Zastosowanie metod uczenia maszynowego do automatycznej detekcji kłamstwa

Bartosz Latosek
Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych
Warszawa, Polska

***Streszczenie*—W niniejszej pracy omówiono zastosowanie metod uczenia maszynowego do automatycznej detekcji kłamstwa, ze szczególnym uwzględnieniem wykorzystania analizy ekspresji twarzy i mikroekspresji. W artykule przedstawiono teorię związaną z emocjami i ich ekspresją na twarzy, bazując na krótkotrwałych, niekontrolowanych mikroekspresjach, które są trudne do fałszowania. Zaprezentowano również proces przetwarzania danych, w tym wykrywanie twarzy, ekstrakcję punktów charakterystycznych oraz ich normalizację. W celu analizy sekwencyjnych danych zaproponowano użycie rekurencyjnych sieci neuronowych, takich jak LSTM i GRU.**

***Słowa kluczowe*—uczenie maszynowe, detekcja kłamstwa, mikroekspresje, analiza emocji, przetwarzanie obrazu, normalizacja danych, sieci rekurencyjne, LSTM, GRU**

I. Pojęcia ogólne

W celu późniejszego zagłębienia się w główną część artykułu, podsumowującą i opisującą dotychczasowe badania, należy zapoznać się z niektórymi zagadnieniami z dziedziny analizy emocji.

Pierwsza część niniejszego paragrafu poświęcona jest krótkiemu wprowadzeniu w zagadnienia związane z rozpoznawaniem emocji. Doświadczanie przeróżnych **emocji** jest zjawiskiem silnie skorelowanym z ludzką naturą, gdyż odgrywają one kluczową rolę w podejmowaniu decyzji, budowaniu relacji społecznych oraz adaptacji do dynamicznie zmieniającego się środowiska. Na potrzeby dalszych rozważań, możemy przedstawić definicję emocji jako wiadomość [1], którą chcemy wykryć i poprawnie sklasyfikować.

W tak przyjętym modelu, rolę sygnału pełnią **ekspresje twarzy**, czyli obserwowane zmiany w mimice twarzy. W klasycznej definicji detekcja emocji sprowadza się właśnie do detekcji ekspresji twarzy, które silnie związane są z przeżyciami. Można więc założyć, że np. następujące po sobie ruchy kącików ust, w połączeniu z uniesieniem brwi mogą (ale nie muszą) być wyznacznikiem odczuwania przez osobę monitorowaną radości. Odpowiednio doświadczona i zdeterminowana osoba jest jednak w stanie sfałszować wyniki modelu opierającego się w całości na sekwencjach ekspresji, poprzez ich symulację bądź powstrzymywanie.

O wiele trudniej jest sfałszować **mikroekspresje**, czyli krótkotrwałe i niekontrolowane zmiany w mimice twarzy [2]. Te mimowolne ruchy mięśni twarzy trwają od *1/25* do *1/5* sekundy, co czyni je trudnymi do świadomego kontrolowania. Najczęściej przywoływany do zobrazowania tego zjawiska przykład dotyczy jednej z pacjentek szpitala psychiatrycznego, w którym przeprowadzano badania. Pacjentka, którą wypisano ze szpitala jakiś czas później doświadczyła nieudanej próby samobójczej. Po zbadaniu nagrań odtwarzanych w zwolnionym tempie zauważono, że tuż przed uśmiechem na twarzy pacjentki (sugerującym poprawę jej stanu psychicznego) pojawiał się wyraz smutku.

Dzięki podbudowaniu podstaw teoretycznych oraz omówieniu kluczowych pojęć, niniejszy paragraf stanowi solidne fundamenty dla dalszej analizy.

II. Cel Pracy

Do zdefiniowania konkretnego celu realizowanej pracy magisterskiej, wymagane jest też zastanowienie się nad znaczeniem kłamstwa w kontekście emocji.

W najbardziej podstawowej formie detekcja kłamstwa na podstawie obrazu twarzy może zostać sprowadzona do wykrywania autentyczności uśmiechu osoby badanej na podstawie jej zdjęcia. Przydatne są tutaj wcześniej omówione zarówno ekspresje twarzy, jak i mikroekspresje - które, w ramach przypomnienia, dużo ciężej sfałszować. Szereg badań przeprowadzonych w celu wynalezienia i optymalizacji rozwiązań, pomocnych w wykrywaniu tak zdefiniowanego kłamstwa sprawił, że w omawianej pracy zdecydowano skupić się na bardziej zaawansowanym zagadnieniu.

W wyniku rozszerzenia prostych danych wejściowych stanowiących zdjęcia o dodatkowy wymiar czasu, otrzymujemy znacznie większe możliwości detekcji kłamstwa. Tak uzyskana sekwencyjność danych pozwala nam na klasyfikacje wyszczególnionych ciągów klatek wideo jako szczere bądź fałszywe. Możliwe staje się nawet przewidywanie intencji osób badanych w czasie rzeczywistym. Dodatkowo, w problemie użyteczne mogłyby okazać się próbki audio pochodzące z nagrań, ale ze względu na wykorzystywane bazy danych i naturę wykrywania anomalii w strumieniu dźwiękowym, autor pracy skupił się wyłącznie na klasyfikacji opartej o obraz wideo.

Wyszczególniając główne zadania, jakie wchodzą w cel niniejszej pracy można wymienić kolejno: przegląd ist-

niejących rozwiązań z dziedziny detekcji kłamstwa, ich ewaluacja oraz wybór najlepszego z nich, ocena jakości wybranego rozwiązania na co najmniej jednej znanej z literatury bazie danych oraz finalnie - próba poprawy jakości opracowanego algorytmu.

III. Bazy danych

Niniejszy paragraf poświęcony został opisowi ogólnej charakterystyki dostępnych źródeł danych oraz krótkiemu zestawieniu najpopularniejszych z nich.

Głównym problemem dotyczącym dostępnych baz danych jest brak nagrań w warunkach rzeczywistych. Dostępne źródła, takie jak **Facial Expression Recognition** [3], czy **Multi-PIE** [4] zawierają nagrania pochodzące kolejno z filmów oraz uzyskane w wyniku odwzorowywania konkretnych emocji przez aktorów.

Dodatkowym problemem są wyidealizowane warunki, w których zostały uzyskane dostępne nagrania. Większość z nich prezentuje osoby grające w dobrym oświetleniu, lub takie których twarz jest niezasłonięta i skierowana prostopadle do kamery. Próba wykorzystania modelu wytrenowanego na danych w takiej formie może spowodować, że będzie on sobie bardzo słabo radził w warunkach rzeczywistych.

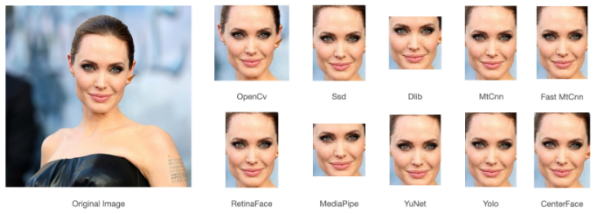
Pomocna okazuje się normalizacja i wstępna obróbka danych treningowych, które będą stanowiły wejście do modelu, a następnie wykorzystanie tego samego procesu do oczyszczenia i przetworzenia rzeczywistych danych testowych.

IV. Strumień przetwarzania danych

Jak wspomniano w poprzednim paragrafie, kluczowym elementem rozwiązania będzie wyspecjalizowany strumień przetwarzania danych, którego zadaniem będzie ekstrakcja i uwydatnienie cech, za pomocą których wybrany model będzie w stanie nauczyć się klasyfikować sekwencje klatek.

Najcenniejsze informacje - m.in. ekspresje i mikroekspresje, w gruncie rzeczy występują na twarzy (niektóre z nich przenoszą się również na gestykulację, która jest cięższa do uchwycenia, więc jako główne medium informacji zdecydowano się przyjąć twarz). Rzadkim zjawiskiem jest, by nagranie składało się z sekwencji zdjęć idealnie wykadrowanej twarzy. W większości przypadków, a w szczególności w danych pochodzących z czasu rzeczywistego wymagane jest, aby odpowiednio przyciąć każdy kadr tak, aby zawierał on w sobie tylko interesujące nas informacje - w tym przypadku samą twarz. W tym celu zdecydowałem się wykorzystać bibliotekę *DeepFace* [5], która do zadania detekcji twarzy wykorzystuje m.in. znany algorytm *Violi-Jonesa* [6].

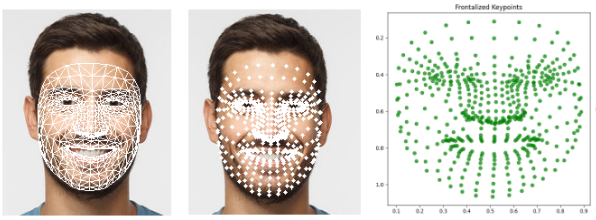
Eksperymenty wykazały, że modele przeznaczone do zadań z wykorzystaniem twarzy radzą sobie znacznie lepiej, gdy twarz jest odpowiednio obrócona [7]. Jest to mało kosztowny krok w potoku przetwarzania danych, a korzyści, które wynikają z jego zastosowania są widoczne.



Rysunek 1. Przykładowe detekcje twarzy pochodzące z oficjalnego repozytorium *Deepface* [5]

Sama operacja opiera się na prostych zależnościach i funkcjach trygonometrycznych, ale wymaga określenia położenia środków oczu na zdjęciu wejściowym.

Kolejnym ważnym krokiem w strumieniu przetwarzania danych jest ekstrakcja punktów charakterystycznych twarzy. Pominięcie tego kroku znacznie wydłużyłoby czas potrzebny na wytrenowanie modelu, ponieważ musiałby on *sam* nauczyć się rozpoznawać obszary pikseli odpowiadające częścią twarzy, a dodatkowo w takich danych wejściowych przeważałyby informacje zbędne (takie jak na przykład piksele zawierające tło). Do ekstrakcji punktów charakterystycznych wykorzystany został moduł *FaceMesh* przystosowany do tego zadania, zawarty w bibliotece *mediapipe* [8]. Zawiera on szereg gotowych, zoptymalizowanych algorytmów umożliwiających estymację *486* punktów charakterystycznych twarzy w *3* wymiarach.

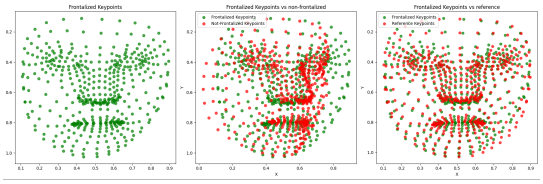


Rysunek 2. Przykładowa ekstrakcja cech z twarzy (grafika autorska)

Punkty charakterystyczne niwelują wpływ złego oświetlenia, a nawet przesłonięcia części twarzy aktora, ale w dalszym ciągu podatne są na nieznormalizowanie wynikające z różnych kątów obrotu twarzy względem kamery. W idealnych warunkach, punkt środkowy twarzy byłby statyczny przez wszystkie klatki filmu, a pozostałe przemieszczałyby się w określonym obszarze zależnym od ich odległości od punktu środkowego. W celu zbliżenia się do takich warunków, można wykorzystać technikę frontalizacji.

Jeden ze sposobów zakłada wykorzystanie punktów referencyjnych, które uzyskuje się w wyniku wykonania algorytmu ekstrakcji cech z twarzy, na obrazku przedstawiającym twarz we wzorcowym położeniu. Następnie, punkty pochodzące z nieznormalizowanych danych nakładane są na punkty referencyjne z użyciem *Analizy Prokrusta* [9].

Uzyskana w wyniku powyższej operacji siatka twarzy może zostać wykorzystana do obliczenia odległości punk-



Rysunek 3. Normalizacja na podstawie punktów charakterystycznych (grafika autorska)

tów kluczowych od punktu centralnego, co samo w sobie może już zostać wykorzystane jako dane wejściowe dla docelowego modelu.

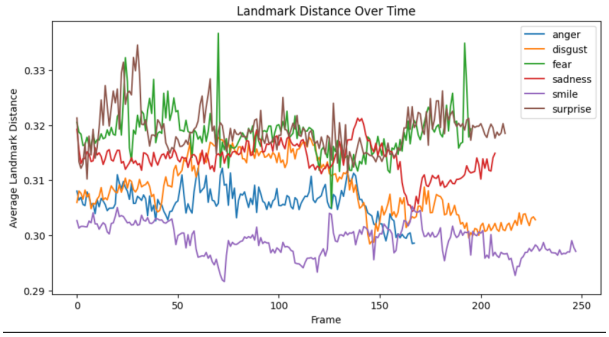
Inne podejście do frontalizacji twarzy zakłada realizację operacji na poziomie zdjęcia (tekstury) twarzy, przed ekstrakcją cech charakterystycznych. Uzyskany w ten sposób obraz wyjściowy może następnie zostać przepuszczony przez dalsze etapy potoku przetwarzania danych tak jak nieznormalizowane zdjęcie, ale z przewagą prostopadłego kąta padania twarzy do kamery.

Rysunek 4. Frontalizacja twarzy pochodząca z pracy [10]

Działanie algorytmu opiera się na detekcji kilku punktów na obrazie wejściowym, rzutowaniu ich na model referencyjny twarzy w *3D* a następnie odpowiednim wygładzeniu powstałych artefaktów przy użyciu symetrii.

Na podstawie położenia znormalizowanych punktów charakterystycznych w czasie, można dostrzec ciekawe zależności w odniesieniu do emocji, które przedstawiały nagrania. Wideo, prezentujące różne ekspresje osób zostały przepuszczone przez omówiony powyżej strumień przetwarzania, w wyniku czego uzyskano szereg czasowy współrzędnych punktów charakterystycznych. Po obliczeniu średniej odległości punktów od punktu centralnego w każdej klatce filmu, uzyskano wykres, przedstawiony na rysunku [5].

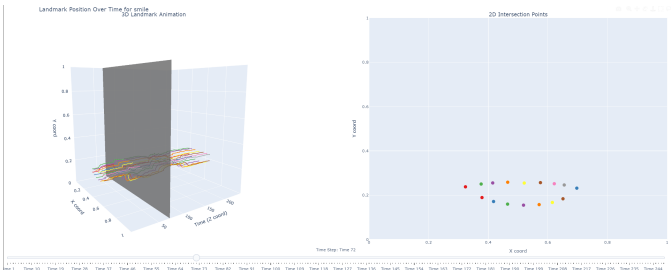
Z powyższego wykresu można wywnioskować, że najbardziej dynamiczne zmiany występują w nagraniu prezentującym strach i zaskoczenie, co jest przewidywalne - podobnie jak to, że smutek cechuje najmniejsza amplituda wartości. Na tej podstawie można założyć, że w szeregu czasowym odległości punktów charakterystycznych zawarte są informacje o przeżywanych emocjach, co może



Rysunek 5. Odległość punktów charakterystycznych w czasie dla różnych emocji (grafika autorska)

zostać wykorzystane do ostatecznego wyboru rodzaju danych wejściowych dla modelu.

W celu uzyskania większej ilości informacji o zależności między przeżywanymi emocjami a pozycją punktów charakterystycznych w czasie, można wykorzystać pozycje tychże punktów na płaszczyźnie *3D*, uwzględniając w nim dodatkowy wymiar - czas. Przedstawiona na rysunku [6] grafika pochodzi z autorskiego programu, który pozwala na przemieszczanie płaszczyzny rzutowania po osi czasu i wizualizację przemieszczenia punktów charakterystycznych.



Rysunek 6. Położenie punktów ust w czasie dla różnych emocji (grafika autorska)

Dzięki takiej reprezentacji danych można upewnić się, że znormalizowane punkty zachowują się zgodnie z przewidywaniami a dodatkowo, zestawiając ze sobą punkty różnych części twarzy jesteśmy w stanie wykryć zależności charakterystyczne dla danych emocji a nawet mikroekspresji.

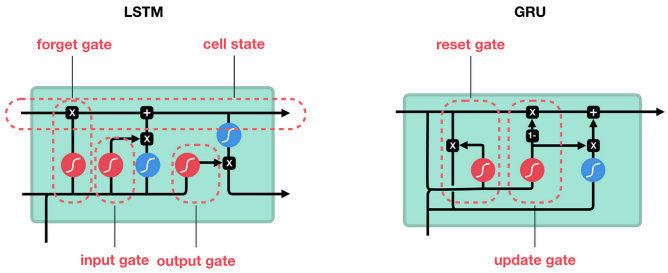
V. Modele detekcji kłamstw

Następnym krokiem, po zrealizowaniu strumienia przetwarzania danych, jest wybór odpowiedniego modelu docelowego do zdefiniowanego zadania.

Specyfika danych sprawia, że naturalnym kandydatem do rozwiązania problemu są **neuronowe sieci rekurencyjne**. Jest to rodzaj sieci neuronowych, wykorzystywany do pracy z danymi sekwencyjnymi. Wyróżnia je posiadanie swojego rodzaju *pamięci* - zachowują kontekst pomiędzy kolejnymi klatkami w filmie. Różne rodzaje sieci rekurencyjnych charakteryzują się różną zdolnością pamiętania

cech długotrwałych i krótkotrwałych, dotyczących sekwencji. Najcenniejsza informacja - mikroekspresje, charakteryzuje się bardzo niskim czasem trwania, w związku z czym cenniejszym kryterium przy wyborze modelu będzie umiejętność wykrywania cech zawartych w krótkich podsekwencjach.

Jeżeli chodzi o główne rodzaje sieci rekurencyjnych, to najczęściej przywołuje się *LSTM* [11] (ang. *Long Short Term Memory*) i *GRU* [12] (ang. *Gated Recurrent Unit*).



Rysunek 7. Porównanie LSTM i GRU [13]

Pierwsza z wymienionych to zaawansowana wersja rekurencyjnej sieci neuronowej, wprowadzająca koncept *3* bramek, umożliwiających lepsze zapamiętywanie i zapominanie zbędnych informacji. Na schematycznym rysunku [7] widzimy kolejno bramkę zapominającą (ang. *forget gate*), za pomocą której obliczane jest jaka część zapamiętanych informacji powinna zostać zapomniana, bramkę wejściową (ang. *input gate*) - decydującą, które informacje z obecnego kroku czasowego powinny zostać zapisane w pamięci krótkotrwałej oraz bramkę wyjściową (ang. *output gate*), za pomocą której kontrolowane są informacje przekazywane do kolejnych kroków czasowych. Taka struktura pozwala sieciom *LSTM* na efektywne przechowywanie informacji przez długi czas, co czyni je odpowiednimi do pracy z długimi sekwencjami.

Sieci *GRU* są podobne w działaniu do *LSTM*, ale dzięki uproszczonej strukturze wykazują lepsze wyniki w trenowaniu przy użyciu mniejszej ilości danych. Przedstawione na schematycznym rysunku [7] bramki służą do kontrolowania odpowiednio tego, które z zapisanych w pamięci informacji powinny zostać zapomniane (*reset gate*) oraz jaka część informacji z obecnego kroku czasowego powinna zostać zapisana (*update gate*). Mimo mniejszej złożoności, *GRU* wykazuje podobną skuteczność do *LSTM* i lepiej sprawdza się w zadaniach wymagających wykrywania cech z krótkich sekwencji.

Wymienione rozwiązania pochodzą z dziedziny *uczenia głębokiego* (ang. *Deep Learning*) wykorzystującej sztuczne sieci neuronowe do analizy dużych i złożonych zbiorów danych. Niewykluczone jest, że do zadania zostaną też wykorzystane inne architektury, pochodzące na przykład z dziedziny *uczenia maszynowego* (ang. *Machine Learning*), czyli na przykład *maszyny wektorów nośnych* czy *drzewa decyzyjne*, ale na obecną chwilę dziedzina problemu wydaje się być zbyt skomplikowana na wykorzystanie mniej złożonych algorytmów.

Wybór odpowiedniej architektury, a następnie jej optymalizacja jest jednym z głównych zadań opisywanej pracy magisterskiej, ale w momencie pisania niniejszego artykułu, nie został jeszcze zrealizowany.

VI. Podsumowanie

Dziedzina detekcji kłamstwa na podstawie obrazu wideo jest obszerna i wymaga zapoznania z zagadnieniami dla niej charakterystycznymi.

Najważniejszym etapem projektu jest stworzenie potoku przetwarzania danych, za pomocą którego nastąpi wstępna ekstrakcja elementów niosących najwięcej informacji o intencji osób nagrywanych. Etap ten służy również normalizacji danych pochodzących z różnych źródeł (a ostatecznie z obrazu w czasie rzeczywistym), tak aby model nie miał trudności z klasyfikacją osób w złym oświetleniu, obróconych względem kamery lub częściowo zasłoniętych. Informacje zebrane w tym etapie, pochodzące np. z analizy położenia w czasie punktów charakterystycznych twarzy posłużą do wyboru ostatecznego formatu danych wejściowych dla modelu.

Wybór odpowiedniej do zadania architektury jest kolejną, istotną częścią problemu, który musi zostać rozwiązany w opisywanej pracy magisterskiej. Przystosowanie *rekurencyjnych sieci neuronowych* do pracy z szeregami czasowymi sprawia, że są one najbardziej obiecującymi kandydatami do tego zadania. Niewykluczone jest jednak wykorzystanie mniej zaawansowanych rozwiązań.

Literatura

[1] B. Martinez and M. F. Valstar. “Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition”. In: Handbook of Pattern Recognition and Computer Vision. Springer International Publishing, 2016, pp. 63–100.

[2] E. A. Haggard and K. S. Isaacs. “Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy”. In: The Study of Emotion. Springer US, 1966, pp. 154–165.

[3] M. Valstar et al. “The first facial expression recognition and analysis challenge”. In: IEEE International Conference on Automatic Face Gesture Recognition and Workshops. 2011, pp. 921–926.

[4] R. Gross et al. “Multi-pie”. In: 8th IEEE International Conference on Automatic Face and Gesture Recognition. 2008, pp. 1–8.

[5] Sefik Ilkin Serengil and Alper Ozpinar. “LightFace: A Hybrid Deep Face Recognition Framework”. In: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE. 2020, pp. 23–27. doi: 10.1109/ASYU50717.2020.9259802. url: https://ieeexplore.ieee.org/document/9259802.

[6] Paul Viola and Michael J Jones. “Robust real-time face detection”. In: International Journal of Computer Vision 57.2 (2004), pp. 137–154.

[7] Sefik Serengil. A Gentle Introduction to Face Recognition in Deep Learning. Accessed: 2024-11-08. 2020. url: https://sefiks.com/2020/05/01/a-gentle-introductionto-face-recognition-in-deep-learning/.

[8] Google AI. MediaPipe Face Landmarker: Face landmark detection solution. [Online; accessed 9-November-2024]. 2024. url: https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker

[9] Procrustes analysis — Wikipedia, The Free Encyclopedia. [Online; accessed 9-November-2024]. 2024.

[10] Tal Hassner et al. “Effective Face Frontalization in Unconstrained Images”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, 2015.
[11] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.
[12] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP 2014. arXiv:1406.1078.
[13] Michael Phi. Illustrated Guide to LSTM’s and GRU’s: A Step by Step Explanation. Towards Data Science. [Online; accessed 9-November-2024]. Sept. 2018.