Uniwersytet Jagielloński w Krakowie

Wydział Fizyki, Astronomii i Informatyki Stosowanej

Łukasz Wójcik

Nr albumu: 1188524

Opracowanie jedno- i wielomodalnych modeli predykcji emocji

Praca magisterska
na kierunku Informatyka Stosowana

Praca wykonana pod kierunkiem:
dr inż. Krzysztofa Kutta
Zakład Technologii Gier

Kraków 2023

Abstract

This in an abstract in English...

Abstrakt

...a to jest abstrakt po polsku.

Spis treści

Wstęp

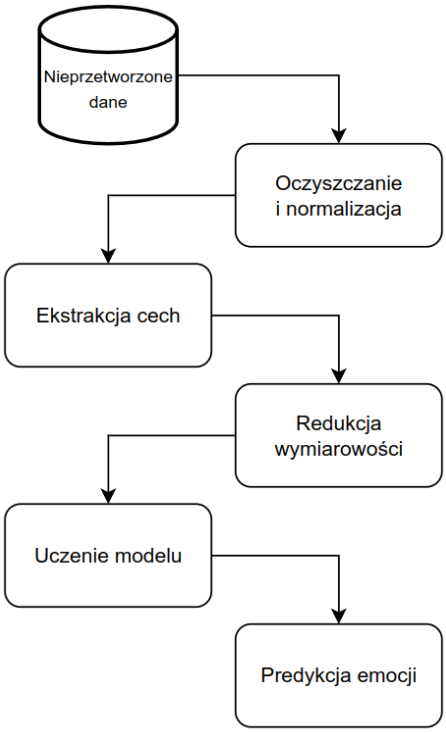
Wstęp...

Rozdział 1

Automatyczna predykcja emocji

1.1 Uogólniony system rozpoznawania emocji

W badaniach nad automatycznym rozpoznawaniem emocji dominują systemy oparte na uczeniu maszynowym lub modelach statystycznych [1, 2]. Powoduje to, że wykonuje się w nich podobne kroki. Dane w takich systemach przechodzą zazwyczaj sekwencyjnie przez tak zwany potok (ang. *pipeline*) [3].



Rysunek 1.1: Ogólny schemat systemu predykcji emocji

Rysunek 1.1 przedstawia uproszczony schemat systemu predykcji emocji. Na początku system otrzymuje nieprzetworzone dane, często nazywane surowymi (ang. *raw*

data). W zależności od źródła mogą to być zdjęcia, filmy, zapisy sygnałów biofizycznych (elektrokardiografia, elektroencefalografia itp.) oraz wiele innych. Są to wartości pochodzące bezpośrednio z sensorów, bazy danych lub publicznie dostępnych zbiorów [1, 4].

Dane wejściowe zazwyczaj nie są wystarczającej jakości dlatego kolejny krok to ich oczyszczanie. Może to być na przykład redukcja szumów w sygnale, odrzucanie skrajnych wartości lub uzupełnianie brakujących. Dodatkowo niektóre modele dają lepsze wyniki po normalizacji danych [3].

Większość modeli nie przyjmuje na wejściu surowych danych, dlatego następnie wykonuje się proces ekstrakcji cech (ang. *feature extraction*). Pozwala on na zmianę danych wejściowych na wartości istotne dla rozpoznawania emocji. Często są to wartości z funkcji i miar statystycznych, takie jak mediana, średnie, odchylenia itp. Przykładowe cechy to geometria twarzy, szybkość mówienia, czas pomiędzy uderzeniami serca [5].

W zależności od metody ilość cech może wynosić ponad 700 [6], dlatego w niektórych systemach kolejnym krokiem jest redukcja wymiarowości. Ma ona na celu zmniejszenie liczby cech wejściowych poprzez łączenie tych silnie skorelowanych, rzutowanie w mniej wymiarowe przestrzenie lub odrzucanie wartości, które nie poprawiają wyników. Powszechnie stosowane podejścia to: analiza głównych składowych (ang. *principal components analysis (PCA)*), grupowanie hierarchiczne (ang. *hierarchical cluster analysis (HCA)*), Gaussian random projection [3].

Po uzyskaniu ostatecznych danych następuje trening modelu, zazwyczaj jest to uczenie nadzorowane [4]. Oznaczanie danych dobywa się na dwa sposoby. W pierwszym każdy wektor danych ma przypisaną kategoryczną emocję, na przykład strach lub złość. W drugim emocje opisane są w przestrzeni wielowymiarowej, zatem zamiast jednej kategorii posiadają zazwyczaj dwie lub trzy wartości. Do klasyfikacji stosuje się różnorakie podejścia, między innymi: support vector machine (SVM), random forest classifier (RFC), stochastic gradient descent (SGD), AdaBoost, k-nearest neighbor (k-NN), hidden Markov models (HMM), linear discriminate analysis (LDA), sztuczne sieci neuronowe [1, 2, 7].

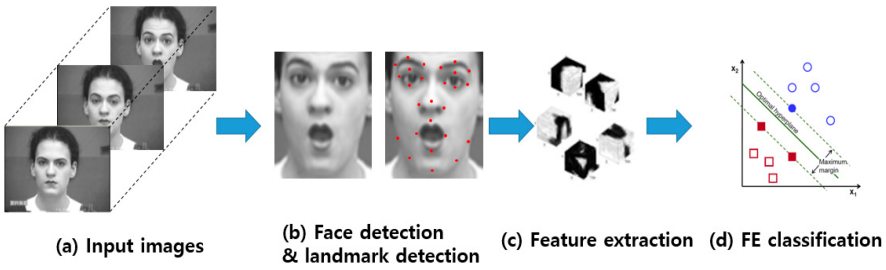
1.2 Metody rozpoznawania emocji

Istnieje wiele sposobów, na podstawie których można wnioskować stan emocjonalny człowieka. Pozwala to na wykorzystanie bardzo zróżnicowanych podejść, od

oceny wyglądu, przez analizę zachowań, aż po pomiary aktywności elektrycznej w organizmie. Poniżej znajduje się opis najczęściej stosowanych metod [1, 2], ich wady oraz zalety.

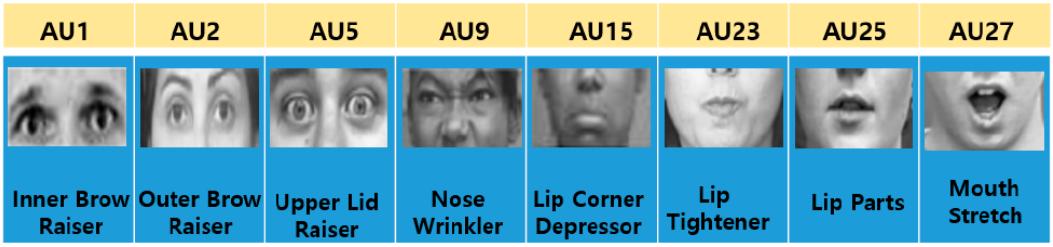
1.2.1 Wyraz twarzy

Ludzka twarz jest bardzo znaczącym źródłem informacji i odgrywa dużą rolę w komunikacji niewerbalnej. Na jej podstawie można oceniać między innymi: płeć, wiek, pochodzenie etniczne, czy stan emocjonalny [5]. Dzięki temu twarz jest bardzo popularnym źródłem w automatycznym rozpoznawaniu emocji, z początkami prac sięgającymi lat 90. XX wieku [5].



Rysunek 1.2: Schemat systemu rozpoznającego emocje na podstawie twarzy. *Źródło: [7]*

Rysunek 1.2 przedstawia ogólny schemat systemu rozpoznającego emocje na podstawie wyrazu twarzy. Na wejściu program otrzymuje pojedyncze zdjęcie lub nagranie zawierające twarz. Pierwszym krokiem jest wykrycie twarzy. W dostarczonym źródle może być ich wiele. Następnie przeprowadza się proces ekstrakcji charakterystycznych miejsc. Po uzyskaniu danych o twarzy zostają one przetworzone przez algorytm uczenia nadzorowanego [5].



Rysunek 1.3: Przykłady różnych Action Units w trzech częściach twarzy. *Źródło: [7]*

Jedno z popularnych podejść rozpoznaje emocje na podstawie Facial Action Coding System (FACS) [8]. Zbiór ten zawiera, w zależności od wersji, od 33 do 44 tak zwanych Action Units (AU). Powstały one przy pomocy stymulacji elektrycznej

mięśni twarzy, które biorą udział w wyrażaniu emocji. Dzięki temu uzyskano obiektywne ruchy mięśni o różnej intensywności zależnej od napięcia prądu. Sam zbiór nie zawiera ścisłego określenia połączeń AU i odpowiadającym im emocjom, a jedynie hipotezy [5].

Inne podejścia bazują na zbiorach, w których osoby były proszone o wyrażenie danej emocji. Pozyskane w ten sposób dane mają jednak wadę w postaci zbyt intensywnego wyrazu twarzy, w dodatku opartych na stereotypach. Osoba poproszona o to, aby pokazała zdziwienie, zazwyczaj wygląda zupełnie inaczej, niż gdy jest naprawdę zdziwiona. Nawet dobrze wyszkolony aktor nie jest w stanie dokładnie odwzorować naturalnej reakcji [5]. Powoduje to, że modele szkolone na takich zbiorach nie są w stanie rozpoznawać emocji wyrażanych w sposób „normalny". Jedną z możliwości zapobiegania temu zjawisku jest tworzenie zbiorów, w których emocje są wywoływane przez prawdziwe zdarzenia, a nie odgrywane przez aktorów [5].

Główną zaletą rozpoznawania emocji na podstawie twarzy jest stosunkowa prostota, aparaty są tanie i powszechnie dostępne. Dodatkowo zbieranie danych nie wymaga kontaktu fizycznego i nie powoduje dyskomfortu. Same wyrazy twarzy dla wielu emocji są uniwersalne między członkami różnych kultur, płci i niezależne od wieku [5].

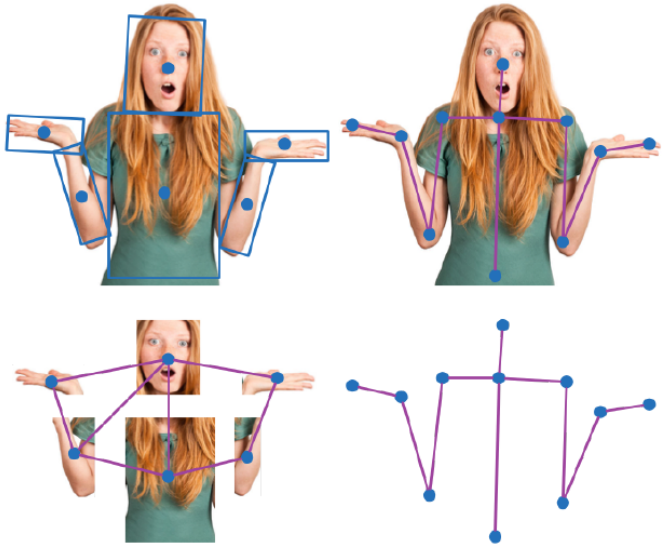
Mimo to z tym podejściem wiąże się wiele problemów. Począwszy od trudności wynikających z samego rozpoznawania twarzy, na przykład różne oświetlenie, czy kąt, pod jakim znajduje się twarz. Następnie pojawiają się problemy związane z emocjami: osoba jest w stanie celowo nie wyrażać żadnych emocji lub mogą być one bardzo nikłe [5].

1.2.2 Postawa ciała i gestykulacja

Drugim bardzo ważnym źródłem informacji o emocjach jest postawa ciała człowieka, jego gesty lub ich brak. Stanowią one znaczną część komunikacji niewerbalnej, ruch dłoni jest drugim co do wielkości źródłem, mówiącym o stanie emocjonalnym, więcej informacji pochodzi jedynie z wyrazu twarzy [9]. Co więcej, postawa ciała pomaga w zmaganiu się z aktualnie odczuwanymi emocjami [10].

Jednym z powszechnie stosowanych sposobów śledzenia ruchu ciała są kamery termowizyjne [5]. Pomiary są możliwe dzięki odblaskowym płytkom, które umieszczane są na odzieży. Pozwala to na zapis ruchu w trójwymiarowej przestrzeni. Tego typu podejście wymaga jednak noszenia specjalnego stroju, a dokładność jest zależna od ilości znaczników. Z tego powodu mierzenie ruchu dłoni, a zwłaszcza palców jest problematyczne. Z drugiej strony zbieranie jest mniej danych, a ich przetwarzanie jest łatwiejsze. Dodatkowo zapewniona jest anonimowość badanych [5].

Dzięki rozwojowi widzenia maszynowego możliwe stało się również używanie zwykłych kamer. Takie podejście zapewnia większą swobodę, nie wymaga specjalnego stroju. Co najważniejsze pozwala na dokładniejsze odwzorowanie ruchów, zwłaszcza palców. To podejście również musi zmagać się z problemami typowymi dla rozpoznawania obrazów: oświetlenie, kolor skóry, ubrania mogą negatywnie wpływać na dokładność [5].



Rysunek 1.4: Sposoby reprezentowania ciała w komputerze: zbiór częsci ciała (lewa strona) oraz reprezentacja szkieletowa (prawa strona). *Źródło: [9]*

Rysunek 1.4 przedstawia dwa sposoby modelowania ludzkiego ciała w systemach komputerowych. Po lewej stronie widnieje model oparty na częściach ciała (ang. *part based model*). Każda część jest rozpoznawana osobno na podstawie wiedzy o budowie ludzkiego ciała. Otrzymywana jest reprezentacja dwuwymiarowa. Po prawej stronie przedstawiono model szkieletowy (ang. *kinematic model*). W tej reprezentacji ciało jest zbiorem wierzchołków połączonych krawędziami, przez co można je reprezentować jako graf [9]. Wierzchołki interpretowane są jako stawy, które posiadają pewne stopnie swobody, odpowiednie dla danej części ciała. Pozwala to na złożoną reprezentację w przestrzeni trójwymiarowej [5].

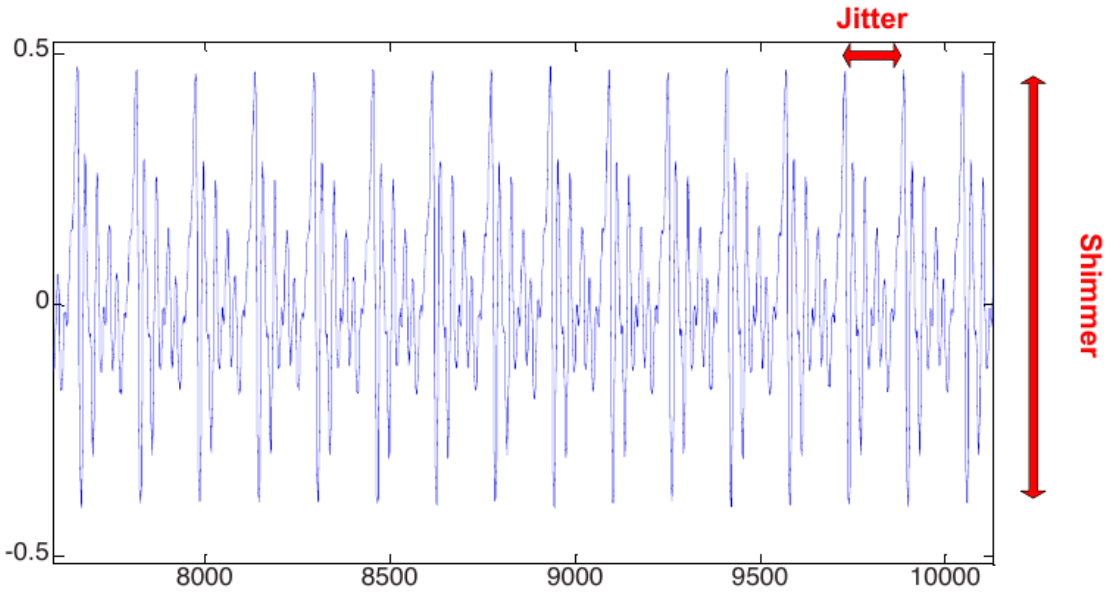
Po uzyskaniu reprezentacji ciała, w systemie następuje proces rozpoznawania postawy, a następnie oceny emocji. Używa się do tego zarówno statycznych obrazów, jak i nagrań ruchu [9, 10].

1.2.3 Mowa

Poza niewerbalnymi źródłami, emocje można również rozpoznawać na podstawie mowy. Ludzki głos stanowi bardzo bogate źródło informacji. Pozwala na wnioskowanie o wieku, płci, stanie emocjonalnym, osobowości, dialekcie i pochodzeniu mówcy [11].

W porównaniu do poprzednich źródeł mowa jest o wiele bardziej podatna na zakłócenia, szum, hałasy w tle. Wymaga więc dokładniejszego procesu oczyszczania. Bardzo ważna jest również normalizacja danych. Zakres podstawowej częstotliwości głosu, który wynosi około 50 — 500 Hz, jest o wiele większy niż różnica między wypowiedzią neutralną i w stanie złości, czyli około 68 Hz [5].

Po oczyszczeniu i normalizacji następuje proces ekstrakcji cech niskiego poziomu (ang. *low-level descriptors (LLD)*). Są to wartości oparte o częstotliwość głosu oraz o zmiany w sposobie wypowiedzi (na przykład szybkość mówienia lub poziom głośności). Sama ilość cech niskiego poziomu nie jest z góry określona i może być różna w zależności od podejścia. Do najpopularniejszych LLD należą: fundamental frequency (F0), Mel-frequency cepstral coefficients (MFCCs), jitter, shimmer, harmonic-to-noise ratio oraz wartości z widma akustycznego [5, 12].



Rysunek 1.5: Przykładowy sygnał mowy z zaznaczonymi jitter i shimmer.
Źródło: [13]

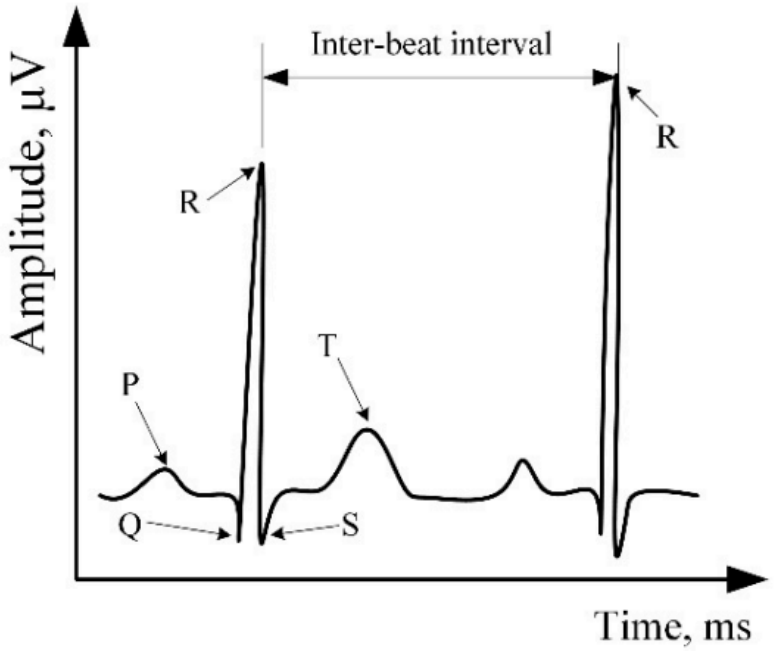
Po uzyskaniu cech niskiego poziomu można zastosować funkcje i miary statystyczne, takie jak średnie i odchylenia, aby otrzymać tak zwane cechy wysokiego poziomu (ang. *high-level descriptors (HLD)*) [5].

Podobnie jak wyrazy twarzy, mowa jest zależna od kultury i pochodzenia osoby. Dodatkowo wyszkolona osoba jest w stanie kontrolować wymowę w taki sposób, aby ukrywać odczuwane emocje lub udawać inne [5].

1.2.4 Sygnały biofizyczne

Emocje wywołują również zmiany, których nie da się zaobserwować za pomocą wzroku lub słuchu. Różne stany emocjonalne wpływają między innymi na szybkość bicia serca, wydzielanie potu, oddech, temperaturę ciała. Są to parametry, które można zmierzyć i wnioskować na ich podstawie odczuwane emocje [5].

Jednym z najpopularniejszych sposobów jest elektrokardiografia (EKG), czyli mierzenie aktywności elektrycznej serca. Do pomiarów używa się elektrod umieszczonych na skórze, najczęściej jest ich 3 lub 12. Analiza sygnału odbywa się na podstawie załamków P, Q, R, S, T [5].



Rysunek 1.6: Przykładowy sygnał EKG z zaznaczonymi załamkami. *Źródło: [2]*

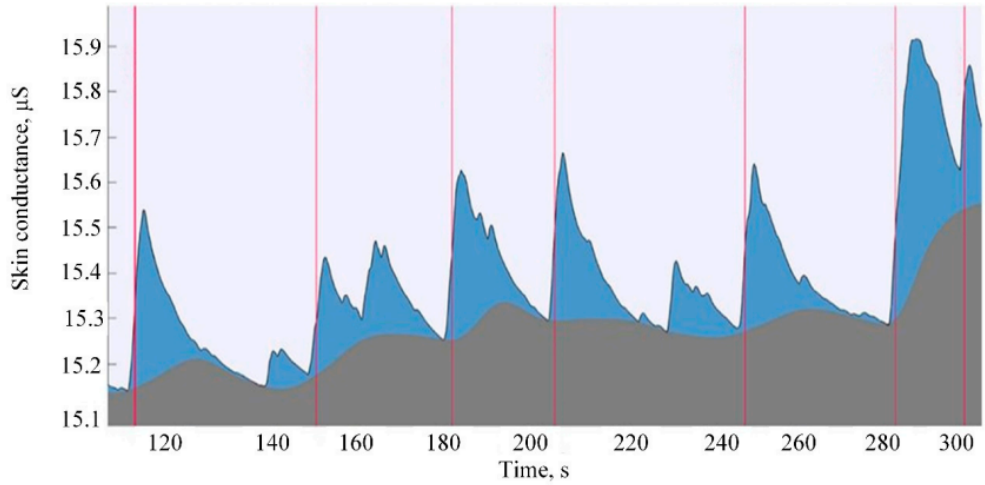
Rysunek 1.6 przedstawia przykładowy sygnał EKG. Jako pierwszy występuje załamek P, który oznacza depolaryzację mięśnia przedsionków. Potem następuje zespół QRS opisujący depolaryzację mięśnia komór. Po nim pojawia się załamek T odpowiadający repolaryzacji komór [2].

W automatycznym rozpoznawaniu emocji najczęściej bierze się pod uwagę zespół QRS oraz odległości między załamkami R (ang. *R-R interval / inter-beat interval*),

które wykorzystuje się w analizie zmienności rytmu zatokowego (ang. *heart rate variability (HRV)*) [5].

Drugim często używanym sygnałem biofizycznym jest reakcja skórno-galwaniczna, powszechnie używa się dwóch skrótów: GSR (ang. *galvanic skin response*) lub EDA (ang. *electrodermal activity*). Opisuje ona zmiany w przewodnictwie skóry spowodowane aktywnością gruczołów potowych. Prowadzi to do różnic w wilgotności i w następstwie do zmiany oporu elektrycznego [2].

Pomiary wykonuje się za pomocą elektrod, które mogą być umieszczone w dowolnym miejscu na skórze. Zazwyczaj wykorzystuje się miejsca najbardziej czułe na zmiany emocjonalne: dłonie oraz podeszwy stóp [5].



Rysunek 1.7: Przykładowy sygnał GSR. Czerwone linie oznaczają momenty pojawiania się stymulantu. *Źródło: [2]*

Rysunek 1.7 przedstawia przykładowy sygnał GSR, który składa się z dwóch głównych komponentów. Szarym kolorem zaznaczono tonic component, który zmienia się powoli i zależy głównie od reakcji na czynniki środowiska (temperatura, wilgotność powietrza itp.). Na niebiesko oznaczono phasic component, przejawiający się jako krótkie piki w odpowiedzi na stan emocjonalny [2].

Poza elektrokardiografią oraz reakcją skórno-galwaniczną stosuje się również wiele innych podejść.

Fotopletyzmografia (ang. *photoplethysmography (PPG)*) jest alternatywnym sposobem mierzenia aktywności serca. Do pomiarów używa się światła, które reaguje na zmiany w ilości krwi w tkankach. Różnice w odbijanym lub przepuszczanym świetle odpowiadają uderzeniom serca [2].

Elektroencefalografia (EEG) jest używana do badania aktywności mózgu na podstawie fal $\delta, \theta, \alpha, \beta, \gamma$. Pomiary odbywają się za pomocą elektrod umieszczonych na głowie. Zazwyczaj używa się 8, 16 lub 32 pary [2].

Elektromiografia (EMG) służy do pomiaru aktywności elektrycznej mięśni. Podczas skurczu mięśni pojawia się napięcie, które można zmierzyć na powierzchni skóry przy pomocy elektrod. EMG jest zazwyczaj stosowane dla mięśni twarzy [2].

Oddychanie jest również sygnałem biofizycznym. Pomiary wykonuje się zazwyczaj za pomocą opaski wokół klatki piersiowej, która mierzy jej ruch wywołany wdechami i wydechami [2].

Dużym problemem sygnałów biofizycznych są zakłócenia związane z aktywnością człowieka. Ruch ma duży wpływ na pracę serca, która nie zmienia się liniowo w stosunku do wysiłku. Kichnięcie powoduje w organizmie reakcję podobną do odczuwania strachu, mimo że osoba kichająca raczej nie jest przestraszona [5].

1.3 Reprezentacja emocji w systemie komputerowym

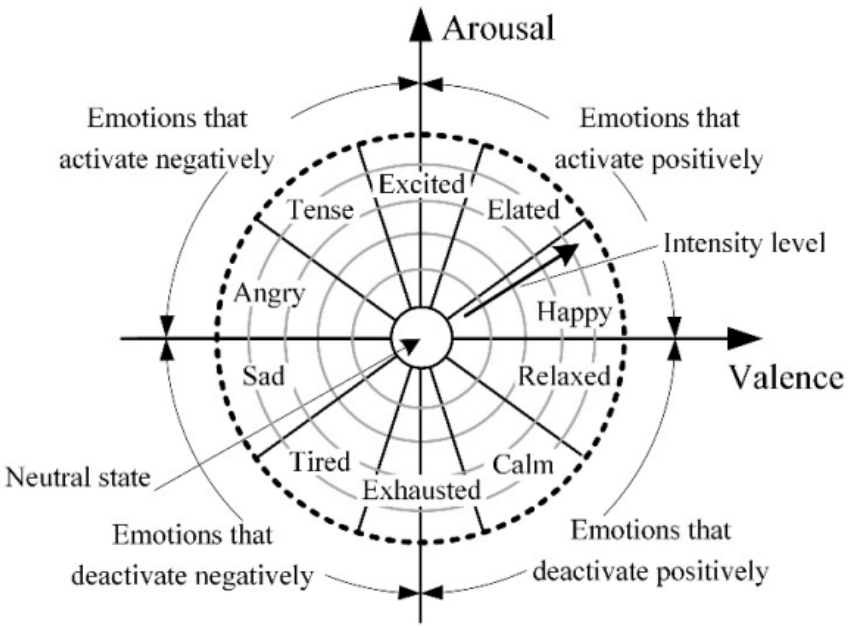
Po uzyskaniu danych należy je przypisać do odczuwanych emocji. Najprostszym sposobem jest przydzielenie im pewnej kategorii, na przykład: strach, złość, radość, smutek itp. W automatycznym rozpoznawaniu emocji ich liczba jest zazwyczaj niewielka i wynosi od 4 do 8 [2]. Czasem są to również klasyfikatory binarne, które przewidują jedynie czy dane należą do danej klasy, czy nie.

Rysunek 1.8: Przykładowa baza danych zawierająca zdjęcia twarzy przedstawiające podstawowe emocje. *Źródło: [14]*

Najczęściej wykorzystuje się kategorie należące do tak zwanych podstawowych emocji. Zostały one zaproponowane między innymi przez Paula Ekmana w 1971 roku [15]. Należą do nich: radość, smutek, złość, zdziwienie, strach oraz wstręt.

Inne popularne podejścia reprezentują emocje za pomocą dwóch lub trzech ciągłych wartości liczbowych. Stany emocjonalne są przedstawione w przestrzeni, dzięki czemu można reprezentować o wiele więcej kategorii, w sposób bardziej płynny i dokładny [5].

Wiele podejść opiera się na dwuwymiarowym modelu zaproponowanym przez Jamesa Russella [16]. Emocje w tym modelu reprezentowane są za pomocą wartości opisujących przyjemność odczuwanej emocji (ang. *valence*) oraz pobudzenie jakie wywołuje (ang. *arousal*). Sam model jest zazwyczaj w kształcie koła, które może być podzielone na wycinki przedstawiające emocje. Punkt leżący w danym wycinku przedstawia odpowiednią emocję [2].



Rysunek 1.9: Kołowy model oparty o torię Russella. *Źródło: [2]*

Dwuwymiarowy model nie jest jednak w stanie wystarczająco rozróżniać niektóre emocje. Przykładowo strach oraz złość są reprezentowane jednakowo: przez wysokie pobudzenie i niską przyjemność. Aby poprawić rozpoznawanie stanów emocjonalnych, inne podejścia dodają trzeci wymiar utożsamiany z dominacją, jaką wywiera dana emocja [5].

Samo przypisywanie emocji do danych odbywa się za pomocą dwóch podejść [5]. Pierwsze z nich opiera się na wyszkolonych obserwatorach, którzy oceniają stan emo-

cjonalny badanej osoby. Nie zawsze jest to jednak możliwe, dlatego drugim powszechnie stosowanym sposobem jest samoocena. Metoda ta jest prostsza i pozwala na klasyfikacje emocji w sygnałach biofizycznych. Jest jednak bardziej zawodna, ponieważ osoba może źle sklasyfikować odczuwaną emocje, lub niedokładnie ocenić moment, w którym do niej doszło [5].

1.4 Modalność w modelach predykcji emocji

Początkowo modele automatycznej predykcji emocji opierały się wyłącznie na jednym źródle informacji, były to zatem systemy jednomodalne. Ten trend był tym bardziej wzmacniany przez skupienie się na rozpoznawaniu emocji na podstawie wyrazu twarzy. Takie podejście ma jednak jedną główną wadę, system nie jest w stanie rozpoznawać emocji, gdy brakuje danych wejściowych. Zdarza się, że twarz jest zakryta, osoba nic nie mówi lub stoi nieruchomo. Aby zapobiec temu problemowi oraz przez chęć uzyskania lepszych wyników rozpoczęto prace nad systemami wykorzystującymi więcej niż jedno źródło informacji [5].

Model wielomodalny to taki, który wykorzystuje co najmniej dwa różne źródła informacji. Może to być twarz oraz mowa, gestykulacja i sygnały biofizyczne, lub wszystkie na raz. Tego typu systemy o wiele rzadziej napotykają problem braku danych oraz zapewniają lepsze wyniki [17].

Systemy wielomodalne zmagają się jednak z innymi problemami. Największy to łączenie ze sobą danych, które wymagają różnych okienek czasowych do analizy. Film może być analizowany na podstawie pojedynczych klatek, jednak sygnały biofizyczne lub mowa wymagają zazwyczaj dłuższych pomiarów, aby dać wartościowe dane [5].

Rozdział 2

Uczenie maszynowe

2.1 Podstawy uczenia maszynowego

W klasyfikacji jako funkcja straty często stosowana jest entropia krzyżowa *(ang. cross-entropy loss)*. Jej ogólny wzór wygląda następująco [4]:

$$H(P, Q) = \int P(x) \log Q(x) dx$$

gdzie P oznacza prawdziwe wartości zbioru testowego $P^*(x, y)$, a Q wartości przewidziane przez model $P_w(y|x)$. Celem uczenia jest zmiana w tak, aby zminimalizować $H(P^*(x, y), P_w(y|x))$.

2.2 Sztuczne sieci neuronowe

Rozdział 3

Część praktyczna

3.1 Zbiór danych

Krótki opis BIRAFFE2.

3.2 Opis systemu

Opis systemu, pipeline, przygotowanie danych itp. Opis użytych metod do ekstrakcji cech, jakieś wzory, na przykład dla k-means.

3.3 Wyniki

Porównanie wyników różnych modeli scikit-learn i TensorFlow oraz modalności, kilka tabelek.

Col1	Col2	Col2	Col3
1	6	87837	787
2	7	78	5415
3	545	778	7507
4	545	18744	7560
5	88	788	6344

Tabela 3.1: Table to test captions and labels.

Posdumowanie

Podsumowanie...

Bibliografia

[1] Ashwini Ann Varghese, Jacob P Cherian, and Jubilant J Kizhakkethottam. Overview on emotion recognition system. 2015. doi:10.1109/ICSNS.2015.7292443.

[2] Andrius Dzedzickis, Arturas Kaklauskas, and Vytautas Bučinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20:592, 01 2020. doi:10.3390/s20030592.

[3] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., 2019. ISBN 9781492032649.

[4] Stuart J Russell and Peter Norvig. *Artificial intelligence a modern approach*. Pearson, 2020.

[5] Rafael Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas. *The Oxford Handbook of Affective Computing*. Oxford University Press, 2015. doi:10.1093/oxfordhb/9780199942237.001.0001.

[6] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Chirag Raman, Jamie Shotton, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. 3d face reconstruction with dense landmarks. 2022. doi:10.48550/ARXIV.2204.02776.

[7] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2), 2018. doi:10.3390/s18020401.

[8] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. doi:10.1037/t27734-000.

[9] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, 12(2):505–523, 2021. doi:10.1109/TAFFC.2018.2874986.

[10] Andrea Kleinsmith and Nadia Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4:15–33, 01 2013. doi:10.1109/T-AFFC.2012.16.

[11] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814, 2021. doi:10.1109/ACCESS.2021.3068045.

[12] Mohammed Abdelwahab and Carlos Busso. Evaluation of syllable rate estimation in expressive speech and its contribution to emotion recognition. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 472–477. IEEE, 2014. doi:10.1109/SLT.2014.7078620.

[13] João Teixeira, Carla Oliveira, and Carla Lopes. Vocal acoustic analysis – jitter, shimmer and hnr parameters. *Procedia Technology*, 9:1112–1122, 12 2013. doi:10.1016/j.protcy.2013.12.124.

[14] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. doi:10.1109/CVPR.2017.277.

[15] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[16] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. doi:10.1037/h0077714.

[17] Sidney D'Mello and Jacqueline Kory. Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, page 31–38. Association for Computing Machinery, 2012. doi:10.1145/2388676.2388686.

Spis rysunków

Spis tabel