# Automatyczna detekcja kłamstwa na podstawie dynamicznej analizy ekspresji twarzy z wykorzystaniem metod głębokiego uczenia

Bartosz Latosek i dr inż. Krystian Radlak

Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska, Warszawa, Polska
https://www.elka.pw.edu.pl

**Streszczenie** Niniejsza praca przedstawia podejście do automatycznej detekcji kłamstw w oparciu o dynamiczną analizę ekspresji twarzy z wykorzystaniem metod głębokiego uczenia. Badania bazują na opracowanym strumieniu przetwarzania danych do ekstrakcji i normalizacji cech pochodzących z ruchów twarzy z nagrań wideo. W pracy wykorzystano trzy zbiory danych dotyczące kłamstwa: *Silesian Deception Dataset*, *Miami University Deception Detection Database* oraz *UvA-NEMO Smile Database*. Eksperymenty z modelami *PyTorch* i *TodyNet* wykazały skuteczność w klasyfikacji autentyczności uśmiechu, ale ograniczoną użyteczność w ogólnej detekcji kłamstw. Szczegółowa analiza z wykorzystaniem *PCA* ujawniła wyraźną separację cech dla klasyfikacji uśmiechów, ale brak zauważalnych wzorców w detekcji kłamstw w scenariuszach rozmów. Wyniki sugerują, że mikro-ruchy twarzy mogą być skutecznym wyznacznikiem autentyczności uśmiechu, ale nie są wystarczające do szerszych zastosowań w wykrywaniu kłamstw, prawdopodobnie z powodu niskiej imersji emocjonalnej osób w badanych zbiorach danych.

**Słowa kluczowe:** detekcja kłamstw · analiza ekspresji twarzy · głębokie uczenie · klasyfikacja szeregów czasowych · wizja komputerowa

## 1 Wprowadzenie

Zagadnienie detekcji kłamstw towarzyszy ludzkości od wieków. Najstarsza odnotowana metoda pochodzi z Chin i jest datowana na rok 1000 p.n.e. Polegała ona na umieszczeniu garści ryżu w ustach podejrzanego – jeśli po przesłuchaniu ryż pozostawał suchy, zostawał on uznany za winnego. Z upływem czasu metody rozpoznawania nieszczerości stawały się coraz mniej prymitywne a zarazem skuteczniejsze.

Przełomem było wprowadzenie w latach 20. XX wieku wariografu przez J.A. Larsona, mierzącego fizjologiczne wskaźniki stresu [1]. Mimo popularności, wariograf okazał się zawodny, co skłoniło do poszukiwań nowych metod - pierwsze prace nad automatyzacją pojawiły się już w 1978 roku [2].

Badania z 2006 roku [3] wykazały, że ludzie rozpoznają kłamstwa z dokładnością jedynie 54%. Uzyskany wynik jest jedynie nieznacznie lepszy od rzutu

monetą, co pokazuje konieczność opracowywania i doskonalenia narzędzi wspierających zautomatyzowaną detekcję kłamstwa.

## 2 Bazy danych

W badaniach wykorzystane zostały trzy zbiory danych: (1) *Silenian Deception Dataset* [4], (2) *Miami University Deception Detection Database* [5] oraz (3) *UvA-NEMO Smile* [6].

Dwie pierwsze z wymienionych pozycji zawierają nagrania studentów wypowiadających zdania, będące prawdą bądź kłamstwem. Zbiór danych *Miami* jest zbalansowany i zawiera po 120 próbek przypisanych do każdej z klas. Z kolei dane pochodzące z *Silesian Deception Dataset*, po podziale nagrań na pojedyncze próbki, tworzą zbiór niezbalansowany, w którym dominuje klasa "kłamstwo". Baza *UvA-NEMO Smile* składa się z wideo przedstawiających osoby w różnym wieku, które uśmiechają się w sposób szczery bądź wymuszony.

Wykorzystane zbiory danych charakteryzują się znaczną różnorodnością - nagrania różnią się kątem ustawienia kamery, warunkami oświetleniowymi, fizjonomią twarzy oraz ekspresywnością osób. Różnice te wykazały konieczność opracowania specjalnego procesu przetwarzania, którego zadaniem jest normalizacja nagrań i ekstrakcja kluczowych cech, niezbędnych w późniejszym trenowaniu modeli.

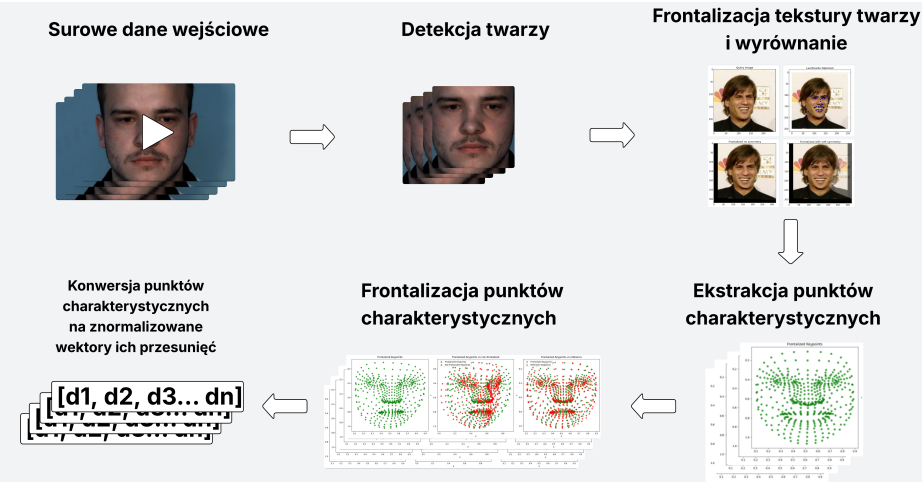
## 3 Strumień przetwarzania danych

W ramach implementacji strumienia przetwarzania skoncentrowano się na integracji zróżnicowanych metod z zakresu wizji komputerowej, które w połączeniu tworzą uniwersalny mechanizm umożliwiający transformację nieprzetworzonych danych wejściowych w uporządkowane sekwencje istotnych cech. Schemat działania stworzonego rozwiązania przedstawia rysunek 1.

Proces przetwarzania obejmuje kolejno: detekcję twarzy w każdej klatce wideo, frontalizację i wyrównanie twarzy w trzech płaszczyznach, ekstrakcję punktów charakterystycznych przy użyciu modelu z pakietu *mediapipe* [7], dodatkową frontalizację punktów metodą analizy Prokrustesa [8] oraz konwersję punktów na wektory przesunięcia względem centralnego punktu twarzy, co zmniejsza wymiarowość danych i zwiększa odporność modelu na zmiany pozycji. Dodatkowym krokiem jest normalizacja przesunięć względem pierwszej klatki, w wyniku czego uzyskujemy ciąg zależności skupiających się na stosunkach przesunięć punktów w czasie, zamiast na ich konkretnych wartościach.

## 4 Testy strumienia przetwarzania

Przed zastosowaniem strumienia na docelowych bazach danych, a następnie wykorzystaniem przetworzonych danych do treningu modeli detekcji kłamstw, zdecydowano zweryfikować jakość uzyskanego rozwiązania na pokrewnym proble-




**Rysunek 1.** Schemat strumienia przetwarzania danych

mie. W tym celu wykorzystano bazę *Ravdess* [9], zawierającą nagrania osób wyrażających jedną z ośmiu podstawowych emocji za pomocą mowy i śpiewu. W założeniu, takie działanie pozwoli na potwierdzenie poprawności działania strumienia oraz analizę korelacji wyekstraktowanych cech z docelową klasą danej próbki.

W eksperymencie przetestowano różne konfiguracje strumienia przetwarzania danych, które posłużyły do treningu prostego klasyfikatora. Model stworzono w bibliotece *PyTorch* [10], wykorzystując prostą architekturę, w postaci warstwy konwolucyjnej z *poolingiem* do ekstrakcji cech oraz dwie warstwy *LSTM*[1] do analizy danych sekwencyjnych. Wyniki uzyskane dla różnych konfiguracji przedstawiono w tabeli 1.

**Tabela 1.** Porównanie metryk dla różnych podejść do detekcji emocji.

| Podejście | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Wszystkie punkty | 0.3472 | 0.3099 | 0.3161 | 0.2989 |
| Wybrane manualnie | 0.4931 | 0.4567 | 0.4621 | 0.4538 |
| Wybrane przez wsp. Shapleya | 0.5509 | 0.5428 | 0.5460 | 0.5379 |
| Połączone (manualne + Shapley) | 0.5417 | 0.5207 | 0.5279 | 0.5153 |
| Przesunięcia punktów | **0.6343** | **0.6261** | 0.6303 | **0.6190** |
| Znormalizowane przesunięcia punktów | 0.6204 | 0.6148 | **0.6311** | 0.6186 |

Na podstawie uzyskanych wyników można zaobserwować, że podejście wykorzystujące wszystkie punkty charakterystyczne twarzy osiągnęło najniższe rezul-

[1] ang. *Long-Short Term Memory*

taty we wszystkich metrykach, co sugeruje, że nadmiar informacji i obecność szumów może negatywnie wpływać na skuteczność modelu. Redukcja liczby punktów — zarówno poprzez manualną selekcję, jak i za pomocą współczynników Shapleya[2], — znacząco poprawiła wyniki, przy czym podejście oparte wyłącznie na wsp. Shapleya okazało się bardziej efektywne niż wariant łączony z manualnym wyborem. Najlepsze rezultaty osiągnięto jednak przy transformacji danych do postaci przesunięć punktów względem środka twarzy. Znormalizowane przesunięcia uzyskały bardzo zbliżone wyniki, a dzięki uwzględnieniu indywidualnych różnic w fizjonomii i ekspresyjności twarzy, podejście to uznano za najbardziej optymalne.

Za pomocą przeprowadzonego eksperymentu jednoznacznie wykazano, że informacje zawarte w punktach charakterystycznych twarzy oraz ich przesunięciach w czasie pozwalają na trening modelu klasyfikującego emocje. Najlepiej poradziły sobie modele trenowane na danych w formie szeregów czasowych odległości punktów od środka twarzy, w związku z czym będzie to główne podejście stosowane w dalszych rozważaniach.

## 5 Detekcja kłamstw

Kolejnym etapem prowadzonych badań była próba wytrenowania modeli klasyfikacyjnych służących do detekcji kłamstw, z wykorzystaniem docelowych baz danych omówionych w sekcji 2. Eksperymenty przeprowadzono na dwóch modelach: (1) Prosty model *PyTorch* oraz (2) *TodyNet* [11], czyli dynamiczna sieć grafowa, przystosowana do klasyfikacji wielowymiarowych szeregów czasowych.

Na potrzeby testów wykorzystano dane w postaci znormalizowanych wektorów przesunięć punktów charakterystycznych twarzy względem punktu centralnego. Dodatkowo, w ramach eksperymentu kontrolnego, zastosowano także zestaw punktów wybranych uprzednio w analizach związanych z klasyfikacją emocji.

### 5.1 Model *PyTorch*

W tej sekcji eksperymentu wykorzystany został model analogiczny do opisywanego w sekcji 4, rozbudowany o dodatkową warstwę konwolucyjną. Ponadto, jedna z warstw typu *LSTM*, została zastąpiona warstwą uwagi (ang. *attention*), co pozwoliło na bardziej selektywne uwzględnianie istotnych fragmentów szeregów czasowych.

Zagregowane wyniki eksperymentu przedstawia tabela 2. Skuteczność modelu *PyTorch* znacząco zależy od charakterystyki zbioru danych — podczas gdy na zbiorze *UvA-NEMO Smile Database* model osiągnął satysfakcjonujące wyniki, na zbiorach dotyczących detekcji kłamstwa (*Miami* oraz *Silesian*) klasyfikacja okazała się nieskuteczna. Zaobserwowane zachowania modelu, takie jak klasyfikowanie wszystkich próbek jako jednej klasy, mogą świadczyć o niedostatecznej

[2] Współczynnik Shapleya to miara wyjaśnialności modelu oparta na teorii gier, pozwalająca ocenić wkład poszczególnych cech w decyzję modelu.

**Tabela 2.** Porównanie wyników modelu *pytorch* dla różnych rodzajów danych wejściowych na trzech zbiorach danych

| Zbiór | Podejście | Accuracy | Precision | Recall | F1-score | TP-rate | FP-rate |
|---|---|---|---|---|---|---|---|
| Miami | Wybrane pkt. | 0.52 | 0.26 | 0.50 | 0.34 | 1.00 | 1.00 |
| | Przesunięcia pkt. | 0.52 | 0.26 | 0.50 | 0.34 | 1.00 | 1.00 |
| Silesian | Wybrane pkt. | 0.28 | 0.14 | 0.50 | 0.23 | 0.00 | 0.00 |
| | Przesunięcia pkt. | 0.71 | 0.35 | 0.50 | 0.41 | 1.00 | 1.00 |
| UvA-NEMO | Wybrane pkt. | 0.68 | 0.69 | 0.69 | 0.68 | 0.77 | 0.39 |
| | Przesunięcia pkt. | 0.71 | 0.71 | 0.70 | 0.70 | 0.63 | 0.21 |

złożoności zastosowanej architektury w kontekście zadania detekcji kłamstwa. W związku z tym, w kolejnym etapie badań zdecydowano się powtórzyć eksperyment z wykorzystaniem bardziej zaawansowanej architektury.

### 5.2 Model *TodyNet*

Sieć *TodyNet* to dynamiczna grafowa sieć neuronowa, zaprojektowana do klasyfikacji wielowymiarowych szeregów czasowych poprzez modelowanie zmiennych jako węzłów grafie z dynamicznie ewoluującymi relacjami. Autorzy wykazali, że TodyNet skutecznie wydobywa ukryte zależności czasowo-przestrzenne i przewyższa inne metody głębokiego uczenia na 26 zbiorach danych UEA[3].

W związku z oficjalną implementacją architektury *TodyNet*, jakość rozwiązania została wyznaczona przez średnia wartość dokładności na przestrzeni epok treningowych modelu. Eksperymenty zostały przeprowadzone na danych w postaci znormalizowanych przesunięć punktów charakterystycznych w czasie.

**Tabela 3.** Porównanie wyników modelu *TodyNet* dla przesunięć pkt. charakterystycznych na trzech zbiorach danych

| Zbiór | Średnia dokładność |
|---|---|
| Miami | 0.49 |
| Silesian | 0.70 |
| Uva-Nemo | 0.74 |

Na podstawie wyników przedstawionych w tabeli 3 można jednoznacznie stwierdzić, że model *Todynet* okazał się nieskuteczny w zadaniu detekcji kłamstwa na zbiorach *Miami* oraz *Silesian*, jednak poradził sobie lepiej od prostego modelu *PyTorch* w problemie klasyfikacji autentyczności uśmiechu na bazie *Uva-NEMO Smile*.

[3] ang. *University of East Anglia*

### 5.3 Wnioski

Na podstawie przeprowadzonych eksperymentów z wykorzystaniem prostego modelu w *PyTorch* oraz bardziej złożonego modelu *Todynet*, można sformułować następujące wnioski:

1. **Modele nie poradziły sobie z detekcją kłamstwa** – zarówno *PyTorch*, jak i *TodyNet* osiągnęły niską skuteczność na zbiorach *Miami* i *Silesian*, często przewidując klasę losowo lub zgodnie z dominującą klasą.
2. **Modele skutecznie rozpoznawały autentyczność uśmiechu** – na zbiorze *UvA-NEMO Smile Database* oba modele osiągnęły dobre wyniki, co wskazuje na istnienie zależności między mimiką twarzy a szczerością uśmiechu.

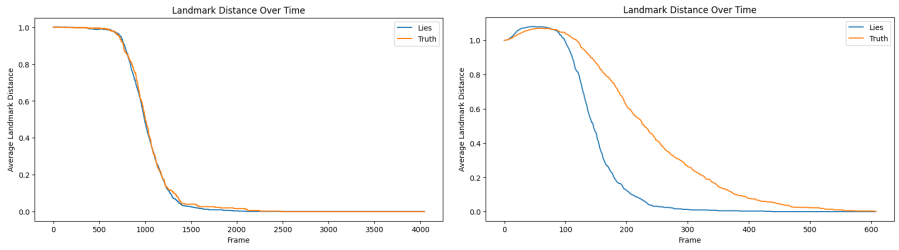
## 6 Analiza danych

Ze względu na niską jakość modeli klasyfikacyjnych uzyskanych w sekcji 5, przeprowadzono analizę korelacji pomiędzy wyekstrahowanymi cechami a klasą docelową dla wybranych zbiorów danych. Ponieważ wyniki dla zbiorów *Miami Deception* i *Silesian Deception* były zbliżone, w analizie wykorzystano jedynie wykresy dotyczące zbioru *Silesian Deception*.

### 6.1 Analiza średnich przesunięć punktów charakterystycznych w czasie

Rysunek 2 przedstawia uśrednione przesunięcia wybranych punktów charakterystycznych w czasie, w zależności od klasy. Dla każdej próbki, przesunięcia zostały znormalizowane względem wartości w pierwszej klatce, a następnie uśrednione po wszystkich punktach charakterystycznych w danej klatce nagrania.

Dla zbioru *UvA-NEMO Smile* zaobserwowano wyraźne różnice w dynamice ruchu punktów charakterystycznych twarzy w zależności od autentyczności uśmiechu, co potwierdza istnienie zależności między cechami a klasą. W przypadku zbiorów *Miami* i *Silesian* takie różnice nie wystąpiły, a trajektorie punktów dla obu klas były niemal identyczne.



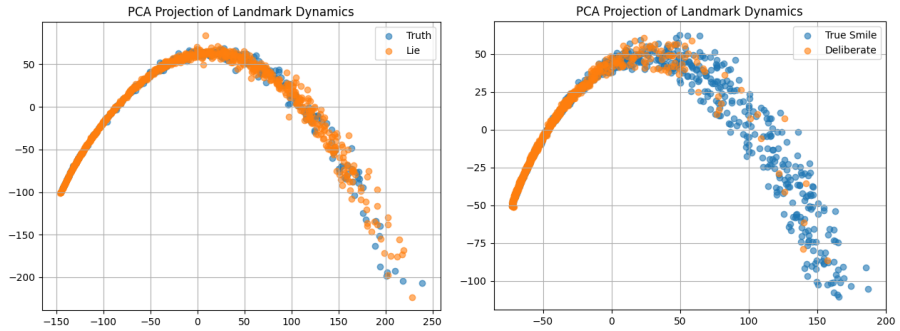

**(a)** Silesian Deception Dataset **(b)** UvA-NEMO Smile Database

**Rysunek 2.** Porównanie średnich przesunięć punktów charakterystycznych

### 6.2 Analiza głównych składowych

Metoda *PCA* [12] (ang. *Principal Component Analysis*) to technika statystyczna służąca do redukcji wymiarowości danych. Przekształca dane do nowego układu współrzędnych, zachowując przy tym najistotniejsze informacje, co ułatwia analizę i wizualizację złożonych zależności.

Na rysunku 3 przedstawiono projekcje danych po redukcji wymiarowości metodą *PCA*, które potwierdzają wyraźne rozdzielenie między uśmiechami autentycznymi i wymuszonymi w zbiorze *UvA-NEMO Smile*. W przypadku zbioru *Silesian* brak jest takiej separacji, co wskazuje na niską informacyjność cech lub zbyt małą ilość danych w zadaniu detekcji kłamstwa.




**(a)** Silesian Deception Dataset **(b)** UvA-NEMO Smile Database

**Rysunek 3.** *PCA* dla średnich przesunięć punktów charakterystycznych

## 7 Podsumowanie

W pracy zaprezentowano podejście do automatycznej detekcji kłamstwa oparte na dynamicznej analizie ekspresji twarzy z wykorzystaniem metod głębokiego uczenia.

Kluczowym elementem było opracowanie uniwersalnego strumienia przetwarzania danych, który pozwala na ekstrakcję i normalizację przesunięć punktów charakterystycznych twarzy. Za jego pomocą następnie przeprowadzono szereg eksperymentów które miały na celu zbadanie zależności pomiędzy ruchami twarzy a prawdomównością osoby nagrywanej. Na podstawie badań wykazano, że przesunięcia punktów charakterystycznych stanowią wartościowe cechy do klasyfikacji emocji oraz autentyczności uśmiechu, ale nie są dobrym wyznacznikiem w przypadku detekcji kłamstwa na bazach *Silesian* oraz *Miami*.

Wymienione bazy danych zostały poddane analizie pod kątem zbadania relacji pomiędzy wyekstraktowanymi cechami a intencjami osoby nagrywanej i wykazano, że podczas gdy zależność ta jest zauważalna w problemie klasyfikacji autentyczności uśmiechu, nie występuje ona w pozostałych bazach danych.

Potencjalną przyczyną zaobserwowanego zjawiska może być niska imersja emocjonalna studentów uczestniczących w tworzeniu zbiorów danych. Uczniowie nie ponosili konsekwencji za wypowiadane kłamstwa, co skutkowało mniejszym zaangażowaniem w próbę ukrycia prawdy. W efekcie brakowało u nich typowych ekspresji i mikroekspresji, które często występują u osób przesłuchiwanych np. w warunkach sądowych.

## Literatura

1. Larson, J.A.: Modification of the Marston Deception Test. Journal of the American Institute of Criminal Law and Criminology, Northwestern University School of Law.
2. Suwa, M., Sugie, N., Fujimora, K.: A preliminary note on pattern recognition of human emotional expression.
3. Bond, C.F., DePaulo, B.M.: Accuracy of deception judgments. *Personality and Social Psychology Review, SAGE Publications Sage CA: Los Angeles, CA.
4. Radlak, K., Bożek, M., Smolka, B.: Silesian Deception Database: Presentation and Analysis.
5. Lloyd, E.P., Deska, J.C., Hugenberg, K., et al.: Miami University deception detection database.
6. Dibeklioglu, H., Salah, A. A., and Gevers, T. (2012). "Are you really smiling at me? Spontaneous versus posed enjoyment smiles," in Proceedings of the European Conference on Computer Vision, Berlin: Springer
7. Google Research, *MediaPipe: Open-source framework for multimodal applied machine learning*
8. Wikipedia contributors. *Procrustes analysis*. 2024. Available at: https://en.wikipedia.org/wiki/Procrustes_analysis
9. Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English.
10. A. Paszke et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*
11. Huaiyuan Liu, Donghua Yang, Xianzhang Liu, Xinglei Chen, Zhiyu Liang, Hongzhi Wang, Yong Cui, and Jun Gu. TodyNet: Temporal dynamic graph neural network for multivariate time series classification. *Information Sciences*, August 2024.
12. I. Jolliffe, *Principal Component Analysis*, in: Encyclopedia of Statistics in Behavioral Science, pp. 1094–1096, 2011.