliczyć rekurencyjnie:

$$\begin{aligned} C(x,y) &= C(x-1,y) + s(x,y) \\ s(x,y) &= s(x,y-1) + I(x,y) \end{aligned} \tag{3.10}$$

Mając tak skonstruowany obraz całkowy, wyliczenie sumy pikseli w interesującym nas obszarze sprowadza się do wykonania kilku operacji dodawania i odejmowania, co ilustruje rysunek 3.7.



Rysunek 3.7: Wyliczanie sumy pikseli. Wartość obrazu całkowego w punkcie 1 to suma pikseli w obszarze A. Wartość w punkcie 2 to $A+B$, w punkcie 3 to $A+C$, a w punkcie 4: $A+B+C+D$. Stąd suma pikseli w D może być obliczona jako $4+1-(2+3)$ (źródło [28])

Metoda ta sprawdza się doskonale w takich zastosowaniach, jak opisywany algorytm lokalizacji twarzy. Raz wyliczony dla danej klatki obraz całkowy pozwala na obliczanie sum prostokątnych obszarów z jednakową, bardzo małą złożonością obliczeniową.

3.4 Wyszukiwanie regionów oczu i ust

Mając wyznaczony (mniej lub bardziej dokładnie) obszar twarzy, można przystąpić do określenia miejsc, w jakich należy poszukiwać potrzebnych do dalszej interpretacji cech. Obszarami takimi są oczy (wraz z brwiami) oraz usta.

Zastosowana metoda opiera się na segmentacji barwnej obrazu twarzy, wyznaczeniu przybliżonego położenia oczu, a następnie wydzieleniu obszarów, zawierających okolice oczu. Znając ich położenie można określić również położenie obszaru zawierającego usta.

Segmentacja oczu wykorzystuje informację, płynącą ze składowych barwnych obrazu twarzy. Tworzony jest obraz, będący różnicą składowej czerwonej i niebieskiej. Opisuje to wzór (3.11).

$$I_{RB} = I_R - I_B \tag{3.11}$$

Jak widać na rysunku 3.8, na tak uzyskanym obrazie obszar oczu jest bardzo ciemny, co odróżnia go od sąsiadujących obszarów skóry.

Pozwala to na przeprowadzenie binaryzacji obrazu I_{RB}, celem wydzielenia ciemniejszych obiektów. Próg binaryzacji należy jednak dobrać ręcznie do aktualnych warunków oświetleniowych. Wynika to z własności samej kamery, której wewnętrzna automatyka stara się często kompensować słabe oświetlenie, co powoduje dużą rozbieżność pomiędzy tymi samymi składowymi barwnymi dla różnego oświetlenia.

Aby usunąć drobne zakłócenia i zmniejszyć ilość obiektów w obrazie, zastosowano operację otwarcia (erozja a następnie dylatacja).

Rysunek 3.8: Obraz twarzy w kolorze, różnica składowych R-B, binaryznacja różnicy R-B, ten sam obraz po operacji otwarcia.

Tak przetworzony obraz twarzy jest następnie poddawany operacji etykietowania, która wydziela na nim wszystkie rozłączne obszary (obiekty). Obliczenia te zaimplementowano zgodnie z podstawowym algorytmem przedstawionym w [22]. Obraz jest przeszukiwany tylko w swej górnej części (2/3 całkowitej wysokości), gdyż właśnie tam powinny znajdować się szukane obszary oczu.

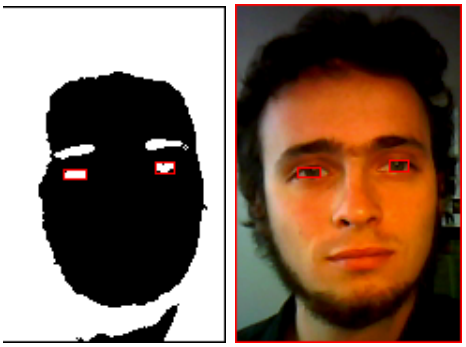
Do dalszej analizy brane są pod uwagę obiekty:

- O wysokości i szerokości większej niż 2 piksele.
- O szerokości mniejszej niż 1/3 szerokości wyznaczonego obszaru twarzy.
- O wysokości mniejszej niż 1/6 wysokości wyznaczonego obszaru twarzy.

Za obszar lewego bądź prawego oka przyjmowany jest najniżej leżący obiekt po danej stronie obrazu. Ma to na celu uniknięcie zinterpretowania ciemnych brwi jako obszaru oka. Wyznaczony w ten sposób obiekt nie zawsze dokładnie odpowiada widocznej części gałki ocznej. Mają na to wpływ cienie pojawiające się na obrazie oraz zmiana realnego kształtu takiego obiektu podczas wykonywania operacji otwarcia. Niemniej jednak obiekt taki można uznać za dobre przybliżenie położenia oka, co pozwoli na dalszą, dokładniejszą analizę.

Algorytm taki działa poprawnie przy pochyleniu głowy na bok o nie więcej niż około 20-30°. Są to jednak wartości właściwe normalnym warunkom działania systemu.

Efekty działania algorytmu wyszukiwania oczu przedstawione są na rysunku 3.9



Rysunek 3.9: Rezultat działania algorytmu wyszukiwania oczu

Mając wyznaczone (nawet niedokładnie) położenie oczu, można określić również miejsce, gdzie znajdują się usta. Co więcej, uzyskany tym sposobem obszar jest stabilny względem twarzy, niezależnie od ruchy samych ust, co bardzo pomaga przy dalszej analizie.

Do wyznaczenia regionu ust zastosowano metodę podobną do opisanej w [15]:

- Niech punkty o współrzędnych (x_l, y_l), (x_r, y_r) będą geometrycznymi środkami obszarów odpowiadającym lewemu i prawemu oku.
- Wyznaczana jest euklidesowa odległość d pomiędzy środkami zlokalizowanych obszarów oczu.

$$d = \sqrt{(x_r - x_l)^2 + (y_r - y_l)^2} \tag{3.12}$$

- Z połowy długości odcinka łączącego oczy poprowadzony jest w dół, prostopadle do niego kolejny odcinek o długości d.
- Koniec tego odcinka przyjmuje się za środek obszaru ust. Jego współrzędne (x_m, y_m) można zatem wyliczyć z równania:

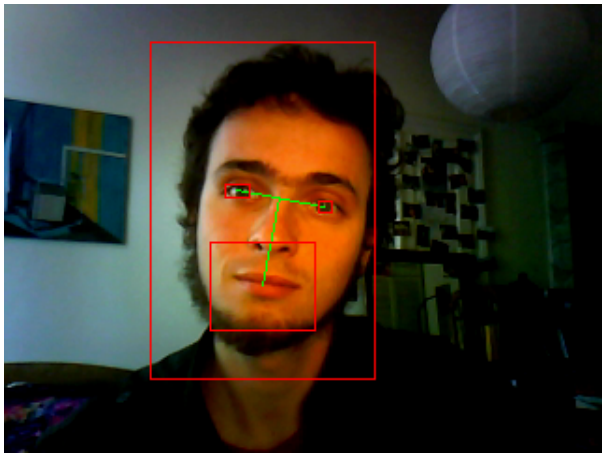
$$\begin{array}{rcl} x_m & = & \frac{x_r + x_l}{2} + (y_l - y_r) \\ y_m & = & \frac{y_r + y_l}{2} + |x_l - x_r| \end{array} \tag{3.13}$$

- Obszar ust wyznaczony jest przez prostokąt o środku w (x_m, y_m), wysokości d i szerokości $1,2 * d$.

Zarazem uzyskiwana jest wartość pierwszej cechy, mogącej świadczyć o wyrażanych emocjach: kąt pochylenia głowy na bok:

$$\alpha = \operatorname{arc\,tg}(\frac{y_l - y_r}{x_r - x_l}) \tag{3.14}$$

Rezultat działania opisanych algorytmów przedstawia rysunek 3.10. Zaznaczono na nim również omawiane odcinki, po których następuje wyznaczenie obszaru ust.



Rysunek 3.10: Rezultat działania algorytmu określania obszaru ust

3.5 Analiza regionów oczu

Znalezione w poprzednim etapie przybliżone położenie oczu jest teraz wykorzystywane do wyznaczenia obszarów, zawierających same oczy oraz brwi. Analiza tych części twarzy pozwoli na wyliczenie ważnych cech, określających mimikę twarzy: kształtu brwi oraz stopnia otwarcia powiek. Własne obserwacje pozwoliły stwierdzić, iż śledzenie położenia właśnie tych elementów najlepiej pozwoli uzyskać informację o wyrażanych gestach. Jednocześnie zrezygnowano z wykorzystania kącików oczu, jako punktów stabilnych, nie zmieniających swego położenia w trakcie ekspresji mimicznej.

Koncepcja zastosowanego rozwiązania (zastosowanie projekcji dla fragmentów obrazu) czerpie inspirację z rozwiązań opisanych w rozdziale 2.2.1, oraz z pracy [16]. Wszystkie użyte wielkości i położenia obszarów zainteresowania ustalono doświadczalnie.

Obszar zainteresowania dla danego oka jest wycinany z kolorowego obrazu całej ramki. Jego granicami jest prostokąt zbudowany wokół środka znalezionego uprzednio obiektu ((x_l, y_l) dla oka lewego i (x_r, y_r) dla prawego) tak, aby punkt ten znajdował się w połowie szerokości obszaru oraz $\frac{1}{3}$ wysokości (w obszarze musi się także znaleźć leżąca nad okiem brew). Prostokąt ten ma wysokość równą odległości d, zaś szerokość równą $\frac{2}{3}d$.

Obraz taki jest następnie obracany o kąt $-\alpha$ (ze wzoru (3.12)), dzięki czemu dalsze obliczenia odbywają się niezależnie od aktualnego pochylenia głowy.

Rysunek 3.11 przedstawia efekt tych przekształceń.

Rysunek 3.11: Wydzielony obszar oka (w skali szarości) po dokonaniu obrotu kompensującego pochylenie głowy.

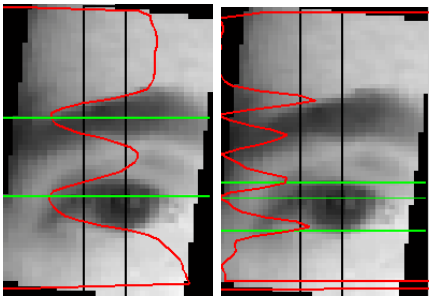
3.5.1 Wyznaczenie położenia oka i powiek. Analiza obszaru oka w przestrzeni R-B.

Aby wyznaczyć dokładne położenie oka, znów dokonuje się przekształcenia obszaru oka do przestrzeni opisanej wzorem (3.11). W takim obrazie dokonuje się analizy pionowego pasa o szerokości równej $\frac{1}{5}$ całego obszaru, położonego w jego środku (pierwszy rys. 3.12).

Dla takiego centralnego pasa wyliczana jest jego projekcja (histogram) wertykalna - funkcja, której wartościami są sumy pikseli obrazu w kolejnych wierszach. Opisuje ją wzór (3.15).

$$p_v(y) = \sum_{x=x_{min}}^{x<x_{max}} I(x,y) \tag{3.15}$$

Tak otrzymaną projekcję poddaje się wygładzeniu za pomocą filtracji dolnoprzepustowej. Dwa największe minima tej funkcji odpowiadają najciemniejszym regionom zawartym w analizowanym pasie. Jednym z nich jest oko, drugim zaś - fragment brwi. Ich współrzędne pionowe zapisywane są do późniejszej analizy.

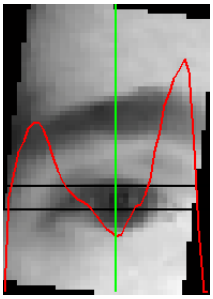


Rysunek 3.12: Określenie położenia oka w pionie, wyznaczenie przybliżonego poziomu brwi oraz określenie położenia powiek

W celu uzyskania informacji o położeniu powiek, należy znów zbadać projekcję wertykalną centralnego regionu - tym razem pod kątem występowania dużych zmian jasności, odpowiadających krawędziom. W bezpośrednim sąsiedztwie środka oka takie wyraźne krawędzie odpowiadają miejscom styku oka z powiekami. Informację o ich położeniu można uzyskać, analizując moduł gradientu projekcji (wzór (3.16)).

$$g(p_v(x)) = |p_v(x+1) - p_v(x)| \tag{3.16}$$

Maksima tej funkcji otaczające wyliczoną współrzędną y oka odpowiadają krawędziom powiek (drugi z rys. 3.12).



Rysunek 3.13: Określenie położenia oka w poziomie

Następnie wyznacza się położenie oka w poziomie. W tym celu analizowana jest projekcja horyzontalna (wzór (3.17)) poziomego pasa o wysokości 5 pikseli, otaczającego wyznaczoną przed chwilą poziomą linię oka. Tak jak poprzednio, dokonywana jest filtracja dolnoprzepustowa projekcji oraz szukanie minimum. Odpowiada ono współrzędnej x oka. Obrazuje to rysunek 3.13.

$$p_h(x) = \sum_{y=y_min}^{y<y_max} I(x,y) \tag{3.17}$$

3.5.2 Określenie kształtu brwi. Analiza obszaru oka w skali szarości.

Mając współrzędne środka oka można przystąpić do określenia kształtu brwi. Tym razem jednak operacje lepiej prowadzić na obrazie w skali szarości, gdyż na takim brwi są lepiej widoczne.

W zaproponowanym rozwiązaniu kształt brwi jest aproksymowany łamaną, składającą się z dwóch prostych odcinków. Jej kształt otrzymuje się, badając projekcje wertykalne pionowych fragmentów obrazu, umieszczonych w określonych miejscach. Można określić to jako próbkowanie współrzędnej y brwi w 3 miejscach.

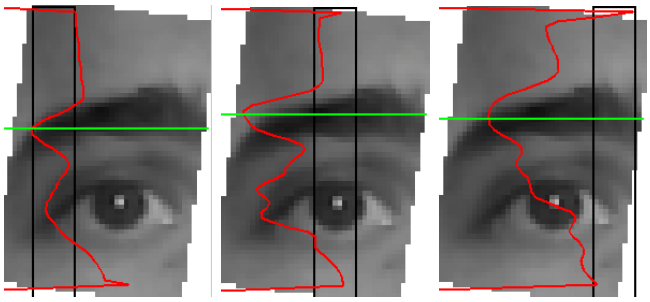
Jeżeli znalezioną wcześniej współrzędną x oka oznaczyć przez x_e, zaś szerokość całego obszaru oka przez w_e, wówczas każdy z pionowych pasów ma szerokość $\frac{1}{5}w_e$ i obejmuje zakres odpowiednio:

- $\langle x_e - \frac{2w_e}{5}, x_e - \frac{w_e}{5} \rangle$ - lewa strona
- $\langle x_e - \frac{w_e}{10}, x_e + \frac{w_e}{10} \rangle$ - centrum

- $\langle x_e + \frac{w_e}{5}, x_e + \frac{2w_e}{5} \rangle$ - prawa strona

Położenie tych pasów zaznaczono na rys. 3.14.

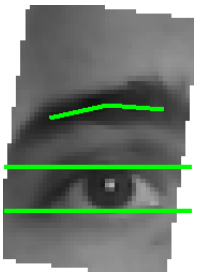
Analiza projekcji wertykalnej z każdego obszaru pozwala wyznaczyć położenie fragmentu brwi w nim się zawierającego. Jest nim minimum leżące najbliżej przybliżonego położenia brwi, wyliczonego w rozdziale 3.5.1 i zaznaczonego na rys 3.12.



Rysunek 3.14: Analiza kształtu brwi w 3 miejscach

Cały etap analizy obszaru oczu pozwala uzyskać dane, obrazowo przedstawione na rysunku 3.15 Do późniejszej analizy gestów mimicznych wykorzystywane są:

- Odległość między górną powieką a dolną.
- Odległości pomiędzy każdym z 3 punktów próbkowania kształtu brwi a poziomą linią oka.

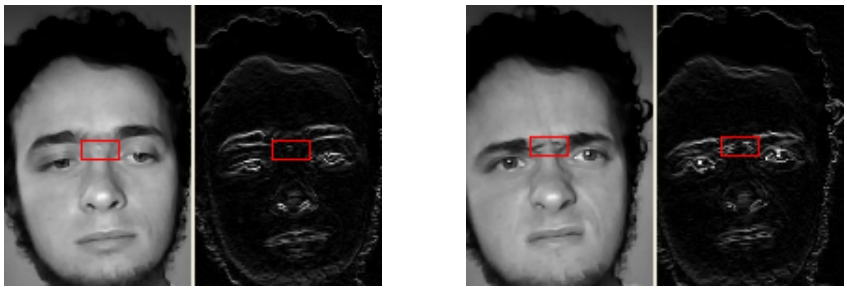


Rysunek 3.15: Rezultat analizy obszaru oka - położenie powiek oraz kształt brwi

3.5.3 Wykrywanie marszczenia brwi. Analiza obrazu gradientowego.

Jednym z bardzo charakterystycznych gestów, mogących pojawić się na twarzy jest zmarszczenie brwi nad nosem. Niesie on wiele informacji o wyrażanych emocjach (szerzej omówiono to w rozdziale 3.7). Dlatego też utworzony system wyposażono w mechanizm pozwalający wykrywać taką cechę.

W celu stwierdzenia istnienia bądź braku zmarszczek nad nosem, analizowany jest odpowiedni obszar obrazu gradientowego twarzy (rys. 3.16). Obraz taki zawiera informację o nagłych zmianach jasności sąsiednich pikseli, co odpowiada krawędziom (np. zmarszczkom). Nagromadzenie w danym rejonie dużej liczby krawędzi powoduje wzrost sumarycznej jasności takiego obszaru. Zastosowanie takiej formy obrazu powoduje uniezależnienie od zmiennego wpływu oświetlenia i cieni na jasność danego regionu.



Rysunek 3.16: Obrazy, w skali szarości i gradientowe, dla twarzy neutralnej i ze zmarszczonymi brwiami. Zaznaczono analizowany obszar.

Dodatkowo, zdecydowano się na zastosowanie tylko gradientu pionowego, a więc przedstawiającego jedynie zmienność jasności pikseli leżących w sąsiednich wierszach. Podyktowane zostało to obserwacją, iż analizowane zmarszczki mają w większości przebieg horyzontalny. Dzięki temu jeszcze lepiej można wychwycić ich obecność.

Obraz gradientu pionowego wyliczany jest z obrazu w skali szarości zgodnie ze wzorem (3.18).

$$I_{grad}(x, y) = |I(x, y) - I(x, y - 1)| \tag{3.18}$$

Aby wzmocnić przyrost bezwzględnej jasności analizowanego obszaru w przypadku pojawienia się zmarszczek, do dalszej analizy obraz taki pomnożono dwukrotnie.

Obszar analizy ma stały rozmiar i jest kwadratem o szerokości 20 i wysokości 10 pikseli. Jego położenie wyznaczane jest, podobnie jak położenie obszaru ust, na podstawie wyliczonego uprzednio położenia oczu (rozdział 3.4).

Jeżeli punkty o współrzędnych (x_l, y_l), (x_r, y_r) są geometrycznymi środkami obszarów odpowiadającym lewemu i prawemu oku, wówczas współrzędne (x_w, y_w) środka obszaru analizy wyliczane są ze wzorów:

$$\begin{aligned} x_w &= \frac{x_r + x_l}{2} \\ y_w &= \frac{y_r + y_l}{2} - \frac{d}{5} \end{aligned} \tag{3.19}$$

gdzie d jest euklidesową odległością pomiędzy oczami, wyliczoną zgodnie ze wzorem (3.14).

Regionu ten ma stosunkowo małą wielkość oraz leży w bliskości linii łączącej oczy. Dlatego też pochylenie twarzy w dopuszczalnych przez system granicach nie wpływa zbytnio na jego położenie i może być pominięte w wyliczeniach.

Analiza opisanego obszaru sprowadza się do wyliczenia sumy jasności pikseli dla obrazu gradientowego. Przy pojawieniu się zmarszczek nad nosem, wartość ta rośnie. Można więc przyjąć pewien próg, powyżej którego można mówić o wykryciu zmarszczek. Wzór (3.20) przedstawia wyliczanie takiej sumy.

$$sum_w = \sum_{x=x_w-10}^{x_w+10} \sum_{y=y_w-5}^{y_w+5} I_{grad}(x, y) \tag{3.20}$$

3.6 Analiza regionu ust

Uprzednio wyznaczony obszar ust (rozdział 3.4) jest obracany o kąt $-\alpha$, podobnie jak obszary oczu. Następnie jest analizowany pod kątem określenia kształtu warg. Usta mogą wyrażać dużą liczbę gestów, przyjmując różne ułożenie. Kluczowym staje się więc zagadnienie wyznaczenia bieżącego ich kształtu. Na wstępie należy jednak zauważyć, iż do określenia większości gestów wystarcza informacja o położeniu kącików ust oraz górnej i dolnej wargi. Subtelne różnice kształtu, tak istotne np. w zagadnieniu czytania z ruchu ust (np. [15]) nie mają już takiego wpływu na rozpoznawany gest mimiczny.

3.6.1 Segmentacja koloru ust

Podstawowym zagadnieniem w określaniu kształtu ust jest ich segmentacja, czyli takie przetworzenie obrazu, które maksymalnie zróżnicuje obszar ust (warg) i otoczenia.

W tym celu wykorzystano algorytm, przedstawiony w [21]. Autorzy postanowili wykorzystać informację płynącą z koloru ust. Zauważono, iż wargi zawsze mają inny odcień niż otaczająca skóra (przesunięty bardziej ku czerwieni). Dzięki temu, stosując odpowiednie przekształcenia, można wydzielić obiekt, reprezentujący usta.

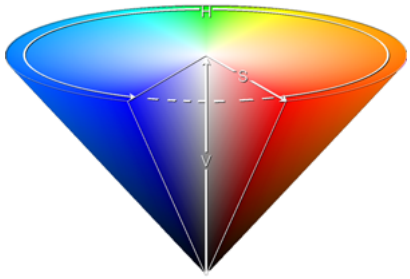
Pierwszym krokiem jest konwersja przestrzeni barwnej obrazu z modelu RGB do HSV (*Hue, Saturation, Value*). Ideę modelu HSV przedstawiono na rys. 3.17.

Obraz w przestrzeni HSV składa się z 3 składowych, podobnie jak obraz RGB. Poszczególne składowe odpowiadają kolejno: odcieniowi (kolorowi), nasyceniu i jasności. Do dalszej analizy wykorzystywana jest pierwsza składowa - H.

Składowa H obrazu obszaru ust przedstawiona jest na rys. 3.18.

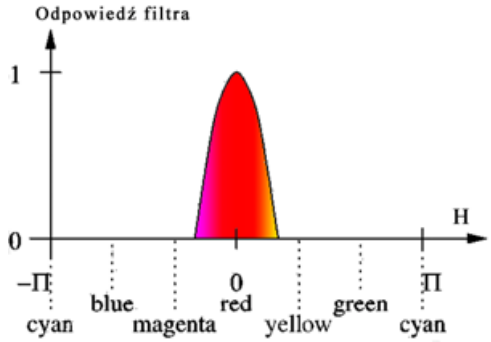
Aby wydobyć z takiego obrazu obszar odpowiadający odcieniowi ust, zastosowano filtr o charakterystyce przedstawionej na rys. 3.19.

Jak łatwo zauważyć, filtr taki pozostawia na obrazie jedynie piksele o odcieniu zbliżonym do czerwonego (kolor ust). Ogólna postać filtra wyraża się wzorem



Rysunek 3.17: Model barw HSV (Źródło: [2])

Rysunek 3.18: Składowa H obszaru ust



Rysunek 3.19: Charakterystyka filtra barwnego (Na podstawie: [21])

(3.21)

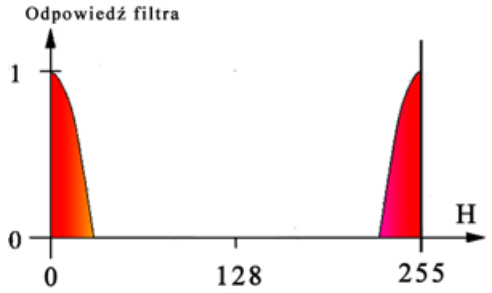
$$f(h) = \begin{cases} 1 - \frac{(h-h_0)^2}{w^2} & |h - h_0| \leqslant w \\ 0 & |h - h_0| > w \end{cases} \tag{3.21}$$

Filtr taki pozwala dowolnie dobierać zarówno położenie przepuszczanego pasma, jak i jego szerokość. Odbywa się to poprzez zmianę wartości parametrów h_0 (pasmo, kolor), oraz w (szerokość pasma).

Zmienność składowej H jest w modelu HSV cykliczna, co widać na rys. 3.17. Dlatego też wartość H przyjęło się wyrażać w stopniach. Kątowi 0°odpowiada barwa czerwona, następnie następują odcienie pomarańczowe, żółte, zielone, niebieskie, fioletowe. Po osiągnięciu wartości 360° składowa H znów przedstawia odcień czerwony. Ponieważ w reprezentacji cyfrowej taka abstrakcyjna skala jest

niemożliwa do osiągnięcia, stosuje się zwykłą skalę liniową. W przypadku skali 256-stopniowej wartość 0 odpowiada kolorowi czerwonemu, 128 - zielonemu, zaś 255 to znów czerwień.

Stanowi to pewne utrudnienie w stosowaniu przedstawionego filtra. Chcąc otrzymać wartości składowej H z bezpośredniego otoczenia czerwieni, należałoby użyć filtra o charakterystyce przedstawionej na rys. 3.20.

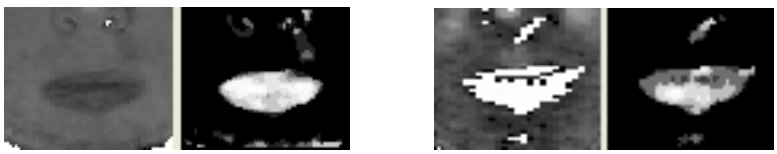


Rysunek 3.20: Charakterystyka filtra barwnego dla rzeczywistej reprezentacji odcieni w skali 256 stopniowej

Dlatego też, aby ułatwić operację filtracji, zastosowano przesunięcie cykliczne wartości pikseli o 128, dzięki czemu kolorowi czerwonemu i zbliżonym odpowiadają wartości bliskie 128.

Dla tak przesuniętej skali barwnej dobrano doświadczalnie parametry działania filtra: $h_0 = 140$ oraz $w = 10$. Ponieważ rejestrowany przez kamerę odcień ust może się zmieniać w zależności od aktualnych warunków oświetleniowych, dlatego też wartości tych parametrów mogą być korygowane ręczenie na bieżąco. Na rys. 3.21 przedstawiono efekt filtracji dla obrazów obszaru ust, pozyskanych przy różnych warunkach oświetleniowych. Dzięki modyfikacji parametrów filtra, otrzymano żądany efekt segmentacji.

Tak uzyskany obraz ust poddaje się operacji binaryzacji z progiem równym połowie skali szarości (128).

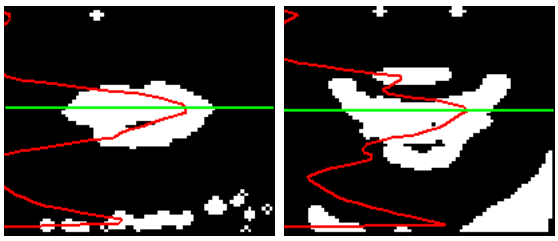


Rysunek 3.21: Składowa H obszarów ust i te same obrazy po przesunięciu barw i filtracji

3.6.2 Analiza kształtu ust

Binarny obraz ust jest poddawany prostej analizie kształtu, co pozwala na określenie wyrażanego gestu.

Na początku szukana jest przybliżona pozycja ust w pionie. W tym celu analizuje się projekcję wertykalną (opisaną wzorem (3.15)) takiego obrazu, poddaną filtracji dolnoprzepustowej. Za przybliżoną współrzędną Y ust przyjmuje się maksimum takiej projekcji (rys 3.22)



Rysunek 3.22: Wyznaczanie przybliżonego poziomu ust dla różnego ich kształtu

Jak widać na rysunku 3.22, oprócz samych ust na takim obrazie mogę pojawić się inne obiekty, reprezentujące najczęściej fragment nosa bądź, jak w przypadku autora, skórę przesłoniętą zarostem. Fragmenty takie mają podobny kolor jak wargi, dlatego też również podlegają segmentacji opisanej w rozdziale 3.6.1.

Aby skompensować ich wpływ na rezultaty analizy projekcji wertykalnej, zastosowano dodatkowe kryterium:

- Pod uwagę brane są dwa największe maksima projekcji.
- Jeżeli maksimum leżące niżej jest ponaddwukrotnie większe niż to wyższe, wówczas przyjmowane jest ono za współrzędną Y ust.

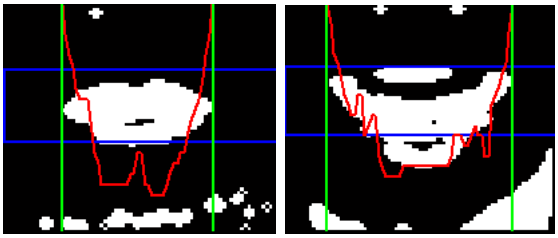
  Odpowiada to sytuacji, gdy na obrazie nie pojawiają się artefakty pochodzące od zarostu na brodzie (niżej), a jedynie małe punkty reprezentujące fragment nosa (wyżej).
- Jeżeli stosunek wartości maksimum dolnego do górnego jest mniejszy niż 2, wówczas za współrzędną Y ust przyjmuje się to drugie.

  Dzieje się tak w przypadku, gdy duże obiekty reprezentujące brodę pojawiają się na obrazie.

Następnym krokiem jest określenie położenia kącików ust. Aby znaleźć ich współrzędne X, analizowany jest prostokątny wycinek obrazu, obejmujący całą jego szerokość i rozciągający się 10 pikseli w górę i w dół od wyznaczonej współrzędnej Y ust. Zaznaczono go na rys. 3.23.

Wyliczając projekcję horyzontalną takiego obszaru (wzór (3.17)), badane są miejsca, w których kończy się obiekt (usta). Poczynając od środka obszaru, szu-

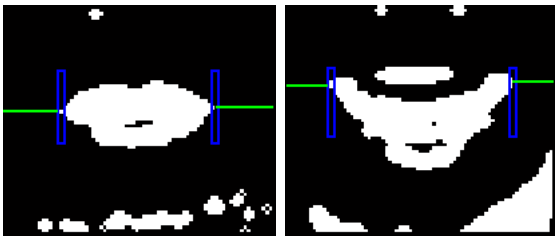
kane są pierwsze punkty (idąc w lewą i w prawą stronę), dla których funkcja projekcji przyjmuje wartość 0. Ilustruje to rysunek 3.23.



Rysunek 3.23: Wyznaczanie położenia w poziomie kącików ust

Wyznaczenie współrzędnych Y kącików odbywa się oddzielnie dla każdego z nich.

W poprzednim kroku była liczona projekcja horyzontalna pewnego regionu w celu określenia bocznych krańców obiektu odpowiadającego ustom. Teraz wystarczy przeanalizować skrajnie lewą i skrajnie prawą kolumnę pikseli tego regionu. Położenie niezerowego piksela odpowiada współrzędnej Y danego kącika. W przypadku gdy kolumna zawiera szereg niezerowych pikseli, za położenie kącika przyjmuje się środek tego bloku (rys. 3.24).



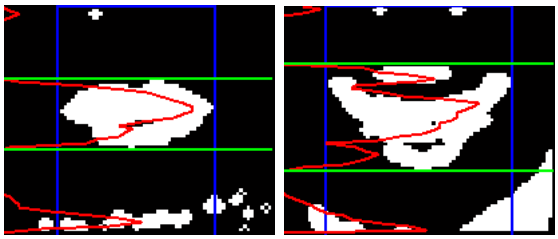
Rysunek 3.24: Wyznaczanie położenia w pionie kącików ust

Ostatnim etapem jest wyznaczenie współrzędnych Y warg. Przyjęto, iż są one reprezentowane przez górne i dolne krańce obiektu odpowiadającego ustom.

W celu wyznaczenia tych krańców badana jest projekcja wertykalna prostokątnego obszaru ograniczonego w poziomie przez współrzędne X kącików. Zaznaczono go na rys. 3.25. Znów szukane są miejsca, dla którego projekcja taka przyjmuje wartości niezerowe. Ponieważ miejsc takich może być kilka, zakłada się iż właściwe jest to, w obrębie którego leży wyznaczona na początku przybliżona oś Y ust.

Zgromadzone do tej pory dane, naniesione obrazowo na analizowany obszar, zaprezentowane są na rysunku 3.26.

Dalsza analiza regionu ust pod kontem prezentowanego gestu obejmuje odległości mierzone od geometrycznego środka obszaru ust. W globalnym układzie współrzędnych twarzy jego położenie to (x_m, y_m), wyznaczone wzorem (3.13), tak

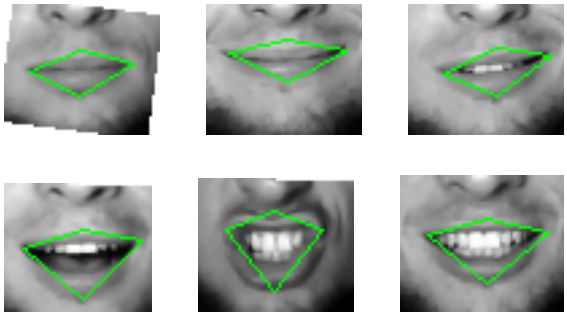


Rysunek 3.25: Wyznaczanie położenia warg

więc badane jest *de facto* położenie punktów charakterystycznych ust względem położenias oczu.

Tak więc w dalszej analizie badane są odległości od środka obszaru ust do:

- Współrzędnych X kącików ust wzdłuż osi X.
- Współrzędnych Y kącików ust wzdłuż osi Y.
- Współrzędnych Y górnej i dolnej wargi wzdłuż osi Y.



Rysunek 3.26: Przykładowe działanie algorytmu określania kształtu ust

3.6.3 Wykrywanie widoczności zębów

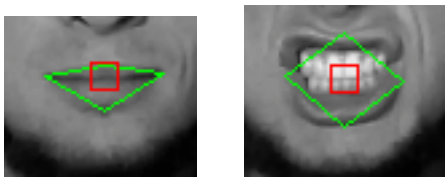
Ostatnią, lecz bardzo ważną cechą twarzy braną pod uwagę przy dalszej klasyfikacji gestów jest fakt pokazania bądź nie zębów. Zależnie od bieżącego kształtu ust ich pojawienie się może oznaczać szeroki, szczery uśmiech bądź też gniew, agresję.

Stwierdzenie, czy zęby są widoczne zrealizowano podobnie jak wykrywanie marszczenia brwi (rozdział 3.5.3), z tym że tym razem badany jest zwykły obraz w skali szarości.

Na podstawie położenia kącików ust wyznaczany jest prostokątny obszar analizy o rozmiarze 10 na 10 pikseli. Jeżeli (x_l, y_l), (x_r, y_r) są współrzędnymi kącików, obszar taki ma środek w punkcie (x_t, y_t) (leży 5 pikseli poniżej środka linii łączącej kąciki):

$$\begin{aligned} x_t &= \frac{x_r+x_l}{2} \\ y_t &= \frac{y_r+y_l}{2} - 5 \end{aligned} \tag{3.22}$$

Rysunek 3.27 ukazuje położenie takiego obszaru na obrazie ust.



Rysunek 3.27: Obszar analizy pojawienia się zębów

Pokazanie białych zębów powoduje gwałtowny przyrost sumarycznej wartości pikseli w takim obszarze, dlatego też można ustalić próg wartości, powyżej którego możemy mówić o widoczności zębów. Wzór (3.23) opisuje wyliczanie takiej sumy (I oznacza obraz w skali szarości).

$$sum_t = \sum_{x=x_t-5}^{x_t+5} \sum_{y=y_t-5}^{y_t+5} I(x,y) \tag{3.23}$$

3.7 Klasyfikacja rozpoznanych cech

Opisane w poprzednich rozdziałach algorytmy służą pozyskaniu wartości pewnych parametrów. Są nimi wzajemne odległości pomiędzy poszczególnymi punktami charakterystycznymi lub też sumaryczne wartości pikseli w danym regionie. Dalsza interpretacja tak pozyskanych wielkości pozwala na wnioskowanie o obecności na twarzy danych gestów mimicznych, co jest zasadniczym celem niniejszej pracy.

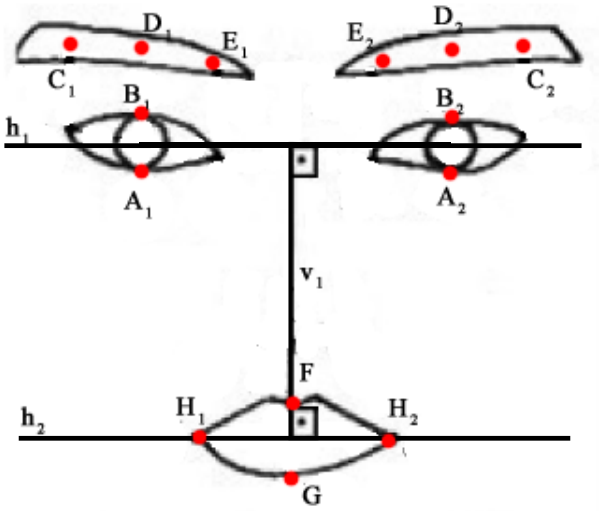
3.7.1 Wyliczanie parametrów twarzy

Dobór parametrów opisywany był już pokrótce na końcu każdego podrozdziału. W tym miejscu nastąpi dokładne ich zestawienie i usystematyzowanie.

Rysunek 3.28 przedstawia rozmieszczenie śledzonych przez system punktów charakterystycznych twarzy wraz z nadanymi im etykietami. Zaznaczone są również linie odniesienia: poziome $h1$ i $h2$ (linia oczu i przybliżona linia ust) oraz pionowa $v1$ (oś twarzy). Oznaczenia te użyte są w tabeli 3.1 do opisania odcinków, których długości system analizuje. Odcinki te (oraz odpowiednie obszary) zaznaczono na rysunku 3.29.

Zastosowany dobór punktów i parametrów jest podobny jak w systemie ISFER, opisanym w rozdziale 2.2.1 (rys. 2.8) Różni się on jednak w wielu miejscach,

głównie z uwagi na zastosowanie innych algorytmów wykrywania cech. Przykładowo, zrezygnowano ze śledzenia kącików oczu i punktów charakterystycznych nosa.



Rysunek 3.28: Oznaczenie punktów charakterystycznych twarzy

Tabela 3.1 zawiera wykaz parametrów, wyliczanych w procesie analizy obrazu twarzy.

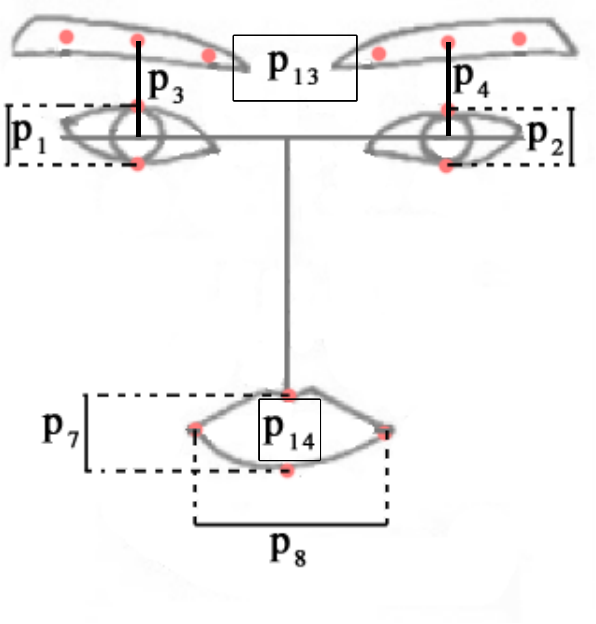
Ponieważ analizowana twarz może zmieniać swoją odległość od kamery, a co za tym idzie, wielkość w kadrze, zatem bezwzględne wartości parametrów również ulegają zmianie przy takim ruchu.

Dlatego też każde nowo pozyskane wartości parametrów od p_1 do p_{11} są poddawane procesowi normalizacji, co przedstawia wzór (3.24). Punktem odniesienia jest bieżąca odległość między oczami d (wzór (3.14)), która jest normalizowana do wartości 100. Wówczas każda z badanych odległości może być traktowana jako procentowa część aktualnej odległości między oczami.

$$p_{norm} = p * \frac{100}{d} \tag{3.24}$$

3.7.2 Wyznaczanie cech twarzy na podstawie parametrów

Analiza samych liczbowych wartości parametrów, nawet po normalizacji, nie jest jeszcze dobrą metodą określania gestów. Każdy użytkownik systemu może mieć nieco inny kształt twarzy i jej poszczególnych elementów. Dlatego też trudno założyć z góry, jakie wartości danego parametru będą odpowiadać konkretnemu zachowaniu.



Rysunek 3.29: Oznaczenie niektórych z mierzonych odległości i obszarów (parametry)

Zastosowano zatem metodę, polegającą na porównywaniu bieżących wartości parametrów z odpowiadającymi im wielkościami dla twarzy neutralnej, pozyskanymi na początku działania systemu. Dzieje się tak ze wszystkimi parametrami, oprócz p_{12}, który jest wartością bezwzględną.

Badane są zatem różnice między wartościami bieżącymi a początkowymi. Jeżeli różnica taka przekroczy odpowiednio ustalony próg (górny lub dolny), wówczas można mówić o odnotowaniu pewnej cechy wyglądu twarzy, związanej z danym parametrem. (tabela 3.2). Pozwala to na zmniejszenie ilości analizowanych informacji, gdyż w większości przypadków nie jest istotna konkretna wartość różnicy parametrów, tylko fakt, czy taka zmiana w ogóle zaszła.

Każdy z parametrów ma inaczej określone progi. Tabela 3.3 zawiera zestawienie wszystkich możliwych wartości przyjmowanych przez wszystkie cechy. Podane progi wartości parametrów zostały ustalone doświadczalnie.

Przyjęcie przez daną cechę wartości różnej od zera można już traktować jako pojawienie się odpowiedniego jednostkowego gestu mimicznego. Gesty bardziej złożone powstają z połączenia gestów jednostkowych.

Tabela 3.1: Zestawienie parametrów wykorzystywanych w systemie. Przez $|AB|$ oznaczono odległość między punktami A i B, natomiast przez $|Ah|$ - odległość punktu A od prostej h.

Nazwa	**Wartość**	**Opis**
p_1	$\|A_1B_1\|$	Otwarcie lewego oka
p_2	$\|A_2B_2\|$	Otwarcie prawego oka
p_3	$\|D_1h_1\|$	Wysokość lewej brwi
p_4	$\|D_2h_1\|$	Wysokość prawej brwi
p_5	$\|C_1h_1\| - \|E_1h_1\|$	Nachylenie lewej brwi
p_6	$\|C_2h_1\| - \|E_2h_1\|$	Nachylenie prawej brwi
p_7	$\|Fh_2\| + \|Gh_2\|$	Otwarcie ust
p_8	$\|H_1v_2\| + \|H_2v_2\|$	Szerokość ust
p_9	$\|H_1v_2\| - \|H_2v_2\|$	Przesunięcie ust na bok
p_{10}	$\|H_1h_2\|$	Uniesienie lewego kącika
p_{11}	$\|H_2h_2\|$	Uniesienie prawego kącika
p_{12}	α (wzór (3.12))	Kąt nachylenia głowy
p_{13}	sum_w (wzór (3.20))	Suma pikseli obrazu gradientowego w obszarze wykrywania zmarszczek
p_{14}	sum_t (wzór (3.23))	Suma pikseli obrazu w skali szarości w obszarze wykrywania zębów

Tabela 3.2: Zbiór wartości przyjmowanych przez cechy

Wartość cechy	**Kiedy występuje**
$+1$	Różnica między parametrem bieżącym a początkowym jest większa od zadanego progu górnego
0	Różnica między parametrem bieżącym a początkowym jest pomiędzy progiem dolnym a górnym
-1	Różnica między parametrem bieżącym a początkowym jest mniejsza od zadanego progu dolnego

Tabela 3.3: Lista cech i przyjmowanych przez nie wartości. Dla każdej cechy podano progi wartości odpowiadającego parametru, po przekroczeniu których cecha osiąga daną wartość.

Cecha	**Wartości cechy**	**Progi**	**Opis (gest)**
f_1	0		Lewe oko normalne
	$+1$	> 1	Lewe oko otwarte szeroko
f_2	0		Prawe oko normalne

Cecha	**Wartości cechy**	**Progi**	**Opis (gest)**
	+1	> 1	Prawe oko otwarte szeroko
f_3	0		Lewa brew normalna
	-1	< -2	Lewa brew opuszczona
	+1	> 7	Lewa brew uniesiona
f_4	0		Prawa brew normalna
	-1	< -2	Prawa brew opuszczona
	+1	> 7	Prawa brew uniesiona
f_5	0		Lewa brew normalna
	-1	< -2	Wewnętrzny koniec lewej brwi uniesiony
	+1	> 2	Zewnętrzny koniec lewej brwi uniesiony
f_6	0		Prawa brew normalna
	-1	< -2	Wewnętrzny koniec prawej brwi uniesiony
	+1	> 2	Zewnętrzny koniec prawej brwi uniesiony
f_7	0		Usta normalne
	-1	< -7	Usta zaciśnięte
	+1	> 7	Usta otwarte
f_8	0		Usta normalne
	-1	< -4	Usta wąskie
	+1	> 10	Usta szerokie
f_9	0		Usta po środku
	-1	< -9	Usta przesunięte w prawo
	+1	> 9	Usta przesunięte w lewo
f_{10}	0		Lewy kącik normalny
	-1	> 5	Lewy kącik opuszczony
	+1	< -5	Lewy kącik uniesiony
f_{11}	0		Prawy kącik normalny
	-1	> 5	Prawy kącik opuszczony
	+1	< -5	Prawy kącik uniesiony
f_{12}	0		Głowa ustawiona pionowo
	-1	$< -10^\circ$	Głowa pochylona w prawo
	+1	$> 10^\circ$	Głowa pochylona w lewo
f_{13}	0		Brak zmarszczek nad nosem
	+1	> 600	Zmarszczki nad nosem
f_{14}	0		Zęby niewidoczne
	+1	> 2000	Zęby widoczne

System jest w stanie zidentyfikować w sumie 24 różne gesty jednostkowe (odpowiadają cechom o wartościach różnych od 0). Lista aktualnie rejestrowanych

gestów zwracana jest użytkownikowi systemu.

3.7.3 Określanie emocji na podstawie gestów

Na podstawie otrzymanej listy gestów system jest w stanie identyfikować również pewne stany emocjonalne, wyrażane mimiką twarzy:

- zadowolenie (uśmiech),
- śmiech (szeroki, serdeczny uśmiech),
- smutek,
- zaskoczenie,
- strach,
- niesmak,
- złość (gniew).

W celu identyfikacji tych emocji skonstruowano odrębne, równolegle działające klasyfikatory:

System regułowy

Jego działanie bazuje na wnioskowaniu z obecności gestów jednostkowych. Oparto się na własnych empirycznych spostrzeżeniach odnośnie cech charakteryzujących dany stan emocjonalny.

Pozytywna identyfikacja danego stanu następuje w chwili, gdy zarejestrowane gesty spełniają listę warunków stawianych każdemu z nich. Tabela 3.4 przedstawia zestawienie tych wymagań.

Tabela 3.4: Warunki logiczne zastosowane do rozpoznawania emocji

Emocje	**Warunek logiczny (lista wymaganych wartości cech)**
Zadowolenie	$f_8 > 0 \wedge f_{10} > 0 \wedge f_{11} > 0 \wedge f_{13} = 0$
Śmiech	$f_8 > 0 \wedge f_{10} > 0 \wedge f_{11} > 0 \wedge f_{13} = 0 \wedge f_{14} > 0$
Smutek	$f_{10} < 0 \wedge f_{11} < 0 \wedge f_3 < 1 \wedge f_4 < 1 \wedge (f_5 < 1 \vee f_6 < 1)$
Zaskoczenie	$f_3 > 0 \wedge f_4 > 0 \wedge f_{13} = 0 \wedge f_5 < 1 \wedge f_6 < 1$

Emocje	**Warunek logiczny (lista wymaganych wartości cech)**
Strach	$f_3 > 0 \wedge f_4 > 0 \wedge f_{13} = 0 \wedge f_5 < 1 \wedge f_6 < 1 \wedge f_{10} < 0 \wedge f_{11} < 0$
Niesmak	$f_{13} > 0 \wedge ((f_9 < 0 \wedge f_{11} < 0) \vee (f_9 > 0 \wedge f_{10} < 0))$
Złość	$f_{14} > 0 \wedge f_8 < 1 \wedge f_7 > 0 \wedge (f_{13} > 0 \vee (f_5 > 0 \vee f_6 > 0))$

Przy konstrukcji reguł zwrócono uwagę na wyniki wstępnych testów systemu, co pozwoliło wykorzystać te cechy, które system rozpoznaje najlepiej i jednocześnie nie uzależniać wyniku od cech często błędnie wykrywanych.

Sieci neuronowe

Kolejne dwa klasyfikatory są próbą implementacji sztucznych sieci neuronowych. W tym wypadku reguły wnioskowania nie są narzucone z góry, sieci same je wypracowują w procesie uczenia.

Obydwie sieci w procesie rozpoznawania emocji są odrębnymi klasyfikatorami. Działają równolegle i niezależnie, co pozwala porównać uzyskiwane przez nie wyniki i wychwycić zalety i wady każdego rozwiązania.

W procesie uczenia pierwszej z sieci podawano na jej wejście zestawy parametrów (tab. 3.1) dla różnych przykładów twarzy z przygotowanego wcześniej zbioru uczącego. Jednocześnie porównywano otrzymane wyjście sieci (aktywację neuronów wyjściowych) z oczekiwanym rezultatem (rodzaj emocji intencjonalnie prezentowanych przez daną twarz). Dzięki algorytmowi uczenia *backpropagation* sieć taka po wielu cyklach uczenia powinna się nauczyć sama rozpoznawać prezentowane jej później zestawy parametrów jako reprezentację konkretnych emocji.

Druga z sieci działa praktycznie identycznie, różni się jedynie formą danych wejściowych. Dla każdej twarzy (zarówno uczącej jak i rozpoznawanej) wpierw dokonuje się klasyfikacji gestów jednostkowych zgodnie z zasadami z tabeli 3.3. Dopiero tak określone wartości tych gestów ($-1, 0$ lub 1) podaje się do odpowiednich neuronów wejściowych sieci.

Użyte zostały identyczne co do struktury sieci typu *backpropagation* z jedną warstwą ukrytą. Liczebność neuronów w poszczególnych warstwach jest następująca:

- 14 - warstwa wejściowa

  Każde wejście odpowiada jednemu parametrowi twarzy (tab. 3.1).

- 10 - warstwa ukryta

Zadanie wyznaczenia ilości neuronów warstwy ukrytej jest przedmiotem zaawansowanych procesów optymalizacji struktury sieci, co nie jest kluczowym tematem niniejszej pracy. Dlatego też w wyznaczeniu ich ilości posłużono się ogólną wskazówką (3.25) [9].

$$N_u = \sqrt{N_{we} * N_{wy}} \tag{3.25}$$

- 8 - warstwa wyjściowa

  Każde z wyjść odpowiada jednemu stanowi emocjonalnemu. Dodatkowe jedno wyjście reprezentuje brak emocji.

Sieci nauczone zostały za pomocą zbioru uczącego. W jego skład weszły po 3 przykłady każdego stanu emocjonalnego (łącznie 24 zestawy danych). Pierwszej z sieci przedstawiano zestawy parametrów (odległości pomiędzy punktami twarzy). Druga z nich na wejście otrzymywała zestawy wykrytych w danym momencie cech - efekt progowania parametrów (tab. 3.3). Obydwie sieci uczono z nauczycielem, zadając żądaną odpowiedź przy danym przykładzie twarzy.

Dodatkowe szczegóły procesu uczenia:

- Neurony posiadają sigmoidalną funkcję aktywacji o współczynniku $\beta = 0,5$.
- Ilość iteracji procesu uczenia została narzucona odgórnie i wynosiła 10000.
- Sieć trenowano ze współczynnikiem uczenia dynamicznie zmniejszającym się wraz z postępem procesu uczenia. Pozwala to na (teoretycznie) lepsze jej nauczenie [26]. Początkowa wartość to 0,6.
- Zastosowano mechanizm *momentum* (bezwładności) w celu uniknięcia dążenia do minimów lokalnych w funkcji błędu sieci [26]. Współczynnik *momentum* również maleje z postępem uczenia. Początkowa wartość to 0,3.

Rozdział 4

Szczegóły implementacji

Wszystkie opisane w poprzednim rozdziale algorytmy i etapy analizy wykorzystuje wykonany program "FaceEmotions". Jest to samodzielna aplikacja, napisana w języku Java. Język ten pozwala na stosunkowo proste utworzenie złożonego systemu o budowie modułowej oraz na w pełni obiektowe podejście do przetwarzania danych. Jednocześnie wbudowana biblioteka *Swing* umożliwia sprawną wizualizację wyników.

W fazie prac nad poszczególnymi algorytmami implementowano je najpierw w środowisku MATLAB. Umożliwiło to szybką kontrolę poprawności ich działania a także łatwą wizualizację otrzymywanych danych i etapów pośrednich przetwarzania. Korzystano przy tym głównie z możliwości przybornika *Image Processing Toolbox*. Dużym ułatwieniem była przy tym swoboda, jaką daje MATLAB przy manipulowaniu macierzami pikseli.

4.1 Budowa programu

Program składa się z szeregu odrębnych klas, wyróżnionych ze względu na pełnione funkcje:

- `FaceEmotions` - główna klasa aplikacji. Odpowiedzialna za tworzenie i kontrolę interfejsu użytkownika (*GUI*).
- `CaptureManager` - klasa odpowiedzialna za akwizycję danych obrazowych z kamery. Wykorzystuje przy tym bibliotekę JMF (*Java Media Framework* [4]).
- `ImageProcessor` - klasa narzędziowa, udostępniająca metody do przeprowadzania wszystkich niezbędnych operacji przetwarzania obrazów. Korzysta przy tym z biblioteki JAI (*Java Advanced Imaging* [3]).

- `FeatureExtractor` - główna klasa odpowiedzialna za przetwarzanie danych obrazowych i ekstrakcję parametrów. Implementuje funkcjonalności służące wyszukiwaniu twarzy oraz regionów oczu i ust.
- `Face` - klasa reprezentująca twarz. Dokonuje przekształceń przestrzeni barwnej, przygotowując obraz obszaru twarzy do dalszej analizy.
- `Eye` - klasa reprezentująca wyodrębniony obszar oka. Udostępnia metody analizujące taki region i zwracające wszystkie niezbędne parametry.
- `Mouth` - klasa o przeznaczeniu podobnym, jak klasa `Eye`, lecz zajmująca się obszarem ust.
- `DecisionModule` - klasa dokonująca interpretacji zebranych parametrów, przechowująca model twarzy neutralnej i porównująca go z aktualnie zarejestrowanymi parametrami.
- `NN` - klasa reprezentująca sieć neuronową oraz jej całą funkcjonalność (uczenie, odczytywanie wyniku klasyfikacji).
- `NNManager` - klasa zarządzająca sieciami neuronowymi. Umożliwia zapis i odczyt z dysku danych uczących dla sieci. Zapisuje i odczytuje również współczynniki nauczonych sieci.
- `FramePanel` - klasa służąca do wyświetlania obrazu wraz z metodami do umieszczania na nim adnotacji, kształtów i wizualizacji danych. Dziedziczy po klasie `DisplayJAI`.

Osobnego wzmiankowania wymagają dwie biblioteki użyte w systemie, bez których jego realizacja nie byłaby możliwa. Są to wspomniane już pakiety *JMF* oraz *JAI*. Obydwa są oficjalnymi produktami firmy SUN, służącymi rozszerzeniu funkcjonalności języka Java o operacje na strumieniowych danych multimedialnych (*JMF*) oraz wszechstronne przetwarzanie i obróbkę obrazów (*JAI*).

Klasa `FaceEmotions` implementuje interfejs `ActionListener` oraz zarządza obiektem typu `Timer`, co pozwala jej na cykliczne wykonywanie odpowiedniego zestawu operacji w zadanym interwale czasowym (250ms).

Poniżej pokrótce opisano cykl działania programu.

Najpierw obiekt klasy `CaptureManager` próbuje przechwycić klatkę obrazu jako obiekt typu `Image`. Jeżeli to się udaje, zostaje ona przekazana do obiektu klasy `FeatureExtractor` celem wykonania wszelkich opisanych wcześniej algorytmów analizy obrazu. Jeżeli jest dostępny model twarzy neutralnej (zainicjalizowano wykrywanie gestów), wówczas wartości odpowiednich parametrów porównywane są w obiekcie klasy `DecisionModule`. Wyniki porównania wypisywane są na ekran wraz z interpretacją (lista zaobserwowanych gestów mimicznych, wykryty

stan emocjonalny). Jednocześnie klasa `NNManager` przekazuje stworzonym przez siebie sieciom neuronowym parametry oraz wyniki ich porównania z odpowiednimi progami. Sieci zwracają informację w postaci odpowiedniego wysterowania swoich wyjść. Wartości te są wypisywane na ekran. Wartość największa jest zaznaczana odrębnym kolorem.

Podczas analizy obrazu przez klasę `FeatureExtractor` na początku wyszukiwany jest obszar głowy (twarzy). Jeżeli takowy zostanie znaleziony, wówczas zostaje utworzony nowy obiekt klasy `Face`, zawierający odpowiedni wycinek obrazu. Dokonywana jest jego konwersja do przestrzeni R-B, a następnie wyszukiwane są regiony oczu. Jeżeli zostaną znalezione, tworzone są obiekty klasy `Eye` odpowiednio dla lewego i prawego oka. Jeżeli znalezione są oba oczy, wówczas tworzony jest także obiekt klasy `Mouth`, zawierający wycinek obrazu zawierający usta.

Po utworzeniu obiektów `Eye` oraz `Mouth` zawierane przezeń obrazy są obracane o odpowiedni kąt a następnie wyliczane są wszystkie opisane wcześniej parametry (odległości punktów charakterystycznych). Podczas wykrywania gestów, klasa `DecisionModule` pobiera te wartości celem porównania z przechowywanym przez siebie modelem. Parametry (początkowe oraz bieżące przechowywane są przez klasę `DecisionModule` w dwóch strukturach grupujących cały zestaw parametrów.

4.2 Opis funkcjonalny programu

Program *FaceEmotions* posiada prosty i czytelny interfejs użytkownika, co znacząco ułatwia jego obsługę. Ogólny widok okna programu przedstawia rysunek 4.1.

Główne menu programu znajduje się z prawej strony okna. Umieszczone są tam przyciski:

- Źródło z kamery - rozpoczęcie akwizycji obrazu z domyślnej kamery podłączonej do komputera.
- Źródło z pliku - wybór i odtworzenie uprzednio zarejestrowanego pliku wideo, z którego ma następować przechwytywanie.
- Start/Stop - Rozpoczęcie/zatrzymanie analizy strumienia wideo.
- Zapisz model - Zapisanie bieżącego układu punktów charakterystycznych twarzy jako neutralnego modelu odniesienia dla wykrywania gestów mimicznych.
- Wyjście - koniec działania programu.

W górnej lewej części okna znajdują się pola, na których wyświetlane są:

Rysunek 4.1: Widok okna programu *FaceEmotions*

- Przechwytywany z kamery obraz, a po rozpoczęciu działania programu - analizowane klatki obrazu. Zaznaczone są tutaj za pomocą ramek wykryte obszary twarzy, oczu i ust. W przypadku spełnienia odpowiedniego warunku w górnej części ramki twarzy pojawia się również nazwa zarejestrowanego stanu emocjonalnego.

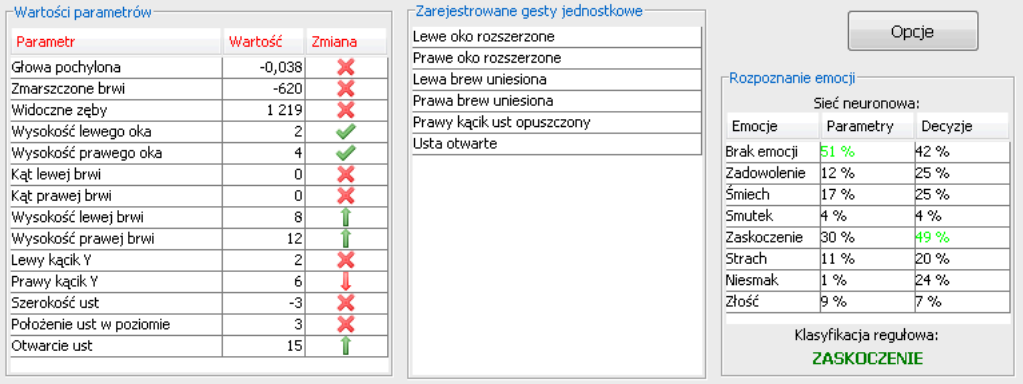
- Znormalizowane do pionu obrazy obszarów oczu i ust z zaznaczonymi wykrytymi punktami charakterystycznymi i łączącymi je liniami.

- Schematyczna wizualizacja analizowanej twarzy - obraz powstały z linii łączących punkty charakterystyczne twarzy.

Dolna lewa część okna programu pełni zamiennie dwie role. Domyślnie prezentowane są tam znormalizowane (wzór (3.24)) wartości różnic poszczególnych parametrów między klatką bieżącą a wzorcową. W przypadku przekroczenia progu zmienności (tabela 3.3) przy danym parametrze wyświetlana jest strzałka (w górę bądź w dół, zależnie od przekroczonego progu).

Po środku wyświetlana jest lista aktualnie rejestrowanych gestów (dla niezerowych wartości cech). Po prawej stronie znajdują się wyniki rozpoznania emocji. Rezultat działania systemu regułowego wyświetlany jest na dole panelu, podczas gdy w jego górnej części znajduje się tabela przedstawiająca procentowe prawdo-

podobieństwa obecności wszystkich gestów. Jest to rezultat działania sieci neuronowych. Przykład z działania programu obrazuje rysunek 4.2.



Rysunek 4.2: Widok listy wartości parametrów, rejestrowanych gestów i emocji

Po kliknięciu przycisku *Opcje*, w dolnej części okna zamiast listy cech zostają wyświetlone suwaki do ustawienia parametrów analizy obrazu (rys. 4.3). Użytkownik może regulować:

- Próg binaryzacji obrazu R-B twarzy dla wykrywania oczu. Umieszczony obok obraz po binaryzacji pozwala dobrać odpowiednią wartość, dla której oczy będą dobrze wyznaczonymi obszarami.

- Parametry filtra barwnego używanego do segmentacji ust (wzór (3.21)) Regulacji podlega zarówno parametr h_0 (odcień, pasmo) jak i w (szerokość filtra). W celu kontroli poprawności dobrania tych parametrów umieszczono z boku obrazy ust: po użyciu filtra barwnego, oraz ten sam obraz po binaryzacji.

Użytkowanie programu przebiega następująco:

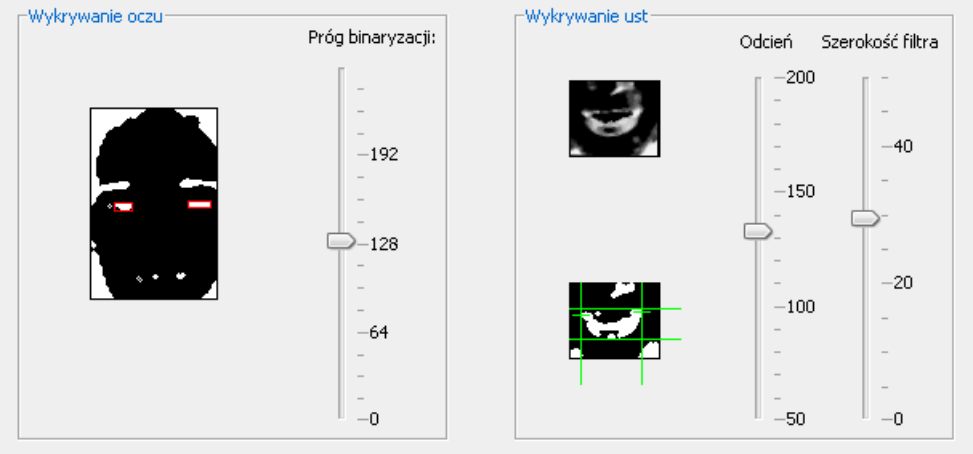
1. Uruchomienie programu.

2. Wybór źródła obrazu wideo.

   a) W przypadku wybrania "Z kamery" należy poczekać chwilę (często dłuższą), aż obraz pojawi się w oknie. Zwłoka ta spowodowana jest działaniem biblioteki *JMF*.

   b) W przypadku kliknięcia "Z pliku" otworzy się wyboru plików w którym należy wskazać wybrany klip wideo.

      System obsługuje dwa formaty: `.mov` oraz `.avi`, przy czym w przypadku tego ostatniego ważna jest zastosowana metoda kompresji. Testy potwierdziły poprawną obsługę jedynie kodeka *Intel IYUV*. Najlepiej



Rysunek 4.3: Widok panelu ustawień parametrów

jednak używać filmów nieskompresowanych. Większość plików `.avi` generowana przez aparaty cyfrowe jest obsługiwana przez system.

3. Po rozpoczęciu rejestracji obrazu przez kamerę należy znów poczekać chwilę, aż jej wewnętrzna automatyka dostosuje parametry rejestracji do aktualnych warunków oświetleniowych. Fakt takiego zachowania kamer wideo powoduje duże zmiany w jakości i kolorystyce obrazów rejestrowanych w różnych warunkach, stąd też konieczna stała się możliwość ręcznego sterowania parametrami przetwarzania.

4. Rozpoczęcie analizy obrazu przyciskiem *Start*. Na tym etapie należy dostroić działanie programu opisanymi regulatorami, ażeby punkty charakterystyczne były znajdowane możliwie najtrafniej.

5. Wciśnięcie przycisku *Zapisz model* powoduje zapisanie bieżących parametrów twarzy jako wzorcowych do porównań. Dlatego też użytkownik powinien przyjąć w tym momencie możliwie neutralny wyraz twarzy.

6. Od tego momentu analizowana jest mimika twarzy w porównaniu z twarzą "wzorcową". Wypisywane są różnice odpowiednich parametrów, zarejestrowane gesty i emocje.

7. Przetwarzania można przerwać w każdym momencie przyciskiem *Stop*, co pozwala na dokładne przyjrzenie się ostatnio zarejestrowanym parametrom i wyrazowi twarzy, który opisują.

Rozdział 5

Wyniki działania i testy

W pierwszej części niniejszego rozdziału zilustrowano przykładami działanie mechanizmu klasyfikacji gestów oraz emocji.

Druga część przedstawia wyniki przeprowadzonych testów systemu. Oddzielnie opisano poszczególne etapy jego działania, co pozwoliło na indywidualną ocenę każdego z zastosowanych algorytmów.

5.1 Przykłady działania klasyfikatora gestów

Poniżej zamieszczono przykłady działania systemu, ukazujące klasyfikowane gesty mimiczne.

Tabela 5.1 przedstawia przykładową twarz o neutralnym wyrazie, będącą wzorem do określania gestów zamieszczonych w tabeli 5.2

Tabela 5.1: Przykład twarzy neutralnej

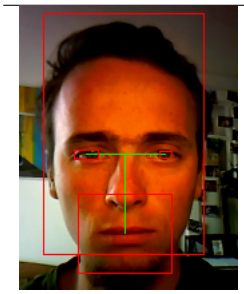
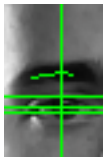
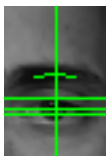
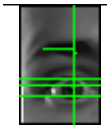
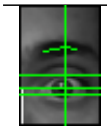
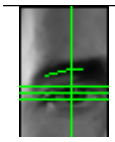
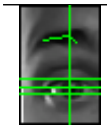
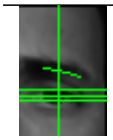
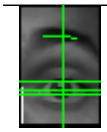
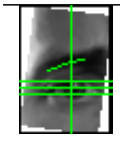
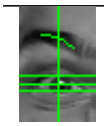
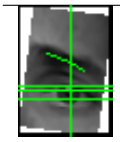
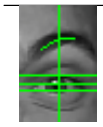
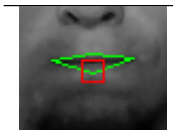
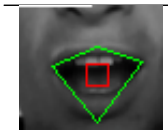
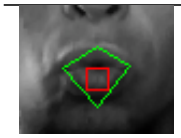
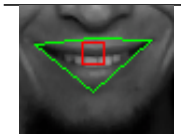
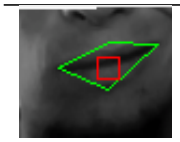
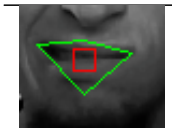
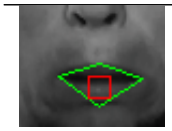
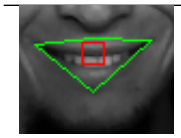
Cała twarz	Lewe oko	Prawe oko	Usta
			

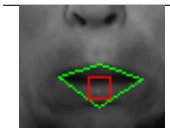
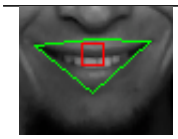
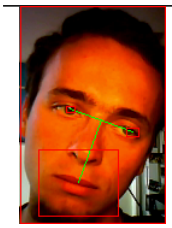
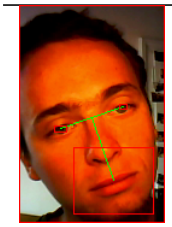
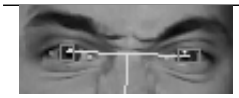
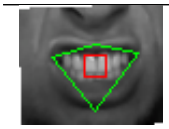
Tabela 5.2 zawiera przykłady zarejestrowania wszystkich dostępnych gestów. Dla każdego przykładu podano wartość różnicy pomiędzy odpowiednim parame-

trem a wartością bazową. Na podstawie tej różnicy system zadecydował o odnotowania danego gestu. Jak łatwo sprawdzić, wartości te przekraczają odpowiednie progi z tabeli 3.3.

Tabela 5.2: Przykłady rozpoznawanych przez system gestów

Obraz	**Nazwa gestu**	**Cecha i wartość**	**Różnica parametrów**
	Lewe oko otwarte szeroko	f1 = 1	5
	Prawe oko otwarte szeroko	f2 = 1	6
	Lewa brew opuszczona	f3 = -1	-3
	Lewa brew uniesiona	f3 = 1	20
	Prawa brew opuszczona	f4 = -1	-4
	Prawa brew uniesiona	f4 = 1	20
	Wewnętrzny koniec lewej brwi uniesiony	f5 = -1	-7
	Zewnętrzny koniec lewej brwi uniesiony	f5 = 1	13

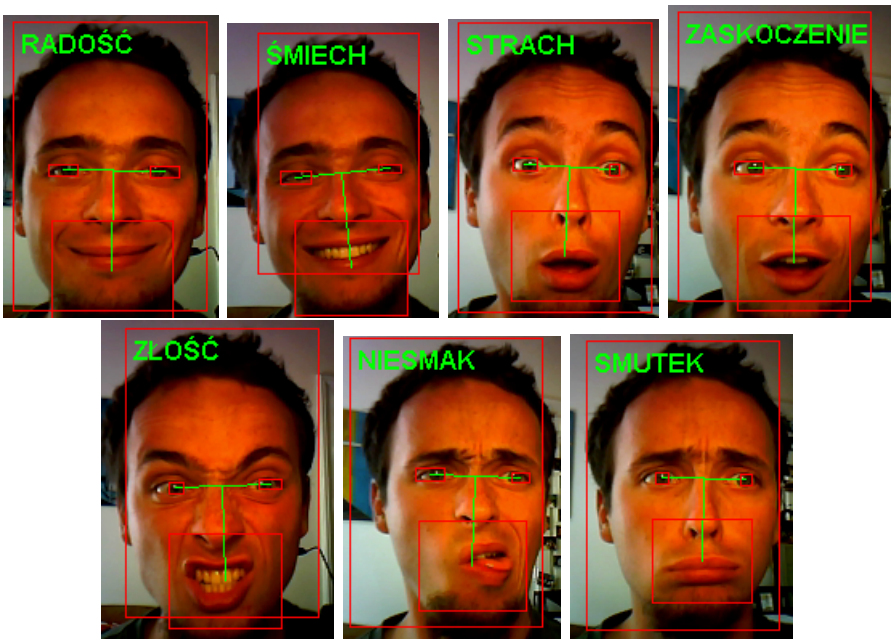
Obraz	**Nazwa gestu**	**Cecha i wartość**	**Różnica parametrów**
	Wewnętrzny koniec prawej brwi uniesiony	f6 = -1	-11
	Zewnętrzny koniec prawej brwi uniesiony	f6 = 1	12
	Usta zaciśnięte	f7 = -1	-13
	Usta otwarte	f7 = 1	30
	Usta wąskie	f8 = -1	-18
	Usta szerokie	f8 = 1	20
	Usta przesunięte w prawo	f9 = -1	19
	Usta przesunięte w lewo	f9 = 1	-11
	Lewy kącik opuszczony	f10 = -1	11
	Lewy kącik uniesiony	f10 = 1	-17

Obraz	Nazwa gestu	Cecha i wartość	Różnica parametrów
	Prawy kącik opuszczony	f11 = -1	15
	Prawy kącik uniesiony	f11 = 1	-18
	Głowa pochylona w prawo	f12 = -1	-0,35
	Głowa pochylona w lewo	f12 = 1	0,333
	Zmarszczki nad nosem	f13 = 1	5510
	Zęby widoczne	f14 = 1	7563

5.2 Testy

Poniżej zamieszczono obserwacje z przeprowadzonych testów działania systemu. Z uwagi na jego budowę zdecydowano się na krótkie oddzielne omówienie każdego stadium przetwarzania. Dzięki temu można lepiej ocenić jakość działania poszczególnych zastosowanych algorytmów. Przy każdym z nich wskazano zauważone sytuacje, w których dany algorytm może nie działać poprawnie. Przykłady poprawnego działania opisane są w rozdziale 5.1.

W dalszej części rozdziału umieszczono jego zasadniczą część - wyniki testów przeprowadzonych dla uprzednio zarejestrowanych sekwencji wideo.



Rysunek 5.1: Przykłady rozpoznawanych przez system emocji

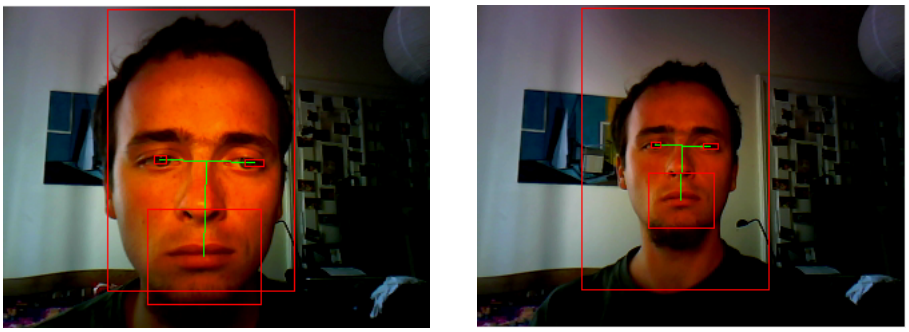
Jednocześnie trudno w tym miejscu podać porównawcze dane odnośnie skuteczności systemu. Wynika to z faktu, iż system pracuje na danych dynamicznie dostarczanych z kamery (lub pliku wideo), nie zaś na statycznych zdjęciach. Twórcy innych opisywanych rozwiązań podając skuteczność swoich systemów opierali się na własnych sekwencjach testowych, być może specjalnie przygotowywanych do współpracy tymi programami. W celu przeprowadzenia obiektywnych porównań należałoby na wejście omawianego systemu podać te same pliki testowe, co z uwagi na ich niedostępność jest niemożliwe.

5.2.1 Śledzenie twarzy

Zastosowany algorytm w znakomitej większości przypadków działa poprawnie, dobrze określając obszar twarzy. Pomocne są przy tym ścisłe ograniczenia, nałożone na warunki pracy systemu (patrz strona 28).

Jedynym zaobserwowanym mankamentem algorytmu jest niewrażliwość na oddalanie głowy od kamery przy zachowaniu stałego jej położenia w płaszczyźnie obiektywu. Podobnie dzieje się przy wykonywaniu bardzo powolnych ruchów głową w dowolnym kierunku. Zobrazowane jest to na rysunku 5.2.

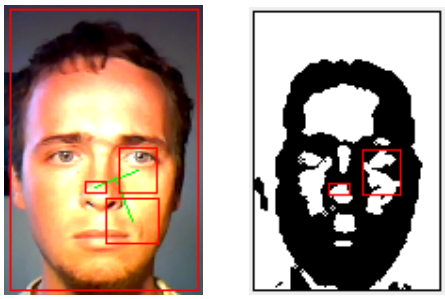
Przyczyną jest działanie mechanizmu, przechowującego poprzednie położenie głowy w przypadku zarejestrowania znikomego tylko ruchu (wzór (3.8)).



Rysunek 5.2: Przykład poprawnego określenia położenia twarzy oraz rezultat powolnego jej oddalania od kamery

5.2.2 Wykrywanie regionów oczu

Algorytm wykrywania oczu oparty o analizę obrazu R-B (str. 33) pozwala na szybką i prostą detekcję oczu. Przy dobrych warunkach akwizycji i po regulacji parametrów trafnie określa położenie oczu. Jednakże jest on dość wrażliwy na warunki oświetleniowe. W przypadku zbyt intensywnego oświetlenia twarzy, na obrazie R-B pojawia się dużo artefaktów w miejscach prześwietlonych (rys. 5.3). Nie pozwalają one na poprawne wykrycie oczu.



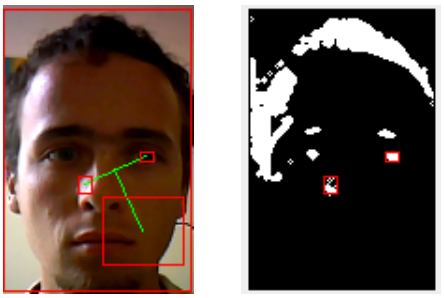
Rysunek 5.3: Przykład błędnej lokalizacji oczu w przypadku prześwietlenia obrazu

Równie niekorzystne warunki sprawia nierównomierne oświetlenie twarzy i pojawianie się cieni i refleksów na nosie (rys. 5.4). Wówczas najczęściej taki obszar mylnie interpretowany jest jako jedno z oczu.

5.2.3 Analiza oczu i brwi

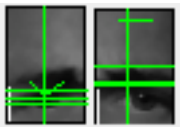
Zaproponowana metoda określania kształtu brwi pomimo swojej prostoty daje w większości przypadków dobre rezultaty. Poprawnie określa położenie trzech punktów węzłowych brwi.

Problemy występują w określonych przypadkach. Jednym z nich jest bardzo niskie opuszczenie brwi, tak, że w profilu jasności obszaru oka (rys. 3.12) dwa



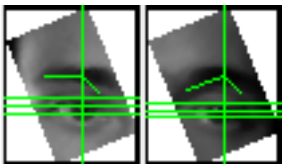
Rysunek 5.4: Przykład błędnej lokalizacji oczu w przypadku bocznego oświetlenia

minima, odpowiadające oku i brwi zlewają się w jedno. Różne zachowania systemu w takim przypadku ilustruje rys. 5.5.



Rysunek 5.5: Przykład błędnej lokalizacji poziomów oka i brwi

Innym przypadkiem, w którym ten mechanizm nie zawsze działa poprawnie jest znaczne pochylenie głowy w bok (rys. 5.6). Wówczas prostokątny obszar wokół oka może nie objąć całej brwi, co wprowadza błędy w określaniu jej kształtu.



Rysunek 5.6: Przykład błędnego rozpoznania kształtu brwi przy mocnym pochyleniu głowy

Dodatkowo, przeprowadzone testy wykazały małą skuteczność w wykrywaniu stopnia otwarcia oczu. Wpływ na taką sytuację ma mała zmienność tego parametru (rzędu 2 - 3 pikseli), co sprawia, że analiza ta staje się podatna na liczne chwilowe zakłócenia. Być może zwiększenie rozdzielczości analizowanego obrazu przyniosłoby poprawę rezultatów.

5.2.4 Analiza ust

Metoda segmentacji ust za pomocą filtra barwnego pozwala dobrze i w prosty sposób określić ich kształt. Jednocześnie jest to metoda dość wrażliwa na warunki akwizycji obrazu, podobnie jak algorytm wykrywania oczu. Mimo to, przy

dobrych warunkach oświetleniowych filtr taki skutecznie wyznacza usta na tle reszty twarzy.

Przy zbyt ciemnym, bądź prześwietlonym obrazie rejestrowany kolor ust (warg) nie odróżnia się znacząco od sąsiadującej z nimi skóry. Możliwe jest też, iż ciemniejszy obszar pod nosem bądź brodą przyjmą jednakowy odcień jak wargi. Sytuację taką przedstawia rys. 5.7.

Rysunek 5.7: Przykład błędnego rozpoznania kształtu ust

Użyta przy tworzeniu i testowaniu systemu prosta kamera internetowa charakteryzuje się dużą zmiennością w prezentowaniu poszczególnych barw. Kamery tego typu, z uwagi na bardzo małe obiektywy kompensują niedobór światła dynamicznym regulowaniem czułości matrycy. Powoduje to ciągłe zmiany w spektrum barwnym rejestrowanego obrazu. Dlatego też określenie parametrów filtra (wzór (3.21)) na wstępie może okazać się niewystarczające (rys. 5.8)i konieczna staje się wówczas dalsza regulacja w trakcie działania programu.

Rysunek 5.8: Przykład błędnego działania filtra barwnego po zmianie oświetlenia

5.2.5 Klasyfikacja gestów

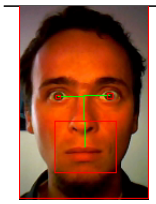
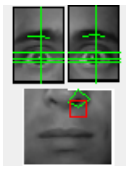
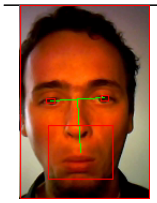
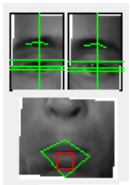
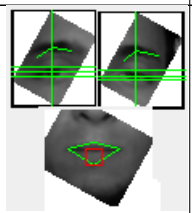
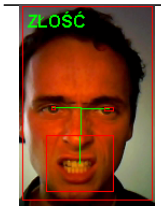
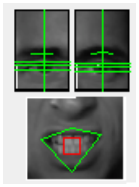
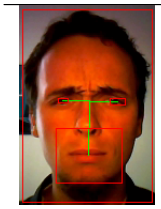
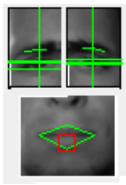
Do celów testowania systemu zarejestrowano przykładowe sekwencje wideo. Na każdej z nich użytkownik starał się przedstawić całą paletę rozpoznawanych gestów. Następnie pliki te podano na wejście systemu, ustawiono omówione wcześniej parametry algorytmów i uruchomiono mechanizm detekcji mimiki.

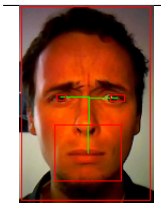
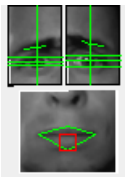
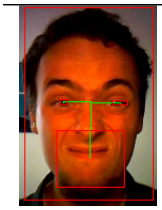
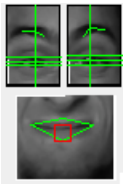
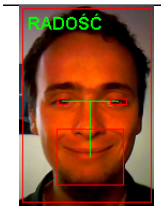
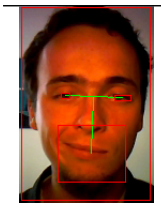
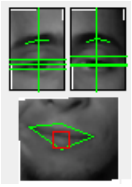
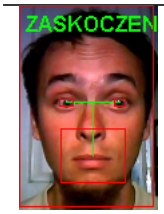
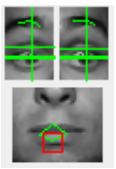
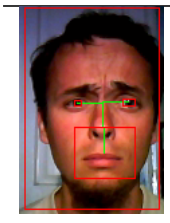
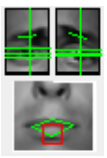
Podczas działania programu wybrano przykładowe klatki obrazu, na których prezentowane są różne gesty (bądź ich zestawy). Jednocześnie zapisano, które z gestów zostały wykryte, które przeoczone, a które zostały fałszywie rozpoznane (wykryte pomimo ich braku). Niektóre z tych klatek przedstawia tabela 5.3.

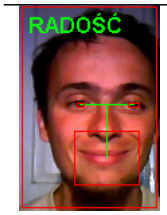
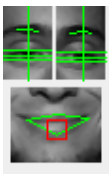
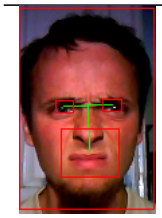
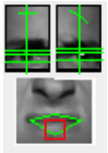
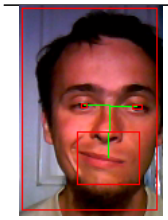
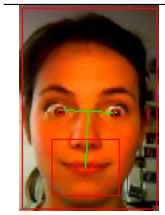
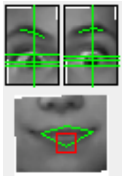
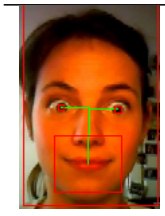
Należy przy tym zauważyć, iż dane te, pomimo zastosowania nagrań zamiast kamery nie są deterministyczne. Przy każdym badaniu ręcznie ustalano parametry działania systemu oraz wskazywano moment prezentowania neutralnego wyrazu twarzy. Dlatego też uzyskane wyniki należy traktować orientacyjnie. Najczęściej

wpływ na permanentnie złe rozpoznawanie danego gestu ma chwilowy błąd w detekcji jakiegoś punktu w momencie ustalania neutralnego wyglądu twarzy.

Tabela 5.3: Wyniki klasyfikacji gestów dla przykładowych twarzy

Twarz	**Rejony analizy**	**Gesty**		
		Zarejestrowane poprawnie	**Zarejestrowane niepoprawnie**	**Pominięte**
			+f5, -f8, -f9, +f10, +f11	+f1, +f2
		-f8, -f10, -f11	+f5, +f7	
		+f12	-f5, +f6	
		-f3, -f4, +f7, +f14	+f5	+f13
		-f3, -f4, +f13		-f5, -f6

Twarz	Rejony analizy	Gesty		
		Zarejestrowane poprawnie	**Zarejestrowane niepoprawnie**	**Pominięte**
		-f3, -f4, -f6, +f13	+f2	-f5
		+f3, +f4, +f5, +f6, +f10, +f11		
		+f8, +f10, +f11	+f5	
		+f9, +f10		
		+f1, +f3, +f4, -f6	-f8, +f9	+f2
		-f5, -f6		

Twarz	Rejony analizy	Gesty		
		Zarejestrowane poprawnie	**Zarejestrowane niepoprawnie**	**Pominięte**
		+f2, +f3, +f4, +f5, +f6, +f10, +f11, +f13	-f8, +f9, +f14	
		+f3, +f4, +f8, +f10, +f11		
		+f13	+f1, +f2, +f3, +f4, -f5, -f6, +f8	+f10, +f11
		+f8, +f9, +f10	+f3, -f5, -f6	
		+f3, +f4, +f5, +f6	-f9, -f10	+f1, +f2
		+f1, +f2, +f3, +f4, +f5, +f6, +f8	+f9	

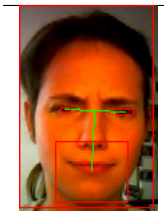
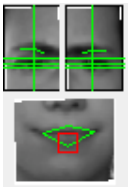
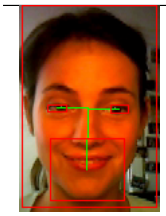
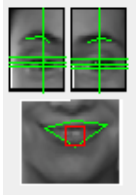
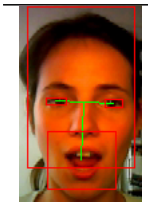
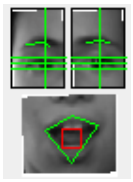
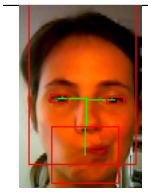
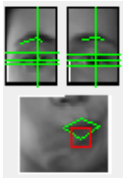
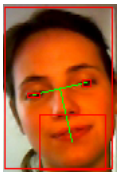
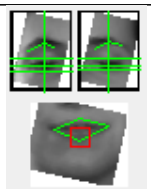
Twarz	Rejony analizy	Gesty		
		Zarejestrowane poprawnie	**Zarejestrowane niepoprawnie**	**Pominięte**
		-f3, -f4, +f5, +f6, +f13		
		+f3, +f5, +f8, +f11		+f10
		+f7	+f5, -f10	
		+f8, +f10, +f11	+f5	
		-f8, -f9	+f1, +f2	
		+f12	+f1, +f2, +f6, -f9, +f11	

Tabela 5.4: Jakość rozpoznawania poszczególnych gestów. Kolumna "Zarejestrowane poprawnie" zawiera procentową skuteczność wykrycia danego gestu, zaś kolumna "Zarejestrowane niepoprawnie" przedstawia procentowy udział fałszywych rozpoznań danego gestu w ogólnej liczbie ich wykryć.

Nazwa gestu	**Zarejestrowane poprawnie**	**Zarejestrowane niepoprawnie**
Lewe oko otwarte szeroko	50 %	50 %
Prawe oko otwarte szeroko	57 %	60 %
Lewa brew opuszczona	100 %	14 %
Lewa brew uniesiona	100 %	25 %
Prawa brew opuszczona	100 %	0 %
Prawa brew uniesiona	100 %	33 %
Wewnętrzny koniec lewej brwi uniesiony	50 %	81 %
Zewnętrzny koniec lewej brwi uniesiony	100 %	62 %
Wewnętrzny koniec prawej brwi uniesiony	60 %	57 %
Zewnętrzny koniec prawej brwi uniesiony	100 %	50 %
Usta zaciśnięte	100 %	66 %
Usta otwarte	100 %	42 %
Usta wąskie	100 %	54 %

Nazwa gestu	**Zarejestrowane poprawnie**	**Zarejestrowane niepoprawnie**
Usta szerokie	88 %	11 %
Usta przesunięte w prawo	100 %	62 %
Usta przesunięte w lewo	100 %	77 %
Lewy kącik opuszczony	100 %	50 %
Lewy kącik uniesiony	81 %	25 %
Prawy kącik opuszczony	100 %	0 %
Prawy kącik uniesiony	90 %	37 %
Głowa pochylona w prawo	100 %	0 %
Głowa pochylona w lewo	100 %	25 %
Zmarszczki nad nosem	87 %	0 %
Zęby widoczne	100 %	50 %
Średnio	**90,1 %**	**38,8 %**

Tabela 5.4 zawiera podsumowanie danych zaprezentowanych w tabeli 5.3. Jak widać jakość rozpoznawania większości gestów jest bardzo dobra, co potwierdza trafność doboru progów ich detekcji. Można jednak zauważyć słabe wyniki w przypadku analizy stopnia otwarcia oka. Powodem jest zbyt mało precyzyjne działanie algorytmu wykrywania położenia powiek, co przy bardzo małych zmianach wartości parametru (kilka pikseli) powoduje niską skuteczność takiego rozwiązania. Podobne wyniki niesie wykrywanie uniesienia wnętrza brwi. Jednakże w tym miejscu należy przyjrzeć się ilości fałszywych rozpoznań. Jest ona dość duża dla tych gestów. Sugeruje to, iż gest jest często wykrywany, lecz nie zawsze wtedy, kiedy rzeczywiście zaistniał. Przyczyną może być źle dobrany próg detektora, bądź wpływ pochylenia głowy w przód lub do tyłu. Takie przechylenia

powodują pozorną zmianę kształtu obserwowanych brwi.

Testy wykazały, iż dość duża ilość gestów jest rejestrowana fałszywie, czyli przy braku ich obecności. Wpływ na to mają zarówno wymienione już czynniki, jak też wspomniane zmiany warunków akwizycji. Przykładowo - zbyt częsta detekcja faktu zaciśnięcia ust może wynikać ze zmiany w czasie ich rejestrowanej barwy, co przy niezmiennych parametrach filtra barwnego prowadzi do zmniejszenia ich wykrywanej wielkości. Adaptacyjna regulacja parametrów analizy twarzy rozwiązałaby ten problem.

Fakt wykrywania pewnej liczby zbędnych gestów nie pozostaje bez wpływu na działanie algorytmów wykrywania emocji na twarzy. Może w szczególności zakłócać pracę klasyfikatorów opartych o sieci neuronowe, dostarczając zaburzonych danych wejściowych.

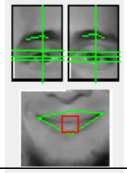
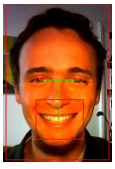
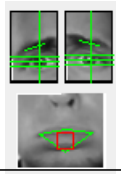
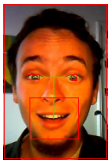
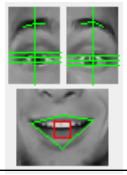
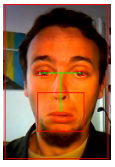
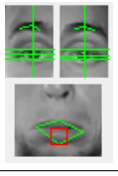
5.2.6 Klasyfikacja emocji

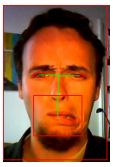
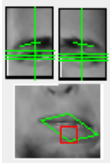
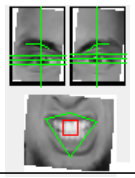
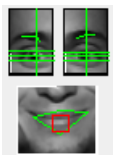
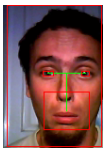
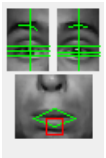
Mechanizm rozpoznawania emocji opiera się na omówionych już klasyfikatorach - regułowym oraz wykorzystującym sieci neuronowe. Skuteczność systemu regułowego, jak i drugiej z sieci w głównej mierze zależy od trafności rozpoznawania gestów jednostkowych. Pierwsza z sieci nie bazuje na rozpoznanych gestach, lecz na bezwzględnych wartościach parametrów.

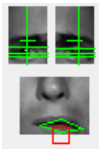
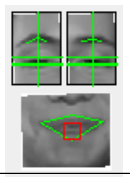
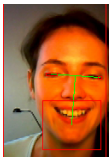
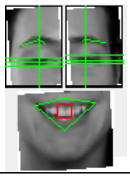
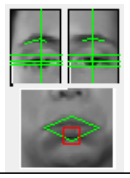
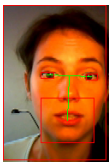
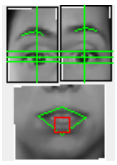
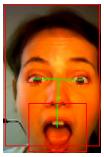
Użyty zestaw reguł wnioskowania (tabela 3.4) powstał na podstawie empirycznych obserwacji gestów typowych dla każdego stanu emocjonalnego. Testy ukazują trafność takiego ich doboru, jednakże należy pamiętać, że niektórzy użytkownicy mogą inaczej wyrażać swoje emocje. Dodatkowym utrudnieniem jest konieczność dość ekspresyjnego i "teatralnego" wyrażania mimiki, aby zwiększyć pewność zarejestrowania kluczowych gestów.

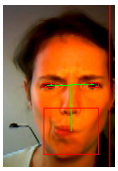
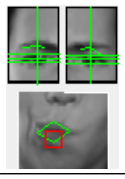
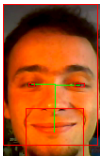
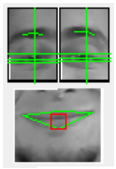
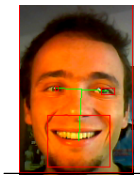
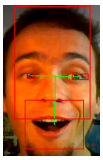
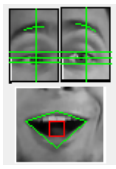
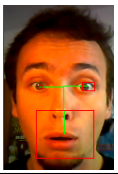
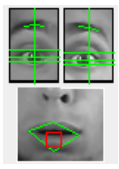
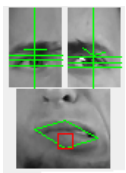
Tabela 5.5 przedstawia wyniki testów rozpoznawania dla przykładowych twarzy. Są to losowo wybrane klatki z trzech filmów testowych oraz pochodzące z kamery. Użytkownicy starali się wyrażać emocje z zestawu rozpoznawanego przez system. Przedstawiono wyniki działania wszystkich trzech klasyfikatorów: regułowego, sieci neuronowej badającej wartości parametrów oraz sieci badającej wartości cech (obecność bądź nie danych gestów jednostkowych).

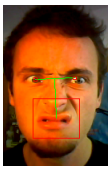
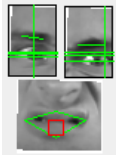
Tabela 5.5: Wyniki klasyfikacji emocji dla testowych twarzy. Dla każdej twarzy podano procentowe prawdopodobieństwo wykrycia gestów przez sieci neuronowe: operującą na parametrach (wyżej) oraz na gestach jednoskowych (nizej).

Twarz	**Rejony analizy**	**Wyrażane emocje**	**Klasyfikacja**							
			Regułowa	**Sieci neuronowe**						
				Zadowolenie	**Śmiech**	**Smutek**	**Zaskoczenie**	**Strach**	**Niesmak**	**Złość**
		Zadowolenie	Zadowolenie	28	56	9	26	9	9	9
				95	24	12	13	4	18	1
		Śmiech	Śmiech	28	56	9	26	9	9	9
				9	89	10	9	9	10	9
		Smutek	Brak	27	11	20	3	4	27	26
				77	24	89	22	13	47	4
		Zaskoczenie	Zaskoczenie	6	9	26	51	81	1	3
				18	14	10	90	8	7	13
		Strach	Brak	14	10	24	12	89	17	0
				11	9	11	14	83	8	6

Twarz	Rejony analizy	Wyrażane emocje	Klasyfikacja							
			Regułowa	Sieci neuronowe						
				Zadowolenie	Śmiech	Smutek	Zaskoczenie	Strach	Niesmak	Złość
		Niesmak	Niesmak	27	11	20	3	4	27	26
				13	12	66	13	9	81	6
		Złość	Złość	11	12	19	10	0	13	88
				5	13	9	20	11	13	89
		Zadowolenie	Brak	19	72	40	61	68	33	53
				30	43	32	38	59	58	44
		Śmiech	Śmiech	35	48	2	23	9	2	10
				62	61	7	18	1	47	2
		Smutek	Smutek	42	49	25	19	76	59	50
				53	65	47	57	68	66	44
		Zaskoczenie	Zaskoczenie	33	55	26	22	78	48	61
				42	57	55	37	60	54	37
		Strach	Brak	33	55	26	22	78	48	61
				44	57	54	35	60	56	42

Twarz	Rejony analizy	Wyrażane emocje	Klasyfikacja							
			Regułowa	Sieci neuronowe						
				Zadowolenie	Śmiech	Smutek	Zaskoczenie	Strach	Niesmak	Złość
		Niesmak	Niesmak	2	1	44	6	9	50	10
				6	16	4	3	36	70	12
		Złość	Brak	10	10	12	11	0	13	89
				30	38	5	9	1	42	12
		Zadowolenie	Brak	61	60	63	42	55	58	35
				79	60	65	49	58	42	61
		Śmiech	Śmiech	35	48	2	23	9	2	10
				9	77	7	11	3	35	8
		Smutek	Brak	50	31	54	26	65	47	50
				46	41	67	62	66	38	45
		Zaskoczenie	Zaskoczenie	12	17	4	30	11	1	9
				25	25	4	49	20	24	7
		Strach	Strach	47	35	24	33	29	57	59
				47	40	72	68	64	36	41

Twarz	Rejony analizy	Wyrażane emocje	Klasyfikacja								
			Regułowa	Sieci neuronowe							
				Zadowolenie	Śmiech	Smutek	Zaskoczenie	Strach	Niesmak	Złość
		Niesmak	Smutek	34	61	29	34	51	64	54
				19	32	41	54	58	67	58
		Zadowolenie	Zadowolenie	28	36	7	18	8	8	30
				90	23	2	1	12	14	11
		Śmiech	Śmiech	28	36	7	18	8	8	30
				8	90	11	6	11	8	15
		Smutek	Brak	11	9	26	37	41	11	0
				3	24	98	6	13	5	5
		Zaskoczenie	Zaskoczenie	45	62	43	13	21	48	69
				58	42	66	52	64	31	36
		Strach	Strach	4	2	31	9	35	32	6
				9	9	10	8	90	10	10
		Niesmak	Niesmak	9	9	10	9	9	89	10
				17	12	4	11	18	60	8

Twarz	Rejony analizy	Wyrażane emocje	Klasyfikacja							
			Regułowa	Sieci neuronowe						
				Zadowolenie	Śmiech	Smutek	Zaskoczenie	Strach	Niesmak	Złość
		Złość	Brak	9	9	8	10	10	10	89
				1	12	4	0	37	39	54

Analizując uzyskane dane, można zauważyć różnice w działaniu poszczególnych klasyfikatorów. Przykładowo: Zaskoczenie było bezbłędnie rozpoznawane przez system regułowy, podczas gdy sieci neuronowe nie dawały zadowalających wyników dla tego gestu. Z kolei sieci często myliły śmiech z zadowoleniem, co było zachowaniem do przewidzenia z uwagi na podobieństwo tych gestów.

Podsumowanie wyników testów przedstawia tabela 5.5. Zawiera ona zestawienie procentowej jakości rozpoznawania danych gestów przez poszczególne klasyfikatory. Podsumowano w niej również średni wynik każdego z klasyfikatorów jak i średnią rozpoznawalność danego gestu.

Jak widać, najlepsze rezultaty osiągnęła sieć neuronowa badająca obecność poszczególnych gestów jednostkowych, wykrytych wcześniej za pomocą progowania parametrów ustalonymi wartościami. Sieć operująca na bezwzględnych wartościach parametrów nie uzyskała zadowalających rezultatów.

Wśród rozpoznawanych emocji najbardziej charakterystyczne i łatwe do identyfikacji okazały się śmiech, niesmak i złość. Wszystkie one są opisywane przez bardzo powtarzalne zestawy gestów lub parametrów. Dodatkowo, sieci neuronowe tak dobrze nauczyły się rozpoznawać śmiech, iż często podobne emocje zadowolenia były mylnie interpretowane właśnie jako śmiech.

Należy przy tym zauważyć, iż nie przeprowadzono optymalizacji struktury wewnętrznej sieci, jej parametrów oraz procesu uczenia. Wszystkie te ustawienia zostały zadane z góry, przy czym przy ich ustalaniu oparto się na wcześniejszych doświadczeniach autora z uczeniem takich sieci. Optymalizacja takich sieci jest zadaniem skomplikowanym, wymagającym bardzo wielu prób i testów, co nie

Tabela 5.6: Jakość rozpoznawania poszczególnych emocji przez klasyfikatory (na podstawie wyników z tabeli 5.5)

Emocje	**Klasyfikator regułowy**	**Sieć badająca parametry**	**Sieć badająca cechy**	**Średnio**
Zadowolenie	50 %	0 %	75 %	**41,7 %**
Śmiech	100 %	100 %	75 %	**91,7 %**
Smutek	25 %	0 %	50 %	**25 %**
Zaskoczenie	100 %	35 %	50 %	**61,7 %**
Strach	50 %	75 %	75 %	**66,7 %**
Niesmak	75 %	100 %	100 %	**91,7 %**
Złość	50 %	100 %	66 %	**72 %**
Średnio	**64,3 %**	**58,6 %**	**70,1 %**	

jest zasadniczym tematem niniejszej pracy. Dlatego też zdecydowano się na dość proste zastosowanie wspomnianych sieci.

Mimo to na poszczególnych przykładach można dostrzec, jak sieci nauczyły się rozpoznawać i generalizować nauczone zbiory wejściowe. Widać to najlepiej na tych przykładach, gdzie oczekiwane wyjście przyjęło wartości zbliżone do 90, zaś pozostałe - około 10. Takie silne wyróżnienie jednego z wyjść świadczy o dobrym rozpoznaniu danego wektora wejściowego, gdyż sieć jest "pewna" jego klasyfikacji.

Na uzyskaną jakość rozpoznawania emocji za pomocą sieci neuronowych mają wpływ także:

- Dobór i ilość twarzy w zbiorze uczącym. Im większa ich ilość i większe zróżnicowanie, tym teoretycznie lepsza zdolność sieci do generalizacji i określania prawdopodobieństwa rozpoznania.

- Jakość wykrywania parametrów zarówno w zbiorze uczącym jak i dla twarzy testowych. Powtarzający się błąd w zbiorze uczącym może pogorszyć jakość rozpoznawania lub wręcz je uniemożliwić.

- Zakresy zmienności poszczególnych badanych parametrów. Parametr o dużej zmienności dla wszystkich twarzy reprezentujących jedną emocję nie nadaje się dobrze do jej sklasyfikowania. Sieć w takich wypadkach powinna zmniejszyć jego udział w rozpoznawaniu i musi się opierać na mniejszej ich liczbie.

Rozdział 6

Podsumowanie

W niniejszej pracy starano się przedstawić zagadnienie automatycznego rozpoznawania i analizy gestów mimicznych wyrażanych przez użytkownika komputera.

Ponieważ problematykę taką można traktować jako eksplorację nowego kanału komunikacji na linii człowiek - komputer, starano się na samym początku przedstawić zagadnienie na gruncie antropologicznym i psychologicznym. Poruszono zagadnienia roli gestów mimicznych w komunikacji, ich rozpoznawania przez ludzi i interpretacji. Powoływano się przy tym na wyniki prowadzonych od dawna badań psychologicznych.

Dalsza część pracy stara się opisać ogólny mechanizm automatycznego rozpoznawania mimiki przez maszynę. Wprowadzono przy tym podział na konkretne, kluczowe stadia takiego przetwarzania, celem wprowadzenia do tematyki i ukazania możliwych rozwiązań lub też potencjalnych trudności.

Następnie opisano dwa duże, działające systemy służące do realizacji opisywanych zadań. Celowo wybrano rozwiązania, korzystające z krańcowo różnego podejścia do tematu analizy obrazu twarzy. Skonfrontowano dokładne podejście analityczne, typowe dla przetwarzania komputerowego z systemem wzorowanym na ludzkim sposobie postrzegania obiektów (w tym twarzy). Obydwa omówione rozwiązania prezentują obiecujące wyniki. Wskazuje to na wielość dróg, którymi może podążać rozwój tego typu systemów, co tym bardziej zachęciło do przeprowadzenia własnych prób i szukania obiecujących metod.

Główną częścią niniejszej pracy było zbudowania działającego systemu rozpoznającego gesty mimiczne. Cel ten został zrealizowany a rezultaty opisane. Przedstawiono ogólny zarys zastosowanej metodologii. Skrystalizowała się ona dopiero podczas prac, po wybraniu najbardziej obiecujących a zarazem możliwie prostych algorytmów. Jednakże dzięki temu zostały wypróbowane różne metody i mechanizmy, co pomimo ich odrzucenia poszerzyło wiedzę autora z zakresu omawianej i pokrewnej tematyki.

Przy wyborze algorytmów przetwarzania obrazu kierowano się w równej mierze jakością ich działania jak i łatwością implementacji i szybkością działania. Przykładowo, zastosowany mechanizm wykrywania położenia głowy w omawianym zastosowaniu sprawdza się doskonale, jednocześnie pracując szybko i wydajnie.

Wykrywanie oczu i ust przeprowadzono z wykorzystaniem segmentacji obrazów w zmienionej przestrzeni barwnej. Opierając się na cytowanych pracach oraz prowadząc własne testy, zdecydowano się na zastosowanie obrazu różnic kanałów czerwonego i niebieskiego dla wykrywania oczu. Okazało się, iż rzeczywiście, na takim obrazie oczy zawsze odróżniają się kolorem od reszty skóry. W przypadku ust zastosowano bardziej wyszukaną filtrację barwną, opisywaną w literaturze. Dzięki temu uzyskano zadowalające rezultaty detekcji warg. Dużym mankamentem zaimplementowanych metod jest ich brak adaptacji do zmiennych warunków rejestracji obrazu. Dlatego też zdecydowano się na ręczną korekcję kluczowych parametrów algorytmów. Dzięki temu jednak użytkownik ma pewność o optymalnym ustawieniu systemu do aktualnych warunków rejestracji.

Użycie filtrów barwnych do określania położenia oczu i warg (kształtu ust) jest szeroko opisywane w literaturze. Zazwyczaj autorzy wskazują je jako najprostsze a zarazem najskuteczniejsze metody. Przeprowadzone badania, jak i wykonana implementacja tylko po części potwierdzają te wyniki. Okazuje się iż kluczową rolę odgrywają warunki akwizycji - jakość kamery, charakterystyka barwna jej obrazu oraz oświetlenie. Omawiane prace prowadzone były zazwyczaj w warunkach laboratoryjnych, na dobrym sprzęcie i dostosowanych do niego warunkach oświetleniowych. Opisywany system projektowany był z myślą o współpracy ze zwykłą kamerą internetową w domowo-biurowym otoczeniu. Dlatego też segmentacja barwna dała dobre wyniki jedynie przy zapewnieniu względnie optymalnych warunków. Być może zastosowanie bardziej "inteligentnych" algorytmów pozwoliłoby uwolnić się od większości zakłóceń, takich jak cienie bądź zmiany kolorów. Byłoby to bliższe ludzkiemu postrzeganiu obiektów. Wszakże coraz silniejszy jest w nauce i inżynierii nurt kopiowania rozwiązań zaczerpniętych z natury.

Problem odnajdywania położenia brwi i kształtu ust zrealizowano o bardzo prostą, lecz skuteczną metodę analizy projekcji obrazu. Dysponując wiedzą o tym, co dany obraz powinien przedstawiać, łatwo uzyskać z nich dużą ilość potrzebnych danych. Spostrzeżenia takie zebrano i wykorzystano w analizatorach poszczególnych obszarów obrazu. Dzięki temu, po zapewnieniu dobrej jakości obrazu, udało się uzyskać więcej niż zadowalające rezultaty w wykrywaniu bieżącego kształtu obiektów. Szczególną uwagę skupiono na brwiach i ustach, gdyż to one, zgodnie ze spostrzeżeniami autora, są najbardziej aktywne przy ekspresji mimicznej.

Jednocześnie zanotowano niską sprawność zastosowanej metody wykrywania położenia powiek (stopnia otwarcia oczu). Jednym z głównych powodów jest mała zmienność takich wartości, co przy obecności zakłóceń może nie wystarczyć

do poprawnego rozpoznania. Inną przyczyną zawodności tej metody jest częste występowanie cieni pod łukiem brwiowym, co znacząco zmniejsza kontrast na linii oko-powieka.

Do analizy gestów przyjęto metodę porównywania położenia punktów węzłowych z ich naturalnym, neutralnym umiejscowieniem. Testy potwierdziły skuteczność i celowość zastosowania takiego schematu postępowania. Pozwoliło to na uniezależnienie się od cech indywidualnych użytkownika, jak i ułatwiło empiryczne wyznaczenie jednego uniwersalnego zestawu wartości, służących do detekcji gestów. Przyjęto metodę progowania przemieszczeń punktów charakterystycznych wspomnianymi wartościami. Podejście takie wydawało się intuicyjne i w istocie, spełniło stawiane mu wymagania.

Napotkano przy tym problem złego lub niedokładnego zarejestrowania neutralnego położenia wspomnianych punktów. Podczas pracy programu mogą wystąpić chwilowe zakłócenia, wynikające na przykład ze zmiany czułości kamery, mrugnięcia oczami itp. Jeżeli w takim momencie zostaną zarejestrowane parametry inicjalizacyjne, wówczas dalsze rozpoznawanie gestów będzie zaburzone.

Analiza dostępnych danych pozwoliła na wyróżnienie zestawu 24 różnych gestów, które system potrafi rozpoznać. Jakość ich identyfikacji nie jest jednakowa, co wynika z opisanych już uwarunkowań.

Dysponując informacją o aktualnie rejestrowanych gestach zdecydowano się na zaimplementowanie dodatkowej funkcjonalności systemu - określaniu emocji wyrażanych mimicznie. Bazując na wspomnianych badaniach antropologicznych opracowano listę 7 stanów emocjonalnych wraz z zestawami gestów, które wchodzą w ich skład. Dzięki temu otrzymano ważną informację w kontekście komunikacji człowiek - komputer.

Zaproponowany zestaw reguł wnioskowania oparto znów na własnych spostrzeżeniach odnośnie ekspresji mimicznej. Pozwolił on w efekcie na określanie stanu emocjonalnego z dokładnością około 64 % (pod warunkiem poprawnego wykrycia pojedynczych gestów). Wynik taki można uznać za dostateczny. Należy przy tym zwrócić uwagę na różną skuteczność w wykrywaniu poszczególnych emocji. System taki okazał się dobrze działać przy detekcji śmiechu (100 %) i zaskoczenia (100 %). Nie sprostał natomiast zadaniu wykrywania smutku (25 %).

Przeprowadzone testy wykazały wszelako trafność założenia o detekcji emocji w oparciu o zestawy gestów, co jest niejako nawiązaniem do omawianego na wstępie systemu FACS oraz zestawu jednostek AU. Pamiętać należy, iż klasyfikator ten bazuje na identyfikacji gestów jednostkowych, dlatego metod usprawnienia takiego systemu regułowego można szukać w ich lepszej detekcji.

Alternatywą dla takiego podejścia są klasyfikatory, będące formą implementacji sztucznej inteligencji. Detekcja emocji za pomocą sieci neuronowej, badającej wartości parametrów nie przyniosła zadowalających rezultatów (średnio 59 % po-

prawnych rozpoznań). Na wynik taki wpływ może mieć wiele czynników, jednak najbardziej prawdopodobne jest nieoptymalnie przeprowadzony proces nauki sieci. Dodatkowo, w zbiorze uczącym znalazło się wiele podobnie reprezentowanych emocji (zadowolenie - śmiech, zaskoczenie - strach). Zmniejszenie ich liczby pozwoliłoby sieci na lepsze wyróżnienie klas w zbiorze uczącym, a co za tym idzie - lepsze rozpoznawanie emocji w testach. Podczas testów sieć taka udowodniła jednak swoją przydatność do detekcji konkretnych emocji: śmiechu, niesmaku i złości (wszystkie 100% poprawnych rozpoznań).

Druga z sieci neuronowych operuje na wyznaczonych wcześniej gestach jednostkowych i stara się klasyfikować ich zestawy w zbiory odpowiadające emocjom. Jest to więc podejście hybrydowe. Łączy empirycznie wyznaczane kryteria detekcji gestów z klasyfikatorem neuronowym. Testy wykazały zasadność takiego podejścia. Uzyskano najlepszą (70 %) średnią jakość klasyfikacji. Co prawda sieć ta nie wyróżniła tak wielu emocji bezbłędnie rozpoznawanych, natomiast jakość wykrywania wszystkich była na poziomie co najmniej 50 %. Znów można pokusić się o przypuszczenie, iż lepsza detekcja gestów jednostkowych pomogłaby jeszcze ulepszyć ten rodzaj klasyfikatora. Mają tu też zastosowanie wszystkie uwagi, wymienione odnośnie usprawnienia jakości nauczania pierwszej sieci.

Możliwe kierunki rozwoju systemu to przede wszystkim usprawnienie mechanizmu wykrywania oczu i ust oraz próby uniezależnienia ich od zmiennych warunków akwizycji. Być może zastosowanie drugiego, pracującego równolegle detektora pozwoliłoby na zebranie ich zalet i wzajemne wykluczenie błędów. Innym ważnym usprawnieniem byłoby dalsze przeanalizowanie zastosowanych reguł decyzyjnych w celu zwiększenia ich uniwersalności i dostosowania do większej liczby form ekspresji mimicznej. Przydatna byłaby tu dyskusja ze specjalistami z dziedziny antropologii i psychologii. Optymalizacja zastosowanych sieci neuronowych jak i procesów ich uczenia na pewno przyniosłaby znaczącą poprawę w jakości klasyfikacji. Korzystną zmianą, ze względu na komfort użytkownika byłoby wprowadzenie pełnej automatyki w zakresie korekcji parametrów działania i ustalania momentu pozyskania obrazu wzorcowego. Sam system mógłby być za to przystosowany do pracy jako aplikacja uruchamiana z przeglądarki www, bądź jako działający w tle plugin np. do komunikatora internetowego.

Bibliografia

[1] http://en.wikipedia.org/wiki/facial_action_coding_system. Źródło internetowe. [cytowanie na str. 5]

[2] http://en.wikipedia.org/wiki/hsl_and_hsv. Źródło internetowe. [cytowanie na str. 42]

[3] http://java.sun.com/javase/technologies/desktop/media/jai/. Źródło internetowe. [cytowanie na str. 56]

[4] http://java.sun.com/javase/technologies/desktop/media/jmf/. Źródło internetowe. [cytowanie na str. 56]

[5] http://mediweb.pl/psyche/wyswietl_vad.php?id=640. Źródło internetowe. [cytowanie na str. 2, 3]

[6] http://niewerbalnie.info/wrodzone_zachowania. Źródło internetowe. [cytowanie na str. 4]

[7] http://pl.wikipedia.org/wiki/mimika. Źródło internetowe. [cytowanie na str. 4]

[8] http://student.bmj.com/back_issues/0404/education/140.html. Źródło internetowe. [cytowanie na str. 4]

[9] http://www.ai.c-labtech.net/sn/sneuro.html. Źródło internetowe. [cytowanie na str. 54]

[10] http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm. Źródło internetowe. [cytowanie na str. 5]

[11] http://www.cyfrografia.pl/sony_dsc_t200.html. Źródło internetowe. [cytowanie na str. 3]

[12] http://www.usatoday.com/tech/columnist/edwardbaig/2007-10-03-sony-cybershot-dsc-t200_n.htm. Źródło internetowe. [cytowanie na str. 3]

[13] T.F. Cootes and C.J. Taylor. Statistical models of appearance for computer vision. Technical report, Imaging Science and Biomedical Engineering, University of Manchester, 2004. [cytowanie na str. i, 10, 11, 12, 18, 21, 22]

[14] P. Ekman, W.V. Friesen, and J.C. Hager. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978. [cytowanie na str. 5, 13, 15]

[15] Mariusz Kubanek. *Metoda Rozpoznawania Audio-Wideo Mowy Polskiej w Oparciu o Ukryte Modele Markowa.* PhD thesis, Politechnika Częstochowska, 2005. [cytowanie na str. 10, 35, 41]

[16] M. Lewicka and K. Stańczyk. System biometryczny identyfikujący osoby na podstawie cech osobniczych twarzy. Master's thesis, AGH Kraków, 2008. [cytowanie na str. 10, 11, 36]

[17] M Pantic, I. Patras, and L. Rothkrantz. Facial action recognition in face profile image sequences. In *IEEE Int'l Conf. on Multimedia and Expo*, page 37–40. IEEE, 2002. [cytowanie na str. 9]

[18] M. Pantic and Leon J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 22(12), 2000. [cytowanie na str. 8, 11, 13, 14]

[19] M. Pantic and L.J.M. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, (18):881–905, 2000. [cytowanie na str. i, 11, 15]

[20] M. Pantic and L.J.M. Rothkrantz. Facial gesture recognition in face image sequences: a study on facial gestures typical for speech articulation. In *Systems, Man and Cybernetics, 2002 IEEE International Conference on*, 2002. [cytowanie na str. 10, 11, 14]

[21] M. Pantic, M. Tomc, and L.J.M. Rothkrantz. A hybrid approach to mouth features detection. In *Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference*, 2001. [cytowanie na str. 12, 41, 42]

[22] J. Park, C.G. Looney, and H. Chen. Fast connected component labeling algorithm using a divide and conquer technique. Technical report, University of Alabama, Tuscaloosa; University of Nevada, Reno. [cytowanie na str. 34]

[23] M Piccardi. Background subtraction techniques: a review, 2004. University of Technology, Sydney. [cytowanie na str. 11, 29]

[24] M. Smiatacz. Automatyczna lokalizacja i śledzenie obiektów na obrazie, 2005. Politechnika Gdańska. [cytowanie na str. 11, 28]

[25] M. Smiatacz and W. Malina. Aktywne modele kształtu i ich biometryczne zastosowania. Technical report, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska. [cytowanie na str. 10, 11, 12, 21]

[26] R. Tadeusiewicz. *Sieci neuronowe.* Akademicka Oficyna Wydawnicza, Warszawa, 1993. [cytowanie na str. 54]

[27] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop.* IEEE, 2006. [cytowanie na str. 8]

[28] J Viola and M.J. Jones. Robust real-time object detection. Technical report, Compaq Cambridge Research Laboratory, 2001. [cytowanie na str. 11, 32, 33]