POLITECHNIKA WARSZAWSKA

DYSCYPLINA NAUKOWA: INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA
DZIEDZINA NAUK: NAUKI INŻYNIERYJNO-TECHNICZNE

Rozprawa doktorska

mgr inż. Weronika Gutfeter

Identyfikacja twarzy na podstawie obrazów wieloujęciowych z zastosowaniem głębokich sieci agregujących

Promotor
prof. dr hab. inż. Andrzej Pacut

WARSZAWA 2022

Streszczenie

W pracy zaproponowano metody rozwiązania problemu tworzenia wzorców biometrycznych twarzy na podstawie zdjęć wieloujęciowych. Przez zdjęcia wieloujęciowe rozumiane są zbiory obrazów twarzy jednej osoby, w których głowa jest ujęta pod różnymi kątami, przy czym zestaw kątów jest zdefiniowany arbitralnie. Przykładem tego typu danych są zdjęcia sygnalityczne wykonywane przez fotografów policyjnych oraz obrazy uzyskane z modeli 3D głowy poprzez ich projekcję na płaszczyznę. Potrzeba rozwiązania tak zdefiniowanego problemu została umotywowana obserwacjami rzeczywistych systemów identyfikacji - zauważono, że wiele z algorytmów biometrii twarzy jest zoptymalizowana pod kątem scenariusza identyfikacji na podstawie zdjęć frontalnych i nie wykorzystuje dostatecznie ujęć skrajnych, takich jak zdjęcia profilowe.

W ramach badań zanalizowano oraz zaimplementowano szereg podejść do identyfikacji zbiorów obrazów twarzy, w których występują zdjęcia niefrontalne, w tym poprzez agregację deskryptorów pojedynczych ujęć (wzorce wytworzone przez modele single-view), poprzez modele agregujące (wzorce wytworzone przez modele multi-view) oraz automatyczną korektę pozy w obrazie. Na potrzeby eksperymentów zaadaptowano modele agregujące stosowane dotąd do klasyfikacji obiektów 3D, w tym model MVCNN (*multi-view convolutional network*) [1] oraz model RotationNet [2]. W kolejnym kroku zaproponowano własny model agregujący z mechanizmem uwagi o nazwie SygnaT. W pracy pokazano, że wszystkie 3 modele agregujące (multi-view) uzyskały lepsze wyniki niż agregacja modeli pojedynczych (single-view). Najwyższe współczynniki dokładności identyfikacji uzyskano dla modelu SygnaT. Była to różnica o ponad 6% dla współczynnika Rank-1 w porównaniu z uśrednionymi deskryptorami single-view oraz o 18% dla współczynnika Rank-1 wyliczonego dla identyfikacji na podstawie zdjęć profilowych.

Wszystkie rozważane podejścia były oparte o głębokie sieci splotowe i zakładały budowę modułową. Jako rdzeń analizowanych rozwiązań przyjęto ten sam enkoder single-view oparty na architekturze ResNet-50 [3] i wytrenowany na bazie VGGFace2 [4] przez autorów pracy. Przyjęcie wspólnego rdzenia ułatwiło miarodajne zestawienie różnych metod ze sobą, a przez modularną strukturę pokazano, że ta część sieci może być modyfikowana lub zamieniona na inną wg preferencji projektantów systemu biometrycznego. Przy omawianiu wyników wzięto pod uwagę różne scenariusze identyfikacji twarzy, w tym również w scenariuszu ze zbiorem otwartym, w którym możliwa jest rejestracja nowych osób bez konieczności ponownego treningu modeli.

Słowa kluczowe: identyfikacja twarzy, biometria, sieci głębokie, sieci agregujące, sieci splotowe, MVCNN, RotationNet, multi-head attention, atencja, rozpoznawanie twarzy 3D, fotografia sygnalityczna

Abstract

The main goal of the thesis is to solve the problem of face identification in the multi-view face images. Datasets that consist of multi-view face images can be obtained in various scenarios. One of the typical usage is the police booking photography, where the specialists are obligated to acquire a set of face images from strictly defined angles for each suspect. Another scenario is also one of the methods employed in the 3D face recognition. As the 3D models are quite heavy and computationally demanding, a typical solution to that problem is to transform 3D data into a set of 2D images showing the object from different views. This approach can be named as multi-view object classification.

After preliminary experiments we observed that most of the state-of-the-art algorithms are optimized to recognize frontal images and there is a significant drop of accuracy when identification is made on sets containing extreme views like full profile pictures. In our work we conducted a survey of methods that can be used for creating biometric templates from image collections with non-frontal faces. We analyzed various approaches like 3D face alignment, single-view templates aggregation and multi-view networks. For our experiments we adapt methods that have been proposed for multi-view object recognition, namely MVCNN [1] and RotationNet [2]. However we should emphasize that there exists a difference in these two domains. Multi-view object recognition is typically based on a closed dictionary of labels and objects are represented by a relatively large number of views while faces are organized in the sets of three or five probes and the possibility of new identifier registration is crucial for the system. For the purpose of the experiments it was necessary to collect face data that is structured in the following manner and to define adequate testing scenarios. We also introduce a new multi-view model in which we apply multi-head attention for data flow aggregation. We named it SygnaT. SygnaT solves some of the limitations of the previous models, as it does not

require a strict number of probes or specific order in the view set. All of the proposed multi-view networks gain higher results than the single-view approaches. For the SygnaT network the Rank-1 accuracy is higher by 6% in a full identification test and by 18% in a test on profile pictures.

Each of the analyzed solutions was built using deep convolutional networks and there is a common backbone for all of them. The backbone is implemented with ResNet-50 [3] architecture. Therefore weights can be transferred from the baseline single-view model which was trained on VGGFace2 [4] to make all the models more unified and easily comparable. We showed that the multi-view networks work in various face identification scenarios, we tested all solutions in closed and open-set tests. Moreover we conducted an experiment with SygnaT on an unstructured dataset (face recognition benchmark IJB-C [5]) and proved that also in this configuration it is better than the single-view model approach.

Keywords: face biometrics, facial identification, deep learning, multi-view classification, multi-view networks, convolutional neural networks, MVCNN, RotationNet, multi-head attention, self-attention, face 3D recognition, police photography

Spis treści

1 Wprowadzenie **1**

1.1 Motywacja 1

1.2 Tezy 3

1.3 Systemy identyfikacji twarzy 4

1.4 Ewaluacja systemów biometrycznych 6

1.5 Eksperckie metody identyfikacji twarzy 9

1.6 Współczesne metody obrazowania twarzy i źródła danych 11

1.7 Bazy danych i scenariusze testowe 21

2 Identyfikacja twarzy w oparciu o sieci głębokie **35**

2.1 Generacja wzorców twarzy z użyciem sieci splotowych 35

2.2 Bazowy model single-view 41

3 Metody stosowane w identyfikacji obrazów wieloujęciowych twarzy **55**

3.1 Wybrane podejścia 55

3.2 Automatyczna korekta pozy 58

3.3 Agregacja modeli 65

3.4 Modele agregujące 72

4 Modele własne do agregacji obrazów wieloujęciowych twarzy **81**

4.1 MVCNN-Sygnalityka 81

4.2 RotationNet-Sygnalityka 90

4.3 SygnaT (Sygnalityka-Transformer) 100

5 Podsumowanie **113**

Materiały dodatkowe **119**

A Projekty powiązane z tematyką pracy 119

Bibliografia **123**

Rozdział 1

Wprowadzenie

1.1 Motywacja

Temat badań jest wynikiem dostrzeżenia luki w metodach identyfikacji twarzy, którą zauważyliśmy w trakcie rozwoju aplikacji do automatycznej analizy materiałów dowodowych w formie obrazów i wideo zbieranych przez policję (krótki opis projektu BIOWIZ znajduje się w rozdziale A „Projekty powiązane z tematyką pracy"). Obrazy, z których korzystają eksperci policyjni to najczęściej dane z różnego rodzaju systemów monitoringu, które powstają bez kooperacji z osobami widocznymi na zdjęciu. W wielu przypadkach konieczne jest porównanie próbek z bazą fotografii sygnalitycznej, która w momencie trwania projektu, była przechowywana w rejestrze KSIP (Krajowy System Identyfikacji Policyjnej). Zdjęcia sygnalityczne składają się z zestawu obrazów głowy, w których poszczególne ujęcia wykonane są z określonych kątów. Oprócz portretów frontalnych zestaw taki zawiera zdjęcia profilowe i półprofilowe twarzy.

Mimo intuicji, że wzorce tworzone na podstawie zbioru zdjęć o zróżnicowanych ujęciach powinni być silniejsze niż wzorce wygenerowane jedynie ze zdjęć frontalnych, odnotowaliśmy wyraźny spadek jakości wyszukania dla zdjęć o skrajnych ujęciach, kiedy używaliśmy metod rozpoznawania twarzy bez zaproponowanych w pracy modyfikacji. Co ciekawe zdjęcia wieloujęciowe wykorzystywano także w pierwszych katalogach portretów osób podejrzanych, które były zbierane przez uważanego za twórcę policyjnych systemów identyfikacyjnych i na-

uki zwanej antropometrią Alphonse'a Bertillona. Tak więc szukanie możliwości zapisania cech twarzy z uwzględnieniem zmienności przestrzennej towarzyszyło systemom identyfikacji biometrycznej praktycznie od początku ich istnienia.

Jednym z rozwiązań problemu jest wykorzystanie w systemie obrazowania 3D. Rejestracja tego typu danych wymaga zazwyczaj dedykowanych urządzeń, takich jak skanery 3D lub kamery głębi. Sama akwizycja obrazów jest stosunkowo trudna i pracochłonna, a na przetwarzanie danych 3D trzeba zarezerwować więcej zasobów obliczeniowych. Autorka pracy zaproponowała we wcześniejszych badaniach mobilny zestaw do pobierania i rozpoznawania twarzy 3D oparty na tablecie komputerowym i przenośnej kamerze głębi Structure.io. Zastosowane metody zostały opisane w artykule [6]. Dzięki skonstruowaniu własnego systemu 3D można było dokonać parę praktycznych obserwacji z użytkowania tego typu danych. O ile sama dziedzina wciąż się rozwija, to w momencie prowadzenia badań napotkaliśmy na kilka problemów. Struktury danych przechowujących informacje przestrzenne są cięższe i wymagają więcej skomplikowanych obliczeń niż ma to miejsce w przypadku danych RGB. Co jednak najważniejsze publicznie dostępne bazy 3D są nieporównywalnie mniejsze niż zbiory zdjęć 2D i stanowiło to blokadę przed wytrenowaniem modeli o porównywalnych wynikach, do tych, które osiągają modele 2D.

Celem naszej pracy było więc zaprojektowanie metody, która będzie potrafiła wykorzystać informację o zmienności przestrzennej twarzy, a jednocześnie nie będzie wymagała drogiej i specjalistycznej aparatury. Inspiracje dla prowadzonych prac stanowiły artykuły dotyczące modeli agregujących wieloujęciowych [1], [2], które opisywały metody klasyfikacji obiektów 3D bez konieczności operowania na danych 3D. Modele te dokonywały klasyfikacji na podstawie zbiorów obrazów wieloujęciowych, które powstały w wyniku projekcji obiektów 3D na płaszczyzny pod różnymi kątami.

1.2 Tezy

Wybór metod stosowanych w pracy ewoluował wraz ze zmianami w dziedzinie wizji komputerowej opartej na sieciach neuronowych, która to jest dziedziną bardzo szybko rozwijaną w ostatnich latach. W trakcie trwania badań pojawiały się nowe metody, modele i biblioteki wspomagające projektowanie własnych rozwiązań. Mimo to można wyróżnić następujące tezy, które są wykazane i wyjaśnione w opracowaniu:

1. Możliwe jest zbudowanie systemu identyfikacji twarzy, który będzie uwzględniał informacje o zmienności przestrzennej twarzy i różnicy w wyglądzie w zależności od pozy bez konieczności zastosowania specjalizowanych urządzeń do akwizycji, jak skanery 3D. Można pokazać, że system, w którym wzorce twarzy są rejestrowane na podstawie zbioru obrazów wieloujęciowych będzie lepszy niż system oparty na zdjęciach frontalnych.
2. Zbiory wieloujęciowe obrazów twarzy wymagają dedykowanych rozwiązań, a składanie informacji z poszczególnych próbek można zrealizować za pomocą modułu agregującego zintegrowanego w strukturze sieci neuronowej w postaci tzw. modelu agregującego multi-view. System identyfikacji wykorzystujący modele multi-view powinien wykazać wskaźniki jakości identyfikacji nie gorsze niż system oparty o modele jednoujęciowe tzw. single-view.
3. Modele agregujące stosowane dotąd do rozpoznawania obiektów 3D należących do zamkniętego słownika obiektów (takie jak MVCNN i RotationNet) można dostosować do zadania identyfikacji biometrycznej twarzy.

1.3 Systemy identyfikacji twarzy

Identyfikacja biometryczna osoby to procedura mająca na celu znalezienie jednoznacznego identyfikatora osoby w **bazie wzorców biometrycznych** na podstawie próbki lub próbek prezentujących wybrany zestaw cech szukanej osoby. W tym miejscu należy ją odróżnić od procesu weryfikacji, której celem jest określenie zgodności próbki (lub zestawu próbek) z wzorcem osoby o zadeklarowanej a priori tożsamości. W praktyce obie procedury mogą zostać zrealizowane z zastosowaniem **funkcji dopasowania**, która na podstawie danych do porównania jest w stanie ocenić ich stopień podobieństwa.

Ocena stopnia podobieństwa może posłużyć do szeregowania listy podpowiedzi, w przypadku systemów identyfikacji, które zwracają wynik przeszukiwania bazy w postaci listy lub do sprawdzenia, czy stopień podobieństwa przekracza obrany próg akceptacji w przypadku weryfikacji tożsamości. Mimo możliwości realizacji procedur identyfikacji i weryfikacji w zbliżony sposób, dokładność systemu weryfikacji nie ma jednoznacznego przełożenia na dokładność systemu identyfikacji (fenomen ten został omówiony w artykule [7]). W związku z tym przy opracowywaniu systemu biometrycznego należy brać pod uwagę scenariusz użycia lub monitorować dokładność systemu w obydwu wariantach. Szczegóły sposobów ewaluacji sprawności systemów identyfikacji i weryfikacji oraz związane z nimi miary jakości zostaną omówione w rozdziale 1.4 „Ewaluacja systemów biometrycznych".

Systemy biometrii twarzy wykorzystują różne metody kodowania cech ludzkiej twarzy, takie jak geometria, barwa i lokalizacja punktów charakterystycznych, które uzyskiwane są z obrazów badanej osoby. Najpopularniejsze ze względu na dostępność danych są systemu bazujące na fotografii cyfrowej (obrazy 2D), aczkolwiek zarówno w literaturze jak i na rynku rozwiązań informatycznych można znaleźć metody wykorzystujące inne typy obrazowania twarzy, np. skany 3D, obrazy głębi, obrazy w świetle podczerwonym, czy zdjęcia termiczne. Metody obrazowania twarzy pod kątem biometrii zostaną omówione w rozdziale 1.6 „Współczesne metody obrazowania twarzy i źródła danych".

Biometria twarzy wraz z daktyloskopią są jednymi z szerzej stosowanych technologii identyfikacji tożsamości, szczególnie w badaniach kryminalistycznych. Jeżeli weźmiemy pod uwagę również nieautomatyczne metody rozpoznawania, to oba podejścia (rozpoznawanie twarzy i

odcisków palców) mogą się poszczycić długa historią zastosowania, o pierwszych kartotekach policyjnych ze zdjęciami twarzy z XIX wieku napisano w rozdziale 1.5 „Eksperckie metody identyfikacji twarzy". Zaletą biometrii twarzy jest stosunkowo niski koszt akwizycji danych (w przypadku klasycznej fotografii 2D) oraz brak wymogu kooperacji ze strony użytkownika - dane biometryczne można uzyskiwać np. z nagrań pochodzących z monitoringu wizyjnego. Konieczność analizy próbek, które pobierane są bez nadzoru, sprawia, że w przypadku metod identyfikacji twarzy kluczowe jest badanie wpływu rożnych czynników zmienności na dokładność identyfikacji. Czynniki, które mają największy wpływ na spadek jakości rozpoznania to: zmienność pozy, zróżnicowanie warunków oświetleniowych, mimika twarzy, zmiany wizerunku spowodowane wiekiem, chorobą, makijażem, zarostem i uczesaniem, a także zakłócenia obrazu, takie jak obecność obiektów częściowo zasłaniających twarz osoby. Spektrum zmienności jest bardzo szerokie. W poniższej pracy został zbadany przede wszystkim wpływ kąta widzenia twarzy (poza głowy) oraz możliwość wykorzystania zdjęć sygnalitycznych, czyli zbiorów złożonych z zestawów zdjęć jednej osoby uchwyconej w różnych pozach w kompleksowym systemie identyfikacji. Definicja zdjęć sygnalitycznych oraz ich charakterystyka wynikająca z rozporządzeń prawnych w Polsce została umieszczona w sekcji 1.6.2 „Zdjęcia sygnalityczne" w rozdziale dotyczącym obrazowania.

W zasadzie większość wczesnych algorytmów klasyfikacji do automatycznego rozpoznawania twarzy, takich jak *twarze własne* [8] i *twarze Fischera* [9] były bardzo czułe na zmienność warunków, w których rejestrowany był obraz twarzy i do poprawnego działania wymagało normalizacji pozycji twarzy w obrazie. Obecnie większość rozwiązań opartych jest o sieci głębokie, które są mniej wrażliwe na translacje i rotację obiektu w obrazie. W dalszym ciągu można stwierdzić jednak, że zestawianie twarzy, które są widoczne w tej samej pozie daje dokładniejsze wyniki porównania (więcej na ten temat w 2.2.4 „Poza jako czynnik wpływający na dokładność identyfikacji twarzy"). Dyskusyjnym jest, czy w przypadku modeli opartych o sieci głębokie należy poddawać obraz korekcji 3D pozycji głowy w celu poprawienia wyników działania. Niektóre źródła [10] podają, że jest to ważne przede wszystkim na etapie treningu sieci w celu wygenerowania większej liczby przykładów uczących, a nie powinno być używane przy ewaluacji - szczególnie przy skrajnie wysokich różnicach w pozie.

Obraz twarzy jest szeroko przyjęty jako metoda wizualnej identyfikacji człowieka. Zdjęcie twarzy jest integralną częścią większości dokumentów tożsamości: od identyfikatorów służ-

bowych, po dokumenty prawa jazdy, dowody osobiste i paszporty. Ze względu na czułość algorytmów rozpoznania na zmienność pozycji głowy powstały krajowe i międzynarodowe standardy dotyczące wykonywania zdjęć do dokumentów tożsamości. Przykładowo w instrukcji wydanej przez Ministerstwo Spraw Wewnętrznych i Administracji z roku 2016[11] wymaga się, żeby twarz na zdjęciu skierowana była w stronę aparatu (en face) - prostopadle do osi aparatu usytuowanej na wysokości pomiędzy nosem a linią oczu (niedopuszczalna jest tzw.„ptasia" lub „żabia" perspektywa).

Ze względu na strukturę państwowych i międzynarodowych dokumentów tożsamości jednym z powszechniej stosowanych scenariuszy identyfikacji twarzy jest **przeszukiwanie bazy wzorców frontalnych**. Jednocześnie policyjne bazy sygnalityczne zawierają w zestawie zdjęć osoby podejrzanej również zdjęcia profilowe. W kontraście do pierwszego scenariusza ten typ identyfikacji będziemy nazywać **przeszukaniem bazy wzorców wieloujęciowych**. W pracy przedyskutowana została również możliwość budowy wzorców twarzy na podstawie obrazów 3D, jednak ze względu na mniejszą dostępność i wyższy koszt tego typu rozwiązań nie została ona rozwinięta w kompletny system. W proponowanych metodach skupiono się przede wszystkim na budowie i przeszukiwaniu baz wieloujęciowych, a celem badań jest optymalizacja rozwiązań pod kątem wykorzystania zbiorów danych złożonych ze zdjęć osoby, której twarz widoczna jest w różnych pozycjach, w tym również w pełnym profilu.

1.4 Ewaluacja systemów biometrycznych

We wstępie wprowadzone zostały pojęcia systemów opartych o identyfikację biometryczną oraz o weryfikację. Dla każdego z tych trybów przyjęto odpowiednie wskaźniki i scenariusze ewaluacji systemów. Standardy oraz definicje związane z ewaluacją pochodzą z międzynarodowych norm biometrycznych: w tym ze standardu ISO/IEC dotyczącego scenariuszy testowych dla systemów biometrycznych *„ISO/IEC 19795 Biometric performance testing and reporting"*[12] oraz z raportu NIST (amerykański instytut standaryzacyjny) opisującego testowanie algorytmów rozpoznawania twarzy *„FRVT - Face Recognition Vendor Test"* [13]. Sprawność systemów wyliczana jest na podstawie funkcji porównania dwóch próbek biometrycznych (lub dwóch wzorców reprezentujących te próbki), która zwraca w wyniku wskaźnik oceny s (*score*). Wskaźnik oceny może określać stopień podobieństwa (im większa wartość, tym próbki bar-

dziej do siebie podobne) lub też niepodobieństwa (odległość - im mniejsza, tym próbki bardziej do siebie podobne).

1.4.1 Ewaluacja systemów biometrycznych w trybie weryfikacji

Scenariusz ewaluacji w trybie weryfikacji opiera się na wynikach porównań jeden-do-jednego. Celem jest sprawdzenie, czy dwie zadane próbki należą do tej samej osoby. Bardzo często chodzi o odpowiedź na pytanie, czy tożsamość osoby reprezentowanej przez daną próbkę zgadza się z zadeklarowaną, a więc z zarejestrowaną wcześniej w systemie. W celu przetestowania systemu weryfikacji analizuje się dwa typy porównań wyliczonych na zbiorze testowym. Pierwszy typ to porównania pozytywne - wewnątrz-klasowe (*genuines*), a drugi to porównania negatywne - próbek należących do różnych klas (*impostors*). W teście badany jest rozkład wyników porównań z rozbiciem na oba podzbiory. W dobrze działającym systemie weryfikacji wskaźnik oceny podobieństwa s dla dwóch próbek należących do tej samej klasy powinien być wysoki, podczas gdy wskaźnik oceny dla próbek obcych powinien być niski. Można wtedy wyznaczyć próg dopasowania θ (*threshold*), który pozwoli wydać decyzję, czy próbki należą do tej samej osoby. Jeżeli wskaźnik oceny s będzie wyższy od $theta$ próbki zostaną uznane za zgodne. Z wyborem określonej wartości progu $theta$ będzie się wiązała pewna liczba błędnie wydanych decyzji, przy idealnej separacji porównań pozytywnych i negatywnych, liczba ta będzie równa 0.

Błąd fałszywego niedopasowania FNMR jest wskaźnikiem, który określa stosunek liczby niewłaściwych porównań par próbek należących do tej samej klasy, które w sposób niewłaściwy zostały uznane za niezgodne do liczby wszystkich porównań wewnątrzklasowych. Błąd fałszywego niedopasowania jest funkcją, która zależy od przyjętego progu dopasowania θ. **Błąd fałszywego dopasowania FMR** wynika ze stosunku liczby porównań par próbek należących do różnych klas, które błędnie zakwalifikowano jako zgodne w stosunku do liczby wszystkich porównań międzyklasowych. Błąd ten również zależy od przyjętego progu dopasowania θ. Wartości błędów FNMR i FMR są funkcjami zależnymi od progu θ. W przypadku, gdy pojawia się potrzeba skalarnej oceny skuteczności systemu, często używanym wskaźnikiem jest **błąd zrównoważony EER** (*Equal-Error-Rate*). Odpowiada on punktowi pracy systemu, w którym wartości błędów FMR oraz FNMR są sobie równe.

Kolejnym często stosowanym sposobem wizualizacji błędów wynikających z niepoprawnej weryfikacji w systemie jest **krzywa ROC** (*ang. Receiver Operator Characteristic*. Krzywa ROC przedstawia zależność wskaźnika poprawnych dopasowań TPR (True Positive Rate) od wartości wskaźnika fałszywych dopasowań FMR i jest wyliczana względem parametru θ. Podobny sposób wizualizacji oferuje metoda krzywej DET (*ang. Detection Error Tradeoff*), z tą różnicy, że na wykresie krzywej DET pokazana jest zależność wskaźnika fałszywych dopasowań FMR od wskaźnika fałszywych niedopasowań FNMR. Dodatkowo wartości pokazane są w zmienionej skali (najczęściej logarytmicznej).

1.4.2 Ewaluacja systemów biometrycznych w trybie identyfikacji

System działający w trybie identyfikacji opiera się na wykonywaniu porównań jeden-do-wielu. Celem jego działania jest przeszukanie pewnej listy referencyjnej wzorców i znalezienie jednego lub wielu, które wykazują największe podobieństwo do zadanego wzorca. Wymaga to porównania wzorca-zapytania ze wszystkimi wzorcami referencyjnymi. Typowo wynik przeszukania zwracany jest w postaci listy rankingowej posortowanej ze względu na wartości oceny podobieństwa s. Lista rankingowa może być ograniczona przez maksymalną długość listy lub przez minimalną wartość podobieństwa podpowiedzi znajdujących się na liście.

W przypadku systemów identyfikacji działających na **zbiorach zamkniętych** zagwarantowane jest, że dla każdej wyszukiwanej próbki-zapytania w bazie referencyjnej istnieje wpis z odpowiadającym jej identyfikatorem. Dla uproszczenia wiele testów porównujących wydajność algorytmów biometrycznych konstruowana jest na zasadzie zbioru zamkniętego. W scenariuszach rzeczywistych o wiele częściej konieczne jest rozwiązanie problemu na **zbiorze otwartym** - przy wyszukiwaniu mogą pojawić się osoby niezarejestrowane w bazie.

Wskaźnik Poprawnej Identyfikacji TPIR (*True Positive Identifaction Rate*) to wskaźnik oznaczający prawdopodobieństwo znalezienia właściwej osoby na liście podpowiedzi na pozycji r mniejszej lub równej K. Wskaźnik wyliczany jest jako stosunek liczby wszystkich zapytań, dla których odpowiadający mu wzorzec jest obecny z pozycją r mniejszą lub równą K oraz z wartością podobieństwa s większą lub równą θ przy założonej długości listy L. W tej pracy przy prezentacji wyników będzie stosowany uproszczony schemat wyliczania wskaźnika TPIR, w którym próg θ wynosi 0 - wyniki nie są filtrowane ze względu na wskaźnik oceny, a lista

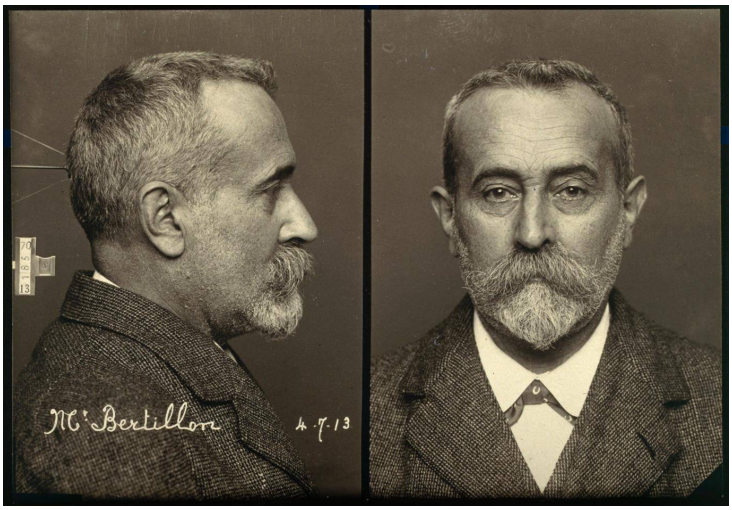
podpowiedzi ma zawsze stałą liczbę pozycji. Dodatkowo przyjęte jest założenie, że wskaźniki podpowiedzi mają bardzo małe prawdopodobieństwo powtórzenia wartości - wartości są w większości unikalne i wtedy pozycja podpowiedzi na liście jest tożsama z jej pozycją rankingową. Wartość TPIR jest wtedy funkcją zależną jedynie od pozycji rankingowej R. TPIR(1), czyli wskaźnik dokładności identyfikacji wyliczony na podstawie pierwszej podpowiedź nazywany jest również **Rank-1**.

Skumulowana krzywa dopasowań CMC (*Cumulative Match Characteristic* jest krzywą przedstawiająca zależność wartości wskaźnika TPIR od pozycji rankingowej R. Krzywa jest funkcją rosnącą, a przy założeniu identyfikacji na zbiorze zamkniętym wartość krzywej CMC dla ostatniej pozycji w rankingu R_N, gdzie N to długość listy do przeszukania wynosi 1. W przypadku biometrii twarzy może istnieć konieczność przeszukania ogromnych zbiorów danych ($N > 1000$), przy czym użytkownik systemu będzie w stanie percepcyjne skorzystać z listy podpowiedzi o długości do 100 elementów, dlatego też typowo krzywa CMC generowana jest tylko dla pierwszych pozycji rankingowych, np. $\{R_i : i \leq 12\}$

1.5 Eksperckie metody identyfikacji twarzy

Identyfikacja człowieka na podstawie twarzy wydaje się jedną z bardziej naturalnych metod sprawdzenia tożsamości, twarz nie wymaga skomplikowanej procedury ujawniania cech, tak jak ma to miejsce w przypadku odcisków palca. Co ciekawe, w zasadzie od samego początku prób stosowania wizerunku twarzy do identyfikacji były one powiązane ze zdjęciami wieloujęciowymi, a więc z wizerunkami, na których głwoa jest uchwycona w różnych pozach.

Jednym z prekursorów zastosowania zdjęć twarzy do weryfikacji osób, w tym przypadku przestępców, był paryski policjant Alphonse Bertillon. W 1881 zaproponował sposób usystematyzowania danych osób zatrzymanych przez policję na podstawie pomiarów cech fizycznych ciała. Metoda wykorzystania pomiarów ciała do identyfikacji została nazwana **antropometrią**. We Francji nie istniał wcześniej żaden inny sposób indeksowania kart przestępców. Alphonse Bertillon wymyślił system oparty na 11 pomiarach ciała. Wśród przyjętych pomiarów znalazły się: wzrost osoby, długość ramienia, długość torsu mierzona w pozycji siedzącej, długość głowy, szerokość głowy, wysokość prawego ucha, długość lewej stopy, długość lewego środko-



Rysunek 1.1: Alphonse Bertillon. Zdjęcie frontalne i z profilu.

wego palca, długość przedramienia, szerokość policzków oraz długość lewego małego palca. Bertillon zainspirowany współczesnymi mu pracami ze statystyki wyliczył, że przy 11 pomiarach prawdopodobieństwo powtórzenia zestawu cech wynosiło 268 435 456 do 1. Dodatkowo notował na kartach kolor włosów, oczu i skóry, a także tatuaże inne znaki szczególne. W celu ułatwienia przeszukiwania zbioru przez policjantów zaproponował sposób indeksowania kart. Na początku dzielił ludzi na 3 grupy względem długości głowy (mała, średnia i duża), potem względem szerokości głowy (również na podobne 3 zestawy), odległości między policzkami itd.

Ponadto Bertillon wykonywał zdjęcia osób podejrzanych w określonych pozach: z frontu i z profilu. Zdjęcia zbierał kamerą umocowaną na ziemi i o określonej pozycji względem krzesła, tak żeby zapewnić podobną pozycję i odległość od osoby fotografowanej przy każdym pomiarze. Można więc uznać, że to od niego pochodzi idea fotografii sygnalitycznej, o której mowa w rozdziale 1.6.2. W tamtych czasach zdjęcia nie były jednak używane do przeszukiwania danych. Karty indeksowane były tylko za pomocą pomiarów. Mimo że system spotkał się na początku z dużą rezerwą, szczególnie ze strony zwierzchników policjanta, to po kilku sukcesach wykazanych w akcji, został zaakceptowany. Do końca roku 1879, w którym został wprowadzony udało się zidentyfikować 49 osób, a 800 w pierwszych 3 latach. Alphonse Bertillon został odznaczony orderem Legii Honorowej po spektakularnym sukcesie z ujęciem Ravachola oskarżonego o serię zamachów bombowych w Paryżu.

Jak na swoje czasy system Bertillona był bardzo innowacyjny, niestety miał też swoje niedociągnięcia Była to przede wszystkim duża niedokładność pomiarów. Jednocześnie na popularności zyskiwała identyfikacja na podstawie linii papilarnych. Identyfikacja odcisków palców została wprowadzona przez Galtona i okazała się być w tym przypadku metodą bardziej dystynktywną, a także niezależną od osoby zbierającej dane. O ile sam Bertillon wykonywał pomiary antropometryczne w dosyć powtarzalny sposób (za pomocą skonstruowanych przez siebie urządzeń), o tyle na większą skalę niedokładności pomiarowe miały duży wpływ na wyniki rozpoznania. Niedokładności były jedną z przyczyn wyparcia antropometrii przez daktyloskopię, która do dziś dominuje w kryminalistyce. Niemniej jednak Alphonse Bertillon pozostawił po sobie wiele koncepcji wykorzystywanych do dziś, takich jak indeksowane karty policyjne, naukę nazywaną antropometrią, procedury badania miejsca zdarzenia i fotografię sygnalityczną.

Wykorzystanie fotografii sygnalitycznej do indeksowania kart policyjnych nie było możliwe w tamtych czasach, ale dzięki rozwojowi algorytmów wizji komputerowej tego typu dane mogą być obecnie używane w systemach identyfikacji.

1.6 Współczesne metody obrazowania twarzy i źródła danych

Automatyczne metody rozpoznawania twarzy należą do szerokiego wachlarza metod przetwarzania obrazów, nazywanych też metodami wizji komputerowej. Mimo że biometria twarzy kojarzy się przede wszystkim z rozpoznawaniem wizerunku na zdjęciach, istnieje wiele różnych technologii uwieczniania twarzy. Metody obrazowania różnią się sposobem rejestracji danych i, co jest z tym związane, urządzeniami rejestrującymi, jak również technologią zakodowania danych. W literaturze dotyczącej biometrii twarzy można odnaleźć algorytmy wykorzystujące następujące rodzaje danych:

1. Obrazy 2D, wśród obrazów 2D możemy wyróżnić bardziej specyficzne typy obrazów, które będą analizowane w pracy:
    - fotografia sygnalityczna - rodzaj zdjęć zbieranych przez służby policyjne, każdej osobie wykonuje się serię fotografii w ściśle określonych warunkach i pod zadanym zestawem kątów.
    - zdjęcia bez restrykcji - odpowiednik angielskiego określenie *„in the wild”*, obrazy

rejestrowane bez kooperacji z użytkownikiem charakteryzujące się wysoką zmiennością czynników takich jak poza, rozdzielczość i jakość obrazu

2. Obrazy 3D pochodzące ze skanerów 3D lub kamer głębi
3. Alternatywne metody obrazowania twarzy, w tym:
    - obrazy zbierane przez urządzenia do termowizji
    - obrazy w bliskiej podczerwieni

Nie wszystkie metody obrazowania zostaną omówione w poniższej pracy, warto jednak zwrócić uwagę, że można je ze sobą łączyć. Projektowanie systemu rozpoznawania w oparciu o dane multimodalne może mieć przewagę nad systemem opartym tylko na jednym rodzaju obrazów. Zapisywanie większej liczby cech twarzy utrudnia fałszerstwo, ponieważ cechy takie jak temperatura, czy struktura 3D twarzy są trudniejsze do skopiowania i wykorzystania w czasie identyfikacji w formie tzw. ataku prezentacji (ang. *presentation attack*).

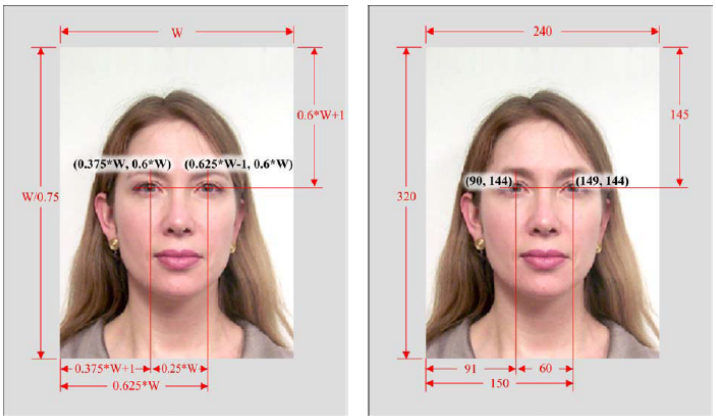
1.6.1 Obrazowanie 2D

Obrazowanie 2D jest jednym z najszerzej stosowanych sposobów rejestracji twarzy. Metody zaproponowane w tej pracy również opierają się na przetwarzaniu obrazów tego typu. Zdjęcia cyfrowe twarzy występują w dokumentach tożsamości, identyfikatorach służbowych, na dyplomach i legitymacjach i jako takie stanowią jedne z największych dostępnych zbiorów danych służące do identyfikacji lub weryfikacji osób. Wykonanie fotografii twarzy, która jest odpowiednia do automatycznej identyfikacji, jest stosunkowo łatwe i obarczone niskim kosztem. Zdjęcia można wykonywać zarówno aparatem cyfrowym, jak i smartfonem. Stąd kolejnym obszernym źródłem danych może być internet. Obrazy, które wchodząc w skład największych zbiorów do treningu sieci głębokich składają się najczęściej ze zdjęć zebranych w wyniku przeszukiwania portali społecznościowych oraz ze zdjęć osób publicznych, które zostały umieszczone w sieci.

Obraz twarzy 2D może być traktowany jako macierz, która zawiera w swoich komórkach informacje o barwie. Format zapisu informacji o barwie różni się w zależności od przyjętej przestrzeni barwnej. Najczęściej stosowanym modelem przestrzeni barwnej, przyjętym również w tej pracy, jest model RGB, który przechowuje informację o kolorach w 3 kanałach, które kodują odpowiednio natężenie barwy: czerwonej, zielonej i niebieskiej.

Obraz 2D jest wynikiem projekcji trójwymiarowego obiektu na płaszczyznę. Wobec tego własności odwzorowanej twarzy, takie jak długość i szerokość zależą od odległości osoby fotografowanej od aparatu, kąta i ogniskowej urządzenia. W celu ujednolicenia sposobu odwzorowania i ulepszenia wyników porównania do obrazów twarzy 2D stosowana jest procedura **normalizacji twarzy** w obrazie (ang. *face alignment*). Operacje, które mogą być wykonywane w ramach normalizacji to:

- skalowanie - zmiana wielkości obrazu do wybranego rozmiaru, np. w taki sposób, żeby odcinek między oczami był określonej długości
- normalizacja pozycji 2D głowy - przesunięcie i obrót obrazu, w taki sposób, żeby określone punkty, np. środki oczu znalazły się w określonych przez współrzędne położeniach
- wyrównanie poziomu jasności obrazu
- normalizacja pozy 3D głowy - korekcja kąta głowy względem jednej, dwóch lub trzech osi obrotu. Operacja może wymagać nieliniowych transformacji obrazu oraz uzupełnienia punktów niewidocznych na obrazie źródłowym



Rysunek 1.2: Ilustracja wytycznych dotyczących pozycjonowania twarzy do dokumentów zdefiniowanych w standardzie ISO/IEC 19794-5:2005(E) [14]. Minimalna szerokość tego typu zdjęcia powinna wynosić 240 pikseli, a pozycje pozostałych elementów w obrazie, takich jak środki oczu, wyznaczane są względem tej szerokości. Po prawej stronie pokazane są przykładowe wymiary dla obrazu o szerokości 240 pikseli.

Ze względu na to, że niektóre z automatycznych systemów wymagają normalizacji obrazu twarzy do poprawnego działania, powstały międzynarodowe normy, które określają wymagania dotyczące fotografii twarzy, która może zostać umieszczona w cyfrowych dokumentach toż-

samości, takich jak paszporty i dowody osobiste. Format zdjęcia w paszportach europejskich powinien być zgodny ze standardami zdefiniowanymi w dokumencie Organizacji Międzynarodowego Lotnictwa Cywilnego ICAO „Biometrics Deployment of Machine Readable Travel Documents, Technical Report" [15] oraz normach ISO/IEC 19794-5:2005 „Biometric Data Interchange Formats – Part 5: Face Image Data"[14]. W Polsce wymogi dotyczące zdjęć do dokumentów opisane są w rozporządzeniu Ministerstwa Spraw Wewnętrznych i Administracji oraz w „Instrukcji wykonywania zdjęć do dokumentów paszportowych oraz dowodów osobistych".

Wykonanie zdjęć twarzy zgodnych z normami wymaga nadzorowania procesu akwizycji i ewentualnej obróbki fotografii przez eksperta. W biometrii twarzy nie zawsze jednak mamy do czynienia z obrazami wykonanymi z intencją wykorzystania ich w systemie identyfikacji, dlatego wyróżniamy dwa podejścia do rejestracji:

1. **akwizycja kontrolowana** - obrazy rejestrowane w ściśle określony sposób, użytkownik współpracuje i otrzymuje instrukcję od fotografa. Przykładem takich zdjęć jest fotografia sygnalityczna wykonywana przez policję, jak również różnego rodzaju zdjęcia związane z dokumentami tożsamości.
2. **akwizycja bez restrykcji** - obrazy zbierane bez narzuconego scenariusza. Na jednym obrazie może pojawić się więcej osób, twarze nie muszą być zwrócone w stronę aparatu. Obecne są zakłócenia, takie jak: częściowe przesłonięcie, czy brak równomiernego oświetlenia.

1.6.2 Zdjęcia sygnalityczne

Zdjęcia sygnalityczne to rodzaj zdjęć wieloujęciowych. Są wykonywane według ściśle określonej procedury przez przez odpowiednie komórki techników kryminalistycznych. Dla każdej z rejestrowanych osób wykonywana jest seria zdjęć, pokazująca głowę fotografowanej osoby pod wybranymi kątami. W Polsce pobiera się fotografie w czterech pozach: jedną frontalną, dwie z półprofilu oraz jedno zdjęcie prawego profilu. Dodatkowo można zrobić zdjęcie pełnej sylwetki oraz zdjęcie pokazujące cechy szczególne.

Obrazy sygnalityczne są przede wszystkim związane z policyjnymi bazami danych. Gotowe zdjęcia umieszcza się na specjalnej karcie albumowej oraz rejestruje w bazie KSIP (Krajowy

Rysunek 1.3: Zestaw zdjęć sygnalitycznych pochodzących ze strony http://poszukiwani.policja.pl/. Na zestaw składa się zdjęcie lewego półprofilu, portret en-face, prawy półprofil i prawy pełny profil. Na dwóch ostatnich obrazach widać tabliczkę informacyjną.

System Informacji Policyjnej) poprzez zeskanowanie zdjęcia. KSIP to centralny zbiór, w którym odnotowywane są informacje o osobach podejrzanych o popełnienie przestępstw ściganych z oskarżenia publicznego, osobach poszukiwanych, bądź usiłujących ukryć swoją tożsamość, a także o zgubionych lub skradzionych przedmiotach.

Album ze zdjęciami sygnalitycznymi tworzony jest w określonych sytuacjach. Według Dziennika Urzędowego Komendy Głównej Policji [16] zdjęcia sygnalityczne to zdjęcia wykonywane:

1. osobie podejrzanej o popełnienie przestępstwa ściganego z oskarżenia publicznego
2. nieletniemu, który dopuścił się czynu zabronionego przez ustawę jako przestępstwo ścigane z oskarżenia publicznego
3. osobie, której tożsamości nie można ustalić lub usiłującej ukryć swoją tożsamość

1.6.3 Dane bez restrykcji (ang. *in the wild*)

Dane bez restrykcji określane są często po angielsku jako „*in the wild*", w poniższej pracy mogą być również określane jako dane niekontrolowane. Można je zdefiniować jako obrazy twarzy zbierane bez kooperacji z użytkownikiem, co do których nie ma żadnych z góry narzuconych ograniczeń, jeśli chodzi o metodę akwizycji. Na zdjęciach bez restrykcji mogą być widoczne twarze pod różnym kątem, na jednym obrazie może być obecnych kilka osób. Nie mają określonych warunków oświetleniowych (zdjęcia wewnątrz i na zewnątrz pomieszczeń), osoby mogą być w ruchu oraz częściowo osłonięte.

Rysunek 1.4: Przykładowe zdjęcia bez restrykcji. Klatki z nagrania pochodzące z monitoringu zewnętrznego. Obrazy należą do zbioru Quiscampi [17].

Podsumowując istnieje długa lista czynników zmienności, które mogą pojawić się na zdjęciach twarzy bez restrykcji:

- zróżnicowane źródła danych (różnice w parametrach urządzeń rejestrujących zdjęcia, takie jak rozdzielczość, ogniskowa i dystorsje)
- mały rozmiar twarzy na zdjęciu (fotografia z daleka, niska rozdzielczość)
- niejednolite tło
- niejednolite oświetlenie
- pełen zakres dopuszczalnych pozycji głowy względem kamery (twarz widoczna z boku lub od góry)
- brak ostrości, zdjęcia poruszone
- twarz częściowo zasłonięta, obecność okularów, elementów ubioru, które przesłaniają obszar twarzy
- wpływ mimiki osoby na zdjęciu

Sprawdzenie wydajności algorytmów rozpoznawania twarzy na danych bez restrykcji stało się w pewnym momencie kluczowe dla rozwoju systemów biometrycznych, ponieważ tylko takie metody mają możliwość działać autonomicznie, bez konieczności nadzoru przez wytreno-

wanych pracowników, np. w systemach monitoringu wizyjnego. Algorytmy wytrenowane na danych zbieranych w określony sposób w laboratorium znacząco odbiegały dokładnością po wdrożeniu w warunkach rzeczywistych. W dziedzinie rozpoznawania twarzy bardzo szybko dostrzeżono potrzebę zbierania tego typu danych, dlatego istnieje wiele zbiorów zawierających dane „*in the wild*".

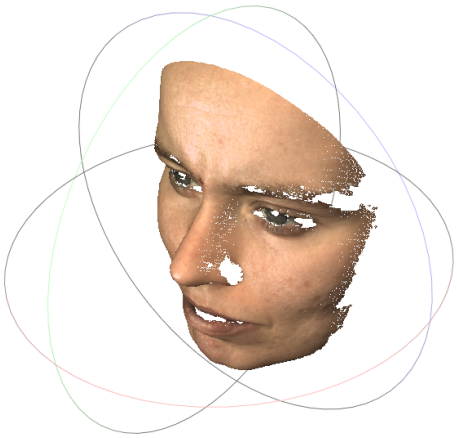
1.6.4 Obrazowanie 3D

Zastosowanie obrazów 3D w biometrii twarzy może być odpowiedzią na wrażliwość algorytmów 2D na pozycję fotografowanego obiektu względem kamery. W obrazach 3D każdy punkt, nazywany w tym przypadku **wokselem**, oprócz standardowych współrzędnych położenia x i y oraz informacji o barwie, charakteryzuje się także dodatkową informacją o głębi, czyli o odległości obiektu od kamery. W przeciwieństwie do obrazowania 2D, gdzie współrzędne punktów podane są w układzie współrzędnych obrazu, dla danych 3D w sposób pośredni lub bezpośredni możliwe do uzyskania są współrzędne rzeczywiste. Skan głowy 3D opisany współrzędnymi rzeczywistymi jest zawsze naturalnej wielkości niezależnie od kąta wykonywania fotografii i odległości od obiektu. Nie wymaga więc skalowania na etapie normalizacji obrazu. Skany 3D zachowują również informacje o strukturze powierzchni twarzy, które dla obrazów 2D są możliwe są do uzyskania jedynie w przybliżonej formie. Obrazy 3D można przechowywać za pomocą rozmaitych struktur danych. Wybór jest zależny od urządzenia obrazującego oraz od metod przetwarzania.

1. **Chmura punktów** (*ang. point cloud*) jest najprostszą strukturą do przechowywania danych 3D. Jest to lista wokseli wraz z przypisanymi do nich wartościami określającymi współrzędne położenia punktu oraz jego barwę (np. *X,Y,Z,R,G,B*). W wersji podstawowej punkty nie są uporządkowane. Większość operacji niezbędnych do wyliczenia cech powierzchni obiektu wymaga znalezienia najbliższych sąsiadów, a przeszukiwanie dużego zbioru punktów w przypadku skanów o wysokiej rozdzielczości jest czasochłonne.
2. **Obraz RGBD** - obrazy te składają się z pary obrazów: klasycznego barwnego obrazu 2D (RGB) oraz odpowiadającej mu **mapy głębi** (*ang. depth map*). Struktura ta wynika z budowy urządzeń do rejestracji danych, takich jak Kinect (skanery światła strukturalnego), które składają się z 2 typów detektorów: pierwszy to kamera RGB, a drugi to detektor

światła podczerwonego emitowanego przez urządzenie. Są to dwa oddzielne źródła danych, które wymagają wzajemnej kalibracji.

3. **Siatka wielokątów** (*ang. polygon mesh*) - jest to format stworzony przede wszystkim na potrzeby grafiki komputerowej. Ma na celu w wygodny sposób opisać własności powierzchni leżących w przestrzeni 3D. Powierzchnie opisywane są przez zestaw wierzchołków, krawędzi i pól, z którymi powiązana jest informacja o barwie. Najprostsza i najczęściej stosowana jest siatka złożona z trójkątów. Siatka jest uporządkowana, a więc wyliczenia własności powierzchni jest znacznie wydajniejsze niż w przypadku chmur punktów.



Rysunek 1.5: Obraz 3D twarzy zarejestrowany skanerem ze światłem strukturalnym Smarttech FaceOne - jest to urządzenie stacjonarne, które w czasie akwizycji rzutuje na obiekt serię pasków w świetle widzialny. Zarejestrowany obraz jest kodowany za pomocą chmury punktów.

Przetwarzanie danych 3D jest bardziej wymagające obliczeniowo niż przetwarzanie danych 2D. Same struktury danych też zajmują więc miejsca na dysku niż obrazy 2D. Istotną częścią przetwarzania wstępnego jest więc zmniejszenie objętości samej struktury. Akwizycja obrazów 3D wymaga dedykowanych urządzeń lub zastosowania metod składania obrazów 2D, takich jak stereowizja lub metody fotogrametryczne.

Wśród urządzeń dedykowanych do zbierania informacji 3D można wymienić skanery ze światłem strukturalnym (może to być światło widzialne lub podczerwone) [18], skanery laserowe oraz kamery ToF (ang. *time-of-flight*). Skanery ze światłem strukturalnym opierają swoją zasadę działania na rzutowaniu na obiekt pewnego wzorca świetlnego. Na podstawie odkształceń

tego wzorca spowodowanych obecnością obiektu w polu widzenia kamery można oszacować odległość obiektu od detektora.

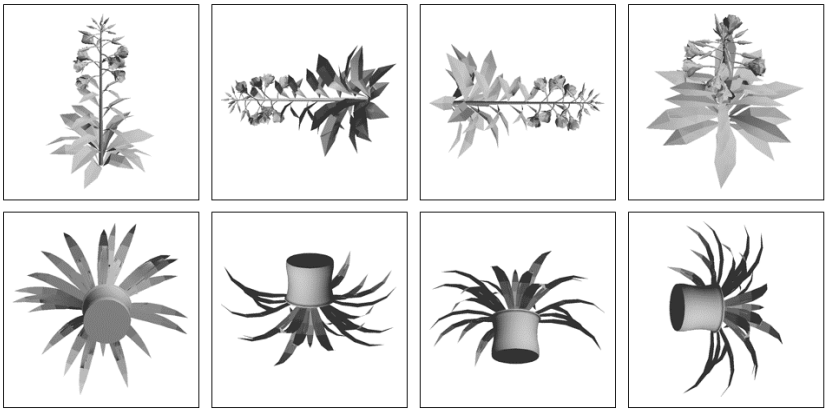
Na etapie badań wstępnych dotyczących rozpoznawania twarzy autorka pracy zaproponowała swój system biometrii twarzy 3D [6]. System był zbudowany na bazie mobilnego zestawu do akwizycji (rys.1.6), który składał się z tabletu PC oraz kamery głębi Structure [19]. Kamera głębi Structure jest rodzajem skanera ze światłem podczerwonym i ma parametry identyczne z urządzeniem do kontroli gier Microsft Kinect [18]. W odróżnieniu od tego drugiego może być stosowana razem z urządzeniem mobilnym, takim jak tablet i smartfon. W eksperymentach z użyciem zaproponowanego zestawu do akwizycji zanalizowano algorytm porównania cech twarzy, które można uzyskać z obrazu 3D, w tym na podstawie krzywizny powierzchni i wartości głębi. Mimo poszerzonego zestawu cech okazało się, że informacja o barwie, która jest dostępna w obrazach 2D, w dalszym ciągu była kluczowa w procesie rozpoznawania.

Rysunek 1.6: Mobilny zestaw do obrazowania 3D twarzy oparty na kamerze głębi Structure. Urządzenie wraz z oprogramowaniem zostało zbudowane przez autorkę i opisane w artykule [6]. Kamera Structure ma parametry podobne do kamery Microsoft Kinect przy mniejszych rozmiarach i możliwości zasilania z urządzenia mobilnego.

Obrazowanie 3D w biometrii twarzy cieszy się zainteresowaniem również ze względu na bezpieczeństwo systemów biometrycznych. Systemy oparte o obrazowanie 2D i nie zawierające dodatkowych zabezpieczeń mogą zostać oszukane przez ataki w formie prezentacji kopii wizerunków przed urządzeniem rejestrującym. Wykonanie falsyfikatu, który zachowywałby trójwymiarowe cechy wyróżniające twarzy, np. w formie maski, jest o wiele bardziej pracochłonne niż w przypadku falsyfikatów pod systemy działające na zdjęciach 2D.

1.6.5 Obrazy wieloujęciowe

Obrazy wieloujęciowe to zestawy obrazów pokazujących ten sam obiekt lub osobę z różnych, ale określonych kątów. Obrazy wieloujęciowe są więc sposobem reprezentacji obiektów 3D poprzez obrazy 2D. Klasycznym zbiorem obrazów wieloujęciowych na których trenowane są niektóre z modeli sieci agregacyjnych omawianych w pracy jest wieloujęciowa wersja bazy ModelNet-40 [20]. Oryginalny zbiór składa się z 40 klas obiektów 3D reprezentowanych za pomocą modeli typu CAD. Przykładowe kategorie obiektów to gitara, kubek, samolot, krzesło. W artykule [1] zaproponowano rozwiązanie problemu klasyfikacji obiektów z zastosowaniem sieci głębokich, które otrzymują na wejściu zestawy projekcji 2D obiektów ze zbioru ModelNet-40. Zestawy widoków są renderowane dla 80 wirtualnych ustawień kamery dla każdego z obiektów. Przykładowe projekcje obiektów przestawione są na rysunku 1.7.



Rysunek 1.7: Przykładowe projekcje obiektów 3D pochodzących z bazy ModelNet-40. W pierwszym i drugim rzędzie widać dwa różne obiekty należące do klasy „roślina". W przypadku większości obiektów należących do zbioru obiekty nie mają zdefiniowanego ustawienia początkowego, tzn. pozycji w której można uznać, że są ustawione frontalnie .

Struktura ModelNet-40 jest istotna, ponieważ część omawianych w pracy metod była zoptymalizowana pod kątem działania na tym właśnie zbiorze danych i niektóre decyzje projektowe twórców wynikały z jego budowy. Ponadto wielkość i właściwości tej bazy mogły stanowić wskazówkę, jakie wymagania powinna spełniać baza wieloujęciowa twarzy. W analogiczny sposób można generować obrazy wieloujęciowe na podstawie bazy twarzy 3D, jednak ze względu na możliwość zastosowania w rzeczywistych systemach identyfikacji o wiele cie-

kawszym rodzajem danych, które również są zgodne z definicją zdjęć wieloujęciowych, były dla autorki zdjęcia sygnalityczne twarzy. Jak zostało już wspomniane we wcześniejszych rozdziałach zdjęcia sygnalityczne mają określony sposób akwizycji i przestawiają głowę fotografowanej osoby w przyjętych odgórnie ujęciach. Temat przetwarzania obrazów 3D twarzy warto mimo to wymienić jako potencjalny kierunek rozwoju opisanych w pracy metod.

1.7 Bazy danych i scenariusze testowe

Podstawowym celem poniższej pracy jest analiza działania metod biometrii twarzy w scenariuszu identyfikacji. Wybór danych oraz sposobu testowania jest podyktowany wymaganiami przyjętego scenariusza. W scenariuszu identyfikacji dane testowe są zazwyczaj dzielone na dwa podzbiory. Pierwszy podzbiór to dane które służą do zbudowania bazy wzorców referencyjnych wśród których będziemy wyszukiwać. Dane należące do tego podzbioru będą w dalszej części nazywane **próbkami galeryjnymi** od angielskiej nazwy bazy referencyjnej, tj. *gallery*. Drugi podzbiór to próbki na podstawie których chcemy dokonać identyfikacji i które będą nazywane **próbkami-zapytaniami** (ang. *query*).

Szczególne wymagania wiążą się z obecnością w bazie zdjęć wieloujęciowych. Ze względu na założenia ich obecność jest wymagana wśród próbek galeryjnych. Próbki galeryjne powinny być poetykietowane z uwzględnieniem kąta pod którym widoczna jest twarz na zdjęciu. Etykiety ujęć pozwalają na podział zbioru na zestawy wieloujęciowe. Liczność takich zestawów pozostaje kwestią badawczą, dlatego dobrze, żeby zbiór danych oferował różne konfiguracje zestawów ujęć. Badając wpływ pozy na rozpoznawanie twarzy potrzebny jest zbiór zdjęć możliwie kontrolowany i pozbawiony innych dystraktorów, które mogłyby wpłynąć na wyniki analizy. Zbiór próbek galeryjnych powinien zawierać dużą liczbę reprezentacji dla każdej z osób, co więcej liczba reprezentacji w konkretnej pozie również powinna być większa niż 2-3. Na próbki-zapytania nie ma narzuconych tak wielu wymagań, ale chcąc zasymulować identyfikację jak w rzeczywistych systemach, potrzebne są próbki bez restrykcji, np. uzyskane z materiałów z monitoringu wizyjnego. Co najważniejsze w każdej z dwóch opisanych powyżej części, a więc galerii i zbiorze próbek-zapytań, powinien znajdować się ten sam zestaw osób, dzięki czemu możemy oba podzbiory ze sobą powiązać. Nieco inne wymagania są związane z treningiem sieci neuronowych, tam ważna jest przede wszystkim liczność i zróżnicowanie

zbioru. Zdjęcia powinny reprezentować osoby w możliwie szerokim zakresie zmienności. Niestety zdobycie bazy danych, która spełniałaby wszystkie powyższe wymagania jest niemożliwa i zbyt czasochłonna.

W eksperymentach omówionych w dalszej części pracy mogą być wykorzystane różne bazy danych w zależności od potrzeb badania. Wybór może się różnić w zależności od tego, czy chcemy testować rozwiązania, wytrenować własny model od zera czy modyfikować sieć, której wagi można zainicjalizować z użyciem mechanizmu transferu wag [21]. Trudno wskazać jeden zbiór danych, który byłby na tyle uniwersalny, żeby móc go zastosować w każdym przypadku. W części zastosowań użyto gotowych powszechnie dostępnych zbiorów danych, których zastosowanie jest istotne ze względu na możliwość porównania z innymi rozwiązaniami opisanymi w literaturze. Z tego powodu w pracy wykorzystano dane i protokół testowy należące do benchmarku IJB-C [5]. Poza tym zmodyfikowano istniejące bazy danych, tak żeby lepiej spełniały wytyczne potrzebne w przeprowadzanych eksperymentach. Zbiory własne mają w nazwie przedrostek „MV", np. MV-MultiPIE oraz MV-Quiscampi, żeby odróżnić jest od niezmodyfikowanych zestawów danych. Zbiory MV-MultiPIE i MV-Quiscampi posiadają podział na zestawy wieloujęciowe, przy czym ze względu na liczbę zestawów tylko ten pierwszy może być zastosowany w treningu modeli. Do trenowania modeli twarzy od zera istotna jest duża liczba zróżnicowanych próbek, tak jak ma to miejsce w zbiorach VGGFace [22] i VGGFace2 [4].

1.7.1 MV-Quiscampi

Pierwszy zbiór testowy został oparty na bazie zdjęć Quiscampi [17], która została zebrana na portugalskim uniwersytecie Beira Interior. Quiscampi jest to zbiór danych dedykowany do identyfikacji twarzy w systemach monitoringu zewnętrznego. Fragment zbioru został po raz pierwszy udostępniony na potrzeby konkursu *International Challenge on Biometric Recognition in the Wild 2016*, którego celem było wyszukanie osób na podstawie zapytań złożonych z obrazów z monitoringu o niskiej jakości. Przy czym baza referencyjna była budowana ze zdjęć zebranych w kontrolowanych warunkach, a każda z osób była reprezentowana przez trzy różne ujęcia. Baza źródłowa składa się głównie z wizerunków pracowników oraz studentów uniwersytetu. W Quiscampi zebrano dane dla 280 osób, przy czym są to dane o zróżnicowanym charakterze.

Wśród materiałów pokazujących daną osobę można wyróżnić następujące rodzaje danych:

1. obrazy całej sylwetki w wysokiej rozdzielczości (2912×4368) wykonane wewnątrz pomieszczenia w trzech pozach względem kamery: frontalnej, z lewego profilu i z prawego profilu. Po dokonaniu detekcji z każdego z obrazów można uzyskać zdjęcie twarzy.
2. nagrania wideo dobrej jakości pokazujące obracającą się głowę osoby na jednolitym zielonym tle wykonane wewnątrz pomieszczenia
3. klatki wycięte z monitoringu wizyjnego, na której jest widoczna dana osoba. Monitoring wizyjny umieszczony jest na zewnątrz budynku, obrazy zbierane były automatycznie, widoczne na obrazach osoby nie mają twarzy skierowanej w stronę kamery, głowa może być częściowo osłonięta lub rozmazana, obserwowane osoby mogą nosić okulary

MV-Quiscampi jest wynikiem reorganizacji danych źródłowych oraz ich dostosowania do założeń przyjętych w prowadzonych badaniach. W badaniach założyliśmy jasny podział na obrazy, z których może zostać utworzona galeria i które pokazują głowę osoby w wielu ujęciach oraz na próbki-zapytania, które nie muszą mieć założonej struktury. Wszystkie twarze widoczne na obrazach dostępnych w zbiorze źródłowym zostały przycięte z wykorzystaniem detektora twarzy wchodzącego w skład pakietu *dlib* 19.17 [23]. Wyniki operacji zostały przejrzane pod względem brakujących lub niedokładnych detekcji - miało to miejsce przede wszystkim w przypadku zdjęć profilowych. Nieprawidłowe detekcje zostały skorygowane lub wykonane ręcznie. W przypadku materiałów z monitoringu, w których na pojedynczych klatkach występowała więcej niż jedna osoba, zostały poprawione identyfikatory osób. Klatki w których nie można było dokonać detekcji, np. dana osoba była widoczna od tyłu zostały zignorowane. Jako że zestaw ujęć w części galeryjnej zbiory był dosyć skromny, tzn. składał się tylko z trzech ujęć, w zbiorze MV-Quiscampi został poszerzony przez wyodrębnienie dodatkowych ujęć z nagrań wideo pokazujących obracającą się głowę. Nagrania wideo zostały pocięte na klatki ręcznie tak żeby otrzymać w wyniku siedem klatek wideo, na których jest widoczna głowa danej osoby:

- w pozie frontalnej
- pod kątem 15° z lewej i prawej strony
- pod kątem 45° z lewej i prawej strony
- pod kątem 90° z lewej i prawej strony

Rysunek 1.8: Przykładowe obrazy pochodzące z bazy Quiscampi (przed detekcją twarzy). W pierwszym rzędzie widać zdjęcia pełnej sylwetki (1), w drugim pięć klatek wyodrębnionych z nagrania wideo (2), w ostatnich rzędach pokazane są wycinki z monitoringu wizyjnego, na których jest widoczna ta sama osoba.

Nie wszystkie osoby w zbiorze posiadały pełen zakres typu danych, tzn. wideo, zdjęcia kontrolowane i zdjęcia z monitoringu. Podsumowując, po wykonaniu wszystkich operacji przetwarzania wstępnego, powstał zbiór MV-Quiscampi, który składa się z trzech podzbiorów:

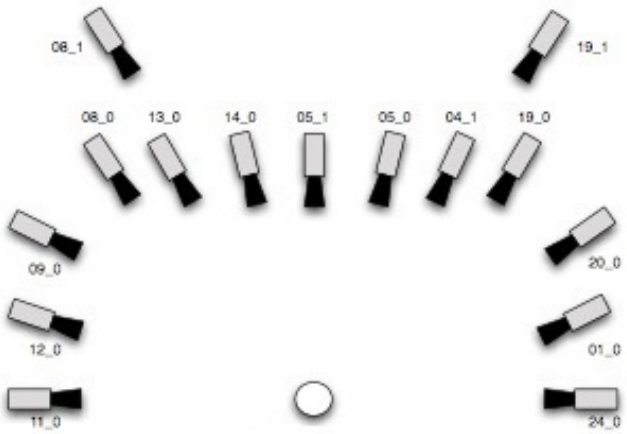
1. Galeria - 807 obrazów twarzy dla 269 osób, każda osoba jest reprezentowana przez 3 możliwe ujęcia głowy: frontalne, lewy profil i prawy profil
2. Galeria-z-Wideo - 1911 obrazów dla 273 osób z czego każda osoba jest reprezentowana przez 7 ujęć: frontalne, $\pm 15^\circ$, $\pm 45^\circ$, $\pm 90^\circ$.
3. Monitoring - 2907 obrazów dla 190 osób, w każdej klasie znajduje się średnio 15,3 zdjęcia, z czego minimalnie może być jedno, a maksymalnie 73. 91% osób ma więcej niż 3 zdjęcia w zbiorze

Zbiór MV-Qusicampi spełnia jedno z wymagań stawianych zbiorom danych wykorzystywanym w pracy - zawiera zdjęcia dwóch rodzajów: wieloujęciowe obrazy, które mogą posłużyć do utworzenia galerii oraz zdjęcia bez restrykcji, które pokazują te same osoby i mogą być wykorzystane jako zapytania do bazy. Po dodaniu wycinków liczba dostępnych ujęć w zestawie obrazów wieloujęciowych wynosi 7. Zbiór galeryjny i zbiór próbek z monitoringu ma dużą część wspólną jeśli chodzi o liczbę osób - 189 osób znajduje się w obydwu podzbiorach. Zestaw danych jest średniej wielkości - ma ponad 100 klas, więc jest dosyć reprezentatywny w kontekście wyników ewaluacji algorytmów identyfikacji. Nie jest jednak wystarczający do treningu sieci głębokich - w zbiorze galeryjnym jest tylko jedna sesja lub dwie jeśli doliczyć wycinki wideo, a to za mało, żeby wyuczyć algorytm różnych konfiguracji ujęć. Ponadto całość zbioru to tylko około 5 tysięcy obrazów.

1.7.2 MV-Multi-PIE

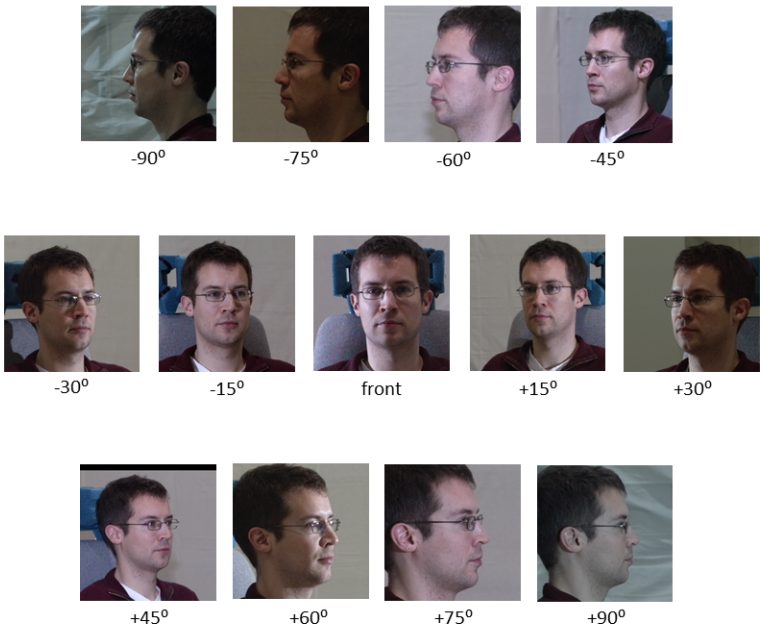
Do stworzenia scenariusza testowego MV-Multi-PIE posłużyła baza obrazów twarzy CMU Multi-PIE zebrana na Uniwersytecie Carnegie Mellon w Pittsburghu, w Stanach Zjednoczonych. Zbiór danych jest udostępniany na zasadach komercyjnych. Baza CMU Multi-PIE to zestaw kontrolowanych zdjęć twarzy, który zawiera ponad 750 tysięcy próbek należących do 337 osób zebranych w czterech sesjach na przestrzeni pięciu miesięcy. Kontrolowanymi czynnikami zmienności były poza głowy, kąt oświetlenia i mimika twarzy, podczas gdy pozostałe warunki pozostawały w miarę możliwości niezmienne, np. jednolite tło wewnątrz pomieszcze-

nia. Każda z fotografowanych osób siedziała na specjalnym stanowisku wokół którego ustawione było 15 kamer. Ustawienia kamery względem stanowiska pokazane są na obrazie 1.9. Obrazy pobierane były w formie sekwencji, w czasie której zmieniał się kąt padania światła na twarze. W zbiorze można wyróżnić 19 różnych ustawień światła. Osoby fotografowane były w czterech sesjach, w ramach każdej sesji zostały utworzone sekwencje ze zmienną mimiką, w tym pokazujące twarz neutralną, smutną, uśmiechniętą, złą, zdziwioną, zniesmaczoną i krzyczącą. Odpowiednio pierwsza sesja zawiera 2 zestawy z różną mimiką, druga sesja zawiera 3 zestawy, trzecia sesja zawiera 3 zestawy i czwarta sesja zawiera 3 zestawy. Baza zajmuje na dysku ponad 305 GB danych. W bazie CMU Multi-PIE źródłowo jest 15 ustawień kamery, z czego 13 to ujęcia na tym samym poziomie co głowa, a 2 to zdjęcia robione z góry. Do utworzenia zbioru MV-Multi-PIE wybrano 13 kątów, w tym samym poziomie, co głowa (zakres wybranych ujęć został przedstawiony na grafice 1.10). Kąty te leżą w zakresie od pełnego lewego profilu do pełnego prawego profilu z odstępem, co $15°$, a więc są to odpowiednio: $-90°$ (lewy pełny profil), $-75°$, $-60°$, $-45°$ (lewy półprofil), $-30°$, $-15°$, $0°$ (portret frontalny), $15°$, $30°$, $45°$ (prawy półprofil), $60°$, $75°$, $90°$ (prawy pełny profil).



Rysunek 1.9: Rozmieszczenie kamer wokół stanowiska do rejestracji twarzy zastosowane przy zbieraniu zdjęć do bazy CMU Multi-PIE na podstawie dokumentacji [24]

Potrzeba organizacji własnego zestawu danych wynikała przede wszystkim z braku zbiorów, w którym istniałby podział na zestawy wieloujęciowe, a dodatkowo można by wybrać wiele tego typu zestawów dla jednej osoby. Zdjęcia pochodzące ze zbioru CMU Multi-PIE idealnie pasowały do tego zastosowania, jako że posiadały jasno określone etykiety klas i zostały pobrane w kontrolowanych warunkach. Baza MV-Multi-PIE została podzielona na zbiory trenu-



Rysunek 1.10: Przykładowe obrazy z bazy MV-Multi-PIE. Twarz na zdjęciach jest widoczna w różnych ujęciach. Jest to 13 kątów w zakresie od lewego pełnego profilu $-90°$ do prawego pełnego profilu $90°$ z odstępem co $15°$

jący, walidacyjny i testowy. Wybór klas do zbiorów został podyktowany kryterium obecności zestawu we wszystkich czterech sesjach, co nie ma miejsca dla wszystkich klas źródłowych. Zróżnicowanie warunków oświetleniowych nie było tematem badań, dlatego wybór ustawień światła był losowy dla każdego zestawu ze zbioru źródłowego. Obrazy w zbiorze MV-Multi-PIE zostały przycięte z zastosowaniem detektora twarzy dlib [23].

1. Część **trenująca** zawiera 120 klas reprezentowanych przez 12 480 plików obrazów. Pliki pochodzą z sesji pierwszej, drugiej i trzeciej, a więc każda osoba ma przypisane 8 sekwencji po 13 ujęć.
2. Część **walidacyjna** zawiera te same 120 klas, co część trenująca. Jest to 4 680 plików pochodzących z czwartej sesji. Każda osoba ma przypisane 3 sekwencje po 13 ujęć.
3. Część **testowa** zawiera 100 nowych klas reprezentowanych przez 3 900 zdjęć pochodzących z czwartej sesji. Dla każdej osoby są to 3 sekwencje po 13 ujęć.

MV-Multi-PIE jest bazą większą niż opisywana wcześniej MV-Quiscampi. Zawiera liczne reprezentacje dla każdej osoby z określoną pozą na każdym obrazie. Warunki w których zebrane zostały zdjęcia są ściśle kontrolowane, co jest jej zaletą, ale również i wadą. W zbiorze brakuje zdjęć trudnych, co uniemożliwia testowanie identyfikacji w warunkach rzeczywistych dla obrazów pobranych bez kooperacji z użytkownikiem.

1.7.3 VGGFace i VGGFace2

VGG-Face2 [4] jest bazą upublicznioną w roku 2018 przez badaczy z Uniwersytetu w Oxfordzie z laboratorium Visual Geometry Group (VGG). Stanowi jedną z podstawowych baz danych, które służą do trenowania sieci neuronowych do rozpoznawania twarzy. Jej popularność wynika przede wszystkich z dużej liczby zróżnicowanych próbek dla każdej klasy, przy jednoczesnym zachowaniu stosunkowo dużego stopnia uporządkowania bazy pod względem prawidłowości etykiet osób widocznych na zdjęciach. Warto też wspomnieć, że jest to druga wersja równie popularnego zestawu obrazów VGGFace [22]. Bazy są udostępniane razem z propozycjami modeli do klasyfikacji opracowanych przez to samo laboratorium, dlatego określenie *VGGFace* może się odnosić zarówno do zbioru danych, jak i wytrenowanego modelu. Pierwsza wersja bazy obrazów VGGFace [22] zawiera 2,6 milionów zdjęć należących do ponad 2622 osób. W artykule [22] dotyczącym zbioru zaproponowano klasyfikację twarzy w oparciu o architekturę VGG-16 i VGG-11 [25] oraz wykorzystano trening z funkcją *triplet loss*.

Baza VGGFace2 jest rozszerzeniem pierwszego zestawu, a powiększenie zbioru miało na celu ulepszenie wyników treningu. Baza zawiera znacznie więcej próbek twarzy - jest to 3,31 milionów obrazów należących do 9131 osób. Każda osoba posiada średnio 362,6 próbek, ale liczności klas mogą znacząco się różnić. Ponadto obrazy charakteryzują się większym zróżnicowaniem pozy, wieku i rasy niż w VGGFace. Zdjęcia zostały pobrane sieci z użyciem mechanizmu Google Image Search. W celu minimalizacji szumu w bazie, dane zostały poddane manualnej oraz półautomatycznej inspekcji. Usunięto obrazy z niewłaściwymi identyfikatorami, duplikaty oraz scalono foldery zawierające dane tych samych osób. Wyniki rozpoznawania twarzy na VGGFace2 zostały przestawione w artykule [4] dla modeli zbudowanych w oparciu o architektury ResNet-50 [3] i SENet-50 [26]. Wagi oraz schematy wszystkich sieci dla obydwu wersji baz są dostępne na stronie laboratorium Oxford VGG [27].

Rysunek 1.11: Przykładowe obrazy pochodzące ze zbioru VGGFace2

Zbiór VGGFace2 ma narzucony podział na część estymacyjną i testową. Część estymacyjna zawiera 8631 klas, a testowa 500. Razem z obrazami dostarczono współrzędne prostokątów detekcji oraz położenie 5 punktów charakterystycznych twarzy (oczy, nos i kąciki ust). Zbiór jest stosunkowo wyrównywany jeśli chodzi o zawartość obu płci: 59,3% osób to mężczyźni. Na zbiorze testowym wyznaczone zostały listy 5-elementowych zestawów danych do testowania klasyfikatora pod kątem wrażliwości na zmianę pozy i wieku. W zestawach dotyczących pozy zestaw składa się albo z samych zdjęć frontalnych albo z samych profilowych (lewych lub prawych). Nie jest to podział adekwatny do przyjętego w poniższej pracy, dlatego etykiety nie zostały wykorzystane.

1.7.4 IARPA IJB-C

Benchmark IARPA Janus Benchmark–C (IJB-C) [5] obejmuje protokoły testowe oraz zbiór danych twarzy i został opracowany przez amerykańską agencję rządową zajmującą się projektami naukowymi z zakresu sztucznej inteligencji IARPA (*ang. Intelligence Advanced Research Projects Activity*) w celu weryfikacji algorytmów rozpoznawania twarzy. Protokoły testowe są też częścią globalnego progamu porównania algorytmów detekcji i rozpoznawania twarzy przeprowadzanym przez amerykański urząd standaryzacyjny NIST [28].

Rysunek 1.12: Przykładowe zdjęcia należące do zbioru IARPA Janus Benchmark - C (IJB-C). W zbiorze jest wiele trudnych ujęć twarzy - o zróżnicowanej pozie i osłoniętych.

IJB-C jest jednym z częściej używanych obecnie w artykułach do prezentacji wyników algorytmów detekcji i rozpoznawania twarzy. Wynika to przede wszystkim z faktu, że jest znacznie większy w sensie liczby próbek i osób występujących w bazie od dotychczas używanych scenariuszy testowych. Zawiera zróżnicowane rodzaje danych, w tym obrazy statyczne i wycinki z materiałów wideo. Część z innych standardowo używanych zbiorów testowych takich, jak np. LFW [29] nie stanowi wyzwania dla obecnie proponowanych modeli - najlepsze modele osiągają na tyle wysokie wyniki, że trudno o porównanie różnych metod ze sobą.

W bazie obrazów należącej do IJB-C starano się utrzymać brak wspólnych osób z popularnymi zbiorami do trenowania sieci takimi jak VGG-Face lub CASIA WebFace. Zbiór składa się z różnego typu materiałów, zarówno obrazów statycznych jak i nagrań wideo. Osoby są pokazane na zdjęciach w zróżnicowanych pozach. Do kolekcji starano się dodać osoby o zróżnicowanych zawodach, starając się uniknąć budowania bazy głównie na podstawie osób znanych, co często ma miejsce w pozostałych zbiorach danych. Obrazy posiadają dodatkowe etykiety i szczegółową informację o osłoniętych częściach twarzy. Benchmark ma również 2 poprzednie wersje IJB-A [30] i IJB-B [31] i stanowi ich rozszerzenie. W sumie baza zawiera 3531 osób, 31 334 obrazów (w tym 10 040 obrazów, które nie zawierają twarzy) oraz 117 542 ramki wycięte z 11 779 materiałów wideo.

W ramach benchmarku udostępnione zostały liczne protokoły testowe i to też stanowi o jego uniwersalności:

- Test 1: Weryfikacja 1:1
- Test 2: Weryfikacja 1:1 z próbkami pojedynczymi
- Test 4: Identyfikacja 1:N
- Test 6: Detekcja twarzy
- Test 7: Grupowanie
- Test 9: Pełen protokół identyfikacja + detekcja (end-to-end) - obrazy statyczne i wycinki wideo
- Test 10: Pełen protokół identyfikacja + detekcja (end-to-end) - nagrania wideo
- Test 11: Pełen protokół identyfikacja + detekcja (end-to-end) - materiały mieszane

Test 1: Weryfikacja 1:1 dotyczy zbioru 3531, dla których wyróżniono 23 124 wzorców. W zbiorze znajdują się obrazy statyczne i wycinki wideo. W ramach protokołu testowego wyróżniono 19 557 porównań pozytywnych i 15 638 932 porównań negatywnych.

Test 2: Weryfikacja 1:1 z próbkami pojedynczymi dotyczy zbioru 3 531, dla których wyróżniono 23 124 wzorców. Dane są identyczne jak w teście 1, ale inaczej wygląda podział na wzorce konieczne do wytworzenia. W ramach protokołu testowego wyróżniono 19 557 porównań pozytywnych i 15 638 932 porównań negatywnych.

Test 4: Identyfikacja 1:N Protokół składa się z 2 list próbek galeryjnych G1 i G2 oraz listy próbek-zapytań wraz z wytycznymi dotyczącymi niezbędnych porównań. Galerie składają się odpowiednio z 5588 obrazów, z których należy utworzyć 1772 wzorców (galeria G1) oraz 6011 obrazów, z których należy utworzyć 1759 wzorców (galeria G2). Lista próbek-zapytań (do wyszukania na listach galeryjnych) ma długość 19 593 wzorców dla 127 152 obrazów.

Pozostałe protokoły zostały omówione szczegółowo w dokumencie źródłowym [5], ale nie były stosowane w tej pracy.

		Liczba osób	Liczba wzorców	Liczba obrazów
Identyfikacja 1:N	Galeria G1	1 772	1 772	5 588
	Galeria G2	1 759	1 759	6 011
	Próbki-zapytania	3 531	19 593	127 152

Tabela 1.1: Podsumowanie właściwości protokołu testowego do identyfikacji 1:N z benchmarku IJB-C

1.7.5 Podsumowanie

W poprzednim rozdziale wprowadzone zostały 4 zbiory danych, które są stosowane w badaniach w zależności od potrzeb. Trudno wyróżnić jeden, który zawierałby zarówno wystarczającą liczbę próbek do przeprowadzenia treningu sieci, jak i odpowiednie etykiety i spełniał warunek wielosesyjności dla zdjęć galeryjnych. Własności poszczególnych zestawów danych zostały podsumowane w tabeli 1.2.

Nazwa	Czy zmodyfikowany?	Zawartość	Zastosowanie
MV-Quiscampi	zmodyfikowana baza Quiscampi, zbiór poddany detekcji, ręcznie skorygowane etykiety i wyniki detekcji, dodatkowe ujęcia wycięte w sposób półautomatyczny z nagrań wideo, dodano etykiety pozy	269 osób w zbiorze galeryjnym • 807 próbek galeryjnych ze zdjęć statycznych (trzy pozy: front i $\pm 90°$) • 1 911 próbek galeryjnych z wideo (siedem póz: front i $\pm 90°$, $\pm 15°$, $\pm 45°$) • 2 907 próbek z monitoringu zewnętrznego, różne warunki oświetleniowe i zakłócenia	możliwy podział na podzbiory: galerię i próbki-zapytania zgodny z założeniami projektu, obecność tych samych identyfikatorów w obu podzbiorach
MV-Multi-PIE	podzbiór bazy CMU Multi-PIE, zbiór poddany detekcji	podział na zbiór trenujący, walidacyjny i testowy, w zbiorze trenującym jest 12 480 obrazów należących do 120 osób	liczne próbki pokazujące głowę tej samej samej osoby w określonej pozie przy jednoczesnej kontroli nad innymi czynnikami zmienności, dostępne również liczne przykłady na których ta sama osoba ma głowę w tej samej pozycji
VGGFace	nie	2,6 miliona obrazów należących do 2622 osób, podział na zbiór trenujący i testowy	zbiór o znacznych rozmiarach stosowany do treningu sieci głębokich
VGGFace2	nie	3,31 milionów obrazów należących do 9131 osób, podział na zbiór trenujący i testowy	zbiór o bardzo dużych rozmiarach stosowany do treningu sieci głębokich
IJB-C	nie	31334 obrazów należących do 3531 osób, dodatkowe etykiety i protokoły testowe	zbiór o znacznych rozmiarach stosowany do ewaluacji algorytmów, gotowe protokoły testowe pokrywające standardowe zastosowania

Tabela 1.2: Specyfika zbiorów testowych i uczących użytych w badaniach

Rozdział 2

Identyfikacja twarzy w oparciu o sieci głębokie

2.1 Generacja wzorców twarzy z użyciem sieci splotowych

We wszystkich opisanych w pracy eksperymentach generacja wzorców twarzy wykonywana jest za pomocą głębokich sieci neuronowych. Zarówno próbki galeryjne, które chcemy zarejestrować w referencyjnej bazie danych, jak i próbki-zapytania, na podstawie których chcemy znaleźć osobę w systemie identyfikacji wymagają zakodowania. W naszej pracy będą analizowane **wzorce proste** - tzn. wzorce, które są tworzone na podstawie jednego obrazu oraz **wzorce złożone** powstałe z wykorzystaniem zestawu obrazów. Przy założeniu, że w systemie jedna osoba może być reprezentowana przez więcej niż jedną próbkę, przechowywanie wzorców złożonych jest o wiele bardziej opłacalne niż przechowywanie wzorców prostych, ponieważ skraca to konieczną do przejrzenia listę referencyjną. Poza tym złożenie wzorca z informacji pochodzącej z wielu próbek pozwala się pozbyć szumów, które mogą być wprowadzone przez pojedyncze próbki. Jako, że tematem pracy jest analiza zbiorów wieloujęciowych, wzorce złożone, które zostały utworzone na podstawie tego typu danych, będziemy nazywać wzorcami **multi-view** (MV). Analogicznie wzorce proste będą nazywane **single-view** (SV), ponieważ są generowane na podstawie jednego ujęcia.

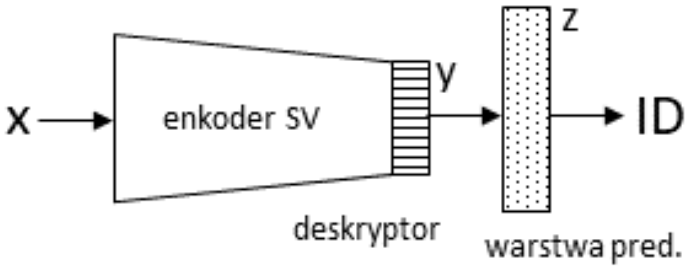
Przez głębokie sieci neuronowe najczęściej rozumiemy perceptrony wielowarstwowe, których celem jest aproksymacja pewnej funkcji f. W przypadku klasyfikatorów jest to odwzorowanie wejścia x na kategorię z, a więc $z = f(x; \theta)$, gdzie θ to parametry sieci (definicja pochodzi z

pracy [32]). W przypadku sieci, których celem jest analiza obrazów naturalnych, korzystamy najczęściej z tzw. **sieci splotowych**. Sieci splotowe (ang. *convolutional neural networks*) są rodzajem sieci głębokich, w których co najmniej jedna z warstwa jest **warstwą splotową** (ang. *convolutional layer*). Warstwy splotowe realizują operację splotu $(x * w)$, która jest działaniem na dwóch funkcjach. W terminologii sieci splotowych druga funkcja nazywana jest często filtrem lub jądrem (ang. *kernel*), a wynik operacji jest określany jako mapa cech (ang. *feature map*). Warstwy splotowe mają własności, które dobrze sprawdzają się przy przetwarzaniu danych o strukturze siatki, takich jak obrazy, ponieważ mogą modelować zależności pomiędzy sąsiadującymi pikselami.

W strukturze sieci głębokich standardowo występują również inne warstwy, w tym **warstwy skalujące** (ang. *pooling*) i **warstwy pełne** (ang. *fully connected*). Działanie warstw skalujących jest kompatybilne z warstwami splotowymi, ponieważ zbierają one wybrane statystyki na podstawie sąsiedztwa o założonym rozmiarze. Może to być warstwa typu maksimum (ang. *max-pooling layer*), która zwraca wartość maksymalną z przyjętego sąsiedztwa, np. z prostokąta o rozmiarze N w przypadku warstwy operującej w dwóch wymiarach. Kolejną częstą stosowaną warstwą skalującą jest warstwa uśredniająca (ang. *average-pooling layer*), która zwraca średnią arytmetyczną dla przyjętego sąsiedztwa. Zestawienie ze sobą warstw splotowych i skalujących przyczynia się do niewrażliwości modelu na niewielkie przesunięcia obiektów w obrazie. Podstawowym typem warstwy, która występuje w sieciach głębokich, a w sieciach splotowych najczęściej znajduje się na końcu jest warstwa pełna. W warstwie pełnej wejście mnożone jest przez macierz wag.

Przez architekturę sieci rozumiemy schemat zawierający kolejność, rodzaj i wielkość warstw, który pozwala na zbudowanie danego modelu. W dalszej części pracy będziemy korzystać przede wszystkim z sieci splotowej o architekturze typu ResNet-50 [3], której schemat budowy wraz z poszczególnymi warstwami został opisany w rozdziale 2.2.1 „Architektura modelu bazowego - ResNet-50". Używając sieci głębokiej w charakterze klasyfikatora ostatnia warstwa sieci pełni funkcję **warstwy predykcyjnej**. Wielkość warstwy predykcyjnej jest zazwyczaj równa liczbie etykiet przyjętych w procesie trenowania sieci. Wyjście o najwyższej wartości jest interpretowane, jako to wskazujące na etykietę klasy. Pomimo trenowania sieci do klasyfikacji zadanego zbioru etykiet, może się pojawić potrzeba ekstrakcji cech analizowanej próbki. Wektor cech nazywany będzie w pracy **deskryptorem**. Deskryptor może zostać zapisany w bazie

danych lub posłużyć do porównania dwóch próbek ze sobą. Bardzo często w charakterze deskryptora wykorzystywane jest wyjście z przedostatniej warstwy sieci splotowej. Część sieci, która dokonuje transformacji wejścia x w deskryptor y nazywana będzie **enkoderem**. Schemat sieci z podziałem na enkoder, deskryptor i warstwę predykcyjną został przedstawiony na grafice 2.1.



Rysunek 2.1: Schemat modelu do klasyfikacji obrazów twarzy. Na podstawie obrazu wejściowego x generowany jest deskryptor y. Na końcu modelu znajduje się warstwa predykcyjna na podstawie której można uzyskać identyfikator osoby. Warstwa predykcyjna ma zazwyczaj długość równą liczbie klas w zbiorze trenującym. W czasie ewaluacji metod kodowania wzorców twarzy wykorzystywane są deskryptory y, które służą jako wzorce proste dla podanych na wejściu sieci obrazów.

Trening głębokich sieci neuronowych odbywa się poprzez aktualizację wartości parametrów sieci θ. Sam algorytm uczenia opiera się na pewnego rodzaju optymalizacji. Celem optymalizacji jest minimalizacja wybranej **funkcji celu** (ang. *loss function*). Nazwa funkcja celu często jest stosowana wymiennie z funkcją kosztu. Ze względu na nieliniowości występujące w sieciach głębokich, większość w praktyce stosowanych funkcji celu jest niewypukła, co wymusza stosowanie określonych procedur w procesie optymalizacji.

Typowy trening sieci głębokich odbywa się iteracyjnie, a określenie wartości funkcji celu odbywa się dla ograniczonego zestawu danych (ang. *batch*). W eksperymentach opisanych w pracy będziemy stosować metodę **spadku gradientu stochastycznego** nazywaną w skrócie SGD (ang. *stochastic gradient descent*). Podstawowym parametrem wpływającym na optymalizację SGD jest prędkość uczenia α (ang. *learning rate*), która określa jakiej wielkości kroki należy

wykonywać w procesie treningu. Aktualizacja parametrów sieci θ_t w kroku t w metodzie SGD ma następującą postać:

$$SGD : \theta_{t+1} = \theta_t - \alpha \nabla_\theta J(\theta_t; x^k, y^k)$$

$\nabla_\theta J(\theta_t; x^k, y^k)$ jest estymatą spadku gradientu związaną z funkcją celu. W celu przyspieszenia procesu uczenia metodę SGD rozszerza się o czynnik pędu γ (ang. *momentum*) [33]. Czynnik ten akumuluje średnie gradientów z poprzednich kroków algorytmu i wymusza kontynuację optymalizacji w ich kierunku.

$$SGD_{momentum} : \theta_{t+1} = \theta_t - (\gamma \nu_t + \alpha \nabla_\theta J(\theta_t; x^k, y^k))$$

Przy trenowaniu sieci neuronowych często stosuje się również metody optymalizacji z adaptacyjną szybkością uczenia takie jak AdaGrad [34] i Adam [35]. W przypadku eksperymentów opisanych w pracy nie zauważono zysku ze stosowania metody Adam zamiast SGD, dlatego zdecydowano się pozostać przy wersji z użyciem SGD.

2.1.1 Trening modeli

Wybór funkcji celu jest kluczowy dla osiągnięcia założonych efektów uczenia, ale wynika też bezpośrednio z zadania, które chcemy rozwiązać. W przypadku rozpoznawania obrazów naturalnych, takich jak zdjęcia twarzy, można wyróżnić dwa podstawowe podejścia do trenowania modeli opartych o sieci neuronowe:

1. **Uczenie klasyfikacyjne** (*classification learning*) - podstawą do określenia poprawności procesu uczenia jest sprawdzenie, czy dana próbka należy (lub należy z wystarczająco wysokim prawdopodobieństwem) do określonej klasy, a więc generuje na wyjściu z sieci identyfikator odpowiedni do danej klasy - przykładem takiego podejścia jest trenowanie sieci z zastosowanie funkcji entropii skrośnej. Uczenie klasyfikacyjne zakłada najczęściej zamknięty zestaw etykiet klasy na wyjściu modelu.
2. **Uczenie metryczne** (*distance learning*) - trenowanie modelu następuje w oparciu o informacje o odległości deskryptorów w przestrzeni reprezentacji. W tej przestrzeni deskryptory należące do tej samej klasy powinny znaleźć się możliwie blisko, a deskryptory obce daleko od siebie. Celem treningu jest więc rozsunięcie deskryptorów wytworzonych na podstawie próbek obcych lub ściągnięcie do siebie deskryptorów tej samej klasy. W tym

podejściu nie ma problemu związanego z ograniczaniem modelu do zamkniętego zbioru etykiet. Przykładem uczenia metrycznego jest funkcja triplet w treningu trójkami [36] lub funkcje celu stosowane w sieciach syjamskich [37].

3. **Metody hybrydowe** Proces uczenia może być sterowany nie jedną, a kilkoma funkcjami celu. Jeżeli funkcje są addytywne względem zestawu próbek uczących, mogą być ze sobą składane. Poziom wpływu każdego ze składników hybrydowej funkcji celu może być regulowany za pomocą dodatkowych parametrów.

Uczenie klasyfikacyjne z zastosowaniem entropii skrośnej

W przypadku treningu, którego celem jest klasyfikacja obrazów, funkcją stosowaną w charakterze funkcji celu jest **entropia skrośna** (ang. *cross-entropy*). Wyjściem z sieci głębokiej, na podstawie którego obliczana jest funkcja celu, jest wyjście z warstwy predykcyjnej Z. Wielkość warstwy predykcyjnej jest zgodna z liczbą etykiet klas w zbiorze uczącym. Entropia skrośna jest zazwyczaj używana w połączeniu z funkcją softmax, która normalizuje wartości wektora neuronów warstwy predykcyjnej.

$$softmax(Z)_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{2.1}$$

Po przetworzeniu funkcją *softmax* suma wektora Z wynosi 1, a poszczególne wartości można interpretować, jako prawdopodobieństwa przynależności do klasy c powiązanej z danym wyjściem. Na etapie ewaluacji identyfikator klasy $\hat{z}$ jest obliczany na podstawie indeksu największej wartości wektora Z. $\hat{z} = \operatorname{argmax} Z$. Funkcja entropii skrośnej (ang. *cross entropy*) $\mathcal{L}_{CE}$ jest zdefiniowana jako:

$$\mathcal{L}_{CE}(Z, c) = -\sum_{j} t_j \log z_j, \quad \text{gdzie } t_j = \begin{cases} 1 & \text{jeśli } j == c \\ 0 & \text{w przeciwnym przypadku} \end{cases} \tag{2.2}$$

Wektor t_j przyjmuje wartość 1, w przypadku gdy dana próbka należy do klasy j, a 0 w przeciwnym przypadku.

Badanie odległości między deskryptorami na etapie ewaluacji

O ile klasyfikacja jest typowym zadaniem w dziedzinie wizji komputerowej, w rozpoznawaniu twarzy, szczególnie w scenariuszu otwartym, konieczna jest możliwość porównania ze sobą dwóch próbek. Standardowo robi się to przez „odcięcie" ostatniej warstwy predykcyjnej modelu i wykorzystanie wyjścia z warstwy przedostatniej, które może być interpretowane jako deskryptor próbki. Długość deskryptora ma stałą wartość, zależną od przyjętej architektury modelu. Dla użytej w pracy sieci ResNet-50 jest to długość równa 2048. Przed zastosowaniem w systemie wektor deskryptora jest normalizowany z użyciem normy ℓ_2. W ogólności deskryptory wytworzone przez sieć są punktami w K-wymiarowej przestrzeni, nazywanej w pracy przestrzenią reprezentacji R^K (ang. *embedding space*). Po normalizacji ℓ_2. punkty te leżą na hipersferze o promieniu 1. Przy porównywaniu ze sobą deskryptorów będziemy stosować odległość kosinusową. Odległość kosinusowa opisana wzorem 2.3 ma tę własność, że jej wartości leżą w zakresie [0,1], przy czym 0 oznacza duży stopień podobieństwa, a 1 sytuację, gdy obiekty są do siebie niepodobne. Warto jednak zwrócić uwagę, że w przypadku znormalizowanych wektorów metrykę tę łatwo przekształcić w odległość euklidesową.

$$d_{cos}(Y_a, Y_b) = 1 - \frac{Y_a \cdot Y_b}{||Y_a||||Y_b||} = 1 - Y_a \cdot Y_b, \quad \text{dla } ||Y_a|| = 1, ||Y_b|| = 1 \tag{2.3}$$

Inne funkcje uczące

Uczenie funkcją entropii skrośnej nie jest jednym podejściem do treningu sieci neuronowych do rozpoznawania twarzy. W podejściach opartych o uczenie metryczne funkcje celu wyliczane są na podstawie wyjścia z warstwy sieci, która odpowiada za wytworzenie deskryptora. Takie podejścia stosowane są w sieciach syjamskich [37] oraz przy uczeniu trójkami [38]. W obydwu podejściach procedura treningu zakłada określony sposób podawania próbek do sieci. W sieciach syjamskich próbki podawana są parami, a funkcja celu określana jest na podstawie ich deskryptora oraz zgodności (czy należą do tej samej klasy, czy do innych). W przypadku uczenia trójkami wejście do sieci formowane jest w zestaw: próbka referencyjna (ang. *anchor*), próbka pozytywna i próbka negatywna względem referencyjnej. Funkcja celu nazywana *triplet loss* wymusza w procesie treningu przysuwanie do siebie próbek należących do tej samej klasy i odsuwanie próbek obcych. Najtrudniejszym wyzwaniem w tych podejściach jest odpowied-

nia konfiguracja sposobu podawania próbek na wejściu, przy dużych zbiorach danych liczba kombinacji próbek jest ogromna, a brak balansu przy ich wyborze wpływa na wyniki uczenia. W tym celu wprowadzane są dodatkowe heurystyki, które ułatwiają wybór zestawów próbek. Przykładowo w uczeniu trójkami rozróżnia się trójki łatwe, semi-trudne i trudne, z czego tylko dwa ostatnie typy wnoszą istotną informację w procesie treningu.

Istnieje też szeroka klasa rozwiązań dedykowanych do trenowania sieci z uwzględnieniem zróżnicowanej pozy próbek źródłowych. Jednym z takich podejść jest wstrzyknięcie do funkcji celu punktów pośredniczących niezależnych od pozy (ang. *pose-invariant proxy embeddings*) zaproponowane w artykule [39]. Punkty pośredniczące *proxy* to punkty leżące w przestrzeni reprezentacji, a więc w tej samej, w której leżą deskryptory próbek. Punkty te pełnią funkcję atraktorów w procesie uczenia i mogą być zestawione z funkcjami celu jak NCA (*neighborhood component analysis* [40]) i *triplet loss*. Temat zastosowania punktów pośredniczących w systemach identyfikacji twarzy został rozwinięty przez autorkę pracy w artykule [41]. Ich zaletą jest możliwość jednoczesnego trenowania na podstawie deskryptorów pochodzących z sieci jedno- i wieloujęciowych. Metodologia nie jest jednak kompatybilna ze wszystkimi modelami omawianymi w poniższej pracy, tj. z sieciami, które nie mają możliwości wytwarzania deskryptorów pojedynczych (single-view) i dlatego nie została zastosowana w dalszej części. Warto jednak zwrócić uwagę na tego typu podejście w kontekście obrazów wieloujęciowych.

2.2 Bazowy model single-view

Punktem wyjścia do prac nad identyfikacją twarzy w systemie z obrazami wieloujęciowymi jest podstawowy model jednoujęciowy, nazywany również *singleview* i oznaczony skrótem SV. Model bazowy SV pełni 2 ważne funkcje w naszych badaniach, po pierwsze stanowi bazę do porównań z zaproponowanymi metodami, a po drugie może zostać wykorzystany wewnątrz architektur pozostałych metod, jako że większość z nich ma konstrukcję modułową. Model SV jest oparty na architekturze ResNet-50 [3]. Decyzja o wyborze architektury została podjęta po analizie metod, które zajmują pierwsze pozycje w rankingach porównawczych algorytmów rozpoznawania twarzy oraz po przeglądzie architektur kompatybilnych z dostępnymi bibliotekami do uczenia maszynowego. Na moment przygotowania tekstu pierwsze miejsca na liście najlepszych modeli dla benchmarku IJB-C [42] zajmowały między innymi modele oparte

o architekturę ResNet-100 z artykułów [43] i [44], ResNet-50 z [45], 64-warstwowa sieć rezydualna z artykułu [46], model SENet (*squeeze and excitation network*), który jest modyfikacją sieci ResNet-50 z artykułu [4]. Z listy tej można wywnioskować, że różnego rodzaju modyfikacje sieci z rodziny ResNet cieszą się popularnością wśród twórców modeli do kodowania obrazów twarzy oraz że stanowią dobrą podstawę do budowania systemów identyfikacji. Można oczywiście dyskutować, że istnieją lepsze modele, szczególnie te większe i głębsze niż ResNet-50, jednakże wybór podyktowany był również wymaganiami sprzętowymi niezbędnymi w trakcie treningu. Głębsze i bardziej rozbudowane sieci mają więcej parametrów, a co za tym idzie potrzebują więcej zasobów obliczeniowych i czasu do wyuczenia. Co więcej w pracy skupiono się na modelach wieloujęciowych, które są jeszcze bardziej wymagające, jako że operują na zestawach zdjęć, a nie na pojedynczych obrazach. Warto również dodać, że sieć ResNet-50 wchodzi w skład większości najpopularniejszych bibliotek do uczenia maszynowego takich jak PyTorch [47], Tensorflow [48] oraz w repozytorium modeli do rozpoznawania twarzy udostępnionych przez laboratorium Oxford VGG [27]. Architektura ta wchodzi też w skład detektorów obiektów, które są silnie rozwijającą się gałęzią wizji komputerowej. ResNet-50 jest domyślny rdzeniem w metodach detekcji zaimplementowanych w bibliotece Mask-RCNN z pakietu torchvision [47] oraz detectron2 [49]. Jest to argument za dużą elastycznością i szerokimi możliwościami wykorzystania tej architektury. Porównanie wydajności i wyników dla poszczególnych architektur bazowych nie leży w zakresie pracy, ponieważ temat jest bardzo szeroki i cały czas rozwijany - porównanie to wymagałoby ciągłej aktualizacji. Warto jednak zaznaczyć, że przy wykorzystaniu procedur opisanych poniżej można wytrenować inny wybrany model w zależności od potrzeb i wymagań systemowych projektu.

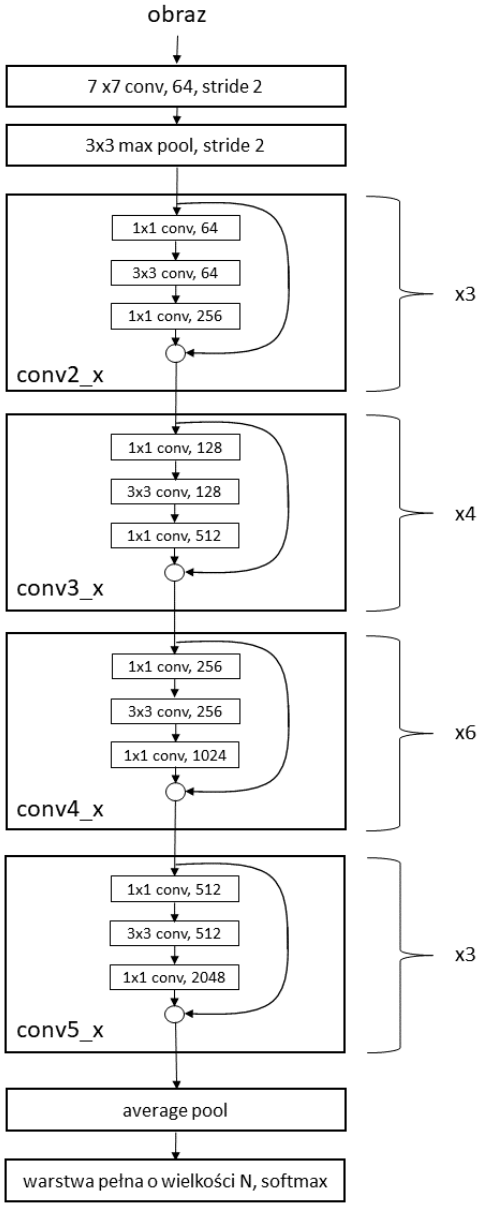
2.2.1 Architektura modelu bazowego - ResNet-50

Modele typu ResNet to tzw. sieci rezydualne, które zostały wprowadzone w artykule [3]. W artykule zaproponowano kilka wariantów modeli typu ResNet, w tym modele o różnych głębokościach: ResNet-18, ResNet-34, ResNet-101 i ResNet-152. W modelu bazowym SV wykorzystano wersję o głębokości 50. ResNet jest typem sieci splotowej, której strukturę można rozbić na powtarzające się bloki nazywane przez autorów *bottleneck blocks*. Między blokami w sieci istnieją dodatkowe połączenia (*skip connections*) realizowane za pomocą funkcji tożsamościowej, które stanowią wyróżnik sieci rezydualnych. Możliwość ominięcia niektórych warstw

poprawia wyniki uczenia sieci oraz pozwala uniknąć problemu zanikającego gradientu. Każdy z bloków zawiera w środku 3 warstwy splotowe, w tym jedną o splocie wielkości 1x1, jedną o splocie 3x3 i ostatnią 1x1. Warstwy o splocie wielkości 1x1 odpowiadają za zmianę wymiarowości map cech. Liczba filtrów splotowych oraz liczność poszczególnych bloków zależy od modelu - dla sieci ResNet-50 została szczegółowo przedstawiona na grafice 2.2 oraz w tabeli 2.1. Ostatnia warstwa sieci jest warstwą pełną i pełni funkcję warstwy predykcyjnej - jej wielkość zależy od liczności klas w zadaniu. Dla zadania klasyfikacji obiektów ze zbioru ImageNet [50], które to zadanie zostało opisane w artykule źródłowym, jest to wielkość równa 1000 klas. Sieci rezydualne w momencie ich wprowadzenia pozwoliły autorom osiągnąć pierwsze pozycje w najważniejszych światowych wyzwaniach z zakresu wizji komputerowej należących do konkursów ILSVRC (ImageNet Large Scale Visual Recognition Challenge) [50] oraz COCO 2015 [51] (detekcja i segmentacja obiektów).

2.2.2 Trening modelu bazowego

Do estymacji wag modelu **bazowy-SV** posłużyła baza VGGFace2 [4]. Baza VGGFace2 składa się z 3,31 milionów obrazów, na których widać twarze należące do 9131 osób. Jest to stosunkowo duża i popularna do trenowania klasyfikatorów twarzy baza. Jej zaletą jest duże zróżnicowanie zawartych w niej obrazów, pod względem pozy, wieku i mimiki. Ponadto zbiór został poddany półautomatycznej inspekcji, dzięki czemu mało jest w nim powtarzających się i źle oznaczonych próbek. VGGFace2 jest drugą wersją równie popularnej bazy zdjęć twarzy VGGFace, która w momencie jej powstania również była często wykorzystywana w treningu sieci neuronowych. Pierwsza wersja VGGFace jest znacznie mniejsza - składa się z 2622 klas (2,6 milionów zdjęć). Więcej szczegółów dotyczących obydwu zbiorów danych zostało omówionych w załączniku 1.7 „Bazy danych i scenariusze testowe". Bazy VGGFace i VGGFace2 zostały zebrane na Uniwersytecie w Oxfordzie, a ich publikacja wiązała się z propozycją modeli sieci neuronowych, które zostały wyuczone z użyciem tych danych. Pierwszy artykuł, który dotyczył bazy VGGFace [52], opierał się na sieci VGG-16 - była to 16-warstwowa sieć splotowa zakończona 2 warstwami pełnymi. Wyniki dotyczące bazy VGGFace2 przeliczane już były na sieciach rezydualnych, w tym dla architektury ResNet-50 i SENet [4].



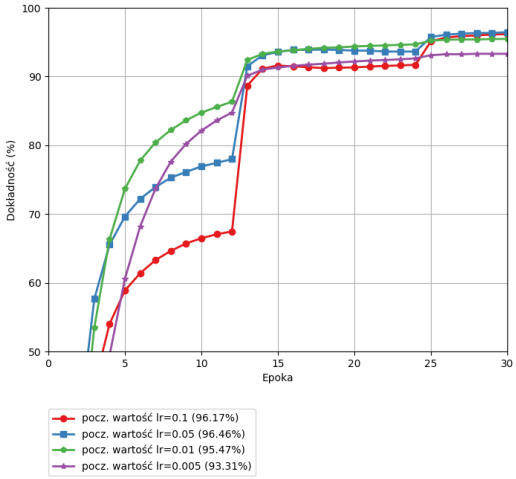
Rysunek 2.2: Schemat sieci ResNet-50 na podstawie opisu [3]. Sieć składa się z bloków określanych w artykule źródłowym jako *bottleneck block*. Bloki zawierają w sobie sieci splotowe o filtrach wielkości 1 oraz 3 oraz połączenia stanowiące skróty między warstwami (*skip connections*).

Nazwa bloku	Rozmiar wyjścia	Schemat bloku
conv1	112×112	7×7, 64, stride 2
		3×3 max pool, stride 2
conv2_x	56×56	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
conv4_x	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
conv5_x	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	1×1	average pool, warstwa pełna o wielkości N, softmax

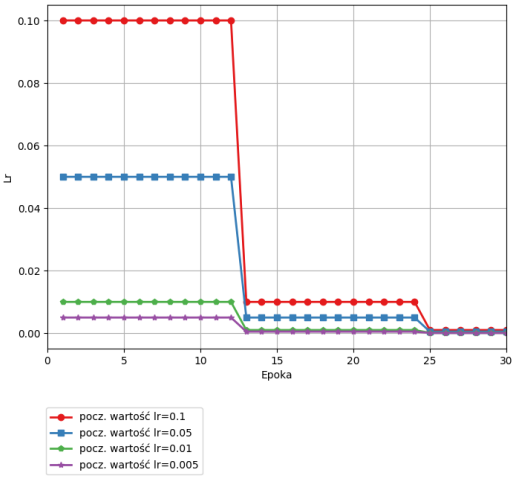
Tabela 2.1: Opis warstw w sieci ResNet-50. W tabeli opisana jest budowa poszczególnych bloków oraz wielkość wyjść w modelu. Opis jest odpowiednikiem grafiki po lewej stronie.

Jako, że w tej pracy również bazowano na sieci ResNet-50 przy trenowaniu modelu sugerowaliśmy się niektórymi ustawieniami parametrów, które zostały opisane w drugim z wymienionych artykułów [4]. Zbiór VGGFace2 posiada arbitralnie przyjęty podział na dane trenujące i testowe. Model **bazowy-SV** trenowany był na części trenującej. W trakcie uczenia obserwowano szereg wskaźników opisujących proces uczenia na zbiorze walidacyjnym, który był utworzony na podstawie części testowej zbioru VGGFace2. W części walidacyjnej znajdowało się 500 klas reprezentowanych przez 2500 obrazów. Lista plików została ograniczona ze względu na konieczność przeliczenia w każdej iteracji (w zbiorze oryginalnie jest ponad 160 tysięcy plików). Zbiór etykiet klas był rozłączny ze zbiorem trenującym, stąd też określony wybór wskaźników. Test końcowy omówiony w kolejnym rozdziale odbywał się na odrębnym zbiorze do ewaluacji (IJB-C). W trakcie walidacji obserwowaliśmy 3 podstawowe wskaźniki: wartość krzywej dokładności identyfikacji CMC dla listy o długości 1 (rank-1) oraz o długości 3 (rank-3).

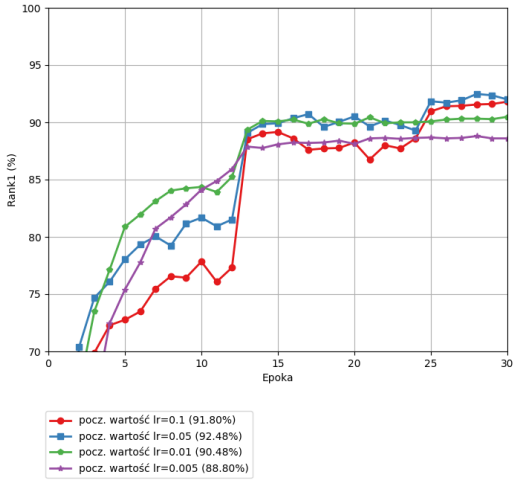
Zaproponowany model **bazowy-SV** był wytrenowany w podejściu klasyfikacyjnym za pomocą entropii skrośnej i funkcji softmax. Przy implementacji rozwiązania wykorzystaliśmy funkcje biblioteki PyTorch [47]. Procedura przetwarzania próbek w trakcie treningu zakłada zmianę wielkości obrazów oraz operacje pozwalające na zróżnicowanie zbioru treningowego w kolejnych iteracjach. Obrazy wejściowe, czyli obrazy twarzy poddane detekcji, skalowane są w ten sposób, że krótsza krawędź obrazu przyjmuje długość 256 z zachowaniem proporcji całego obrazu. Następnie z przeskalowanego obrazu wycinany jest losowy wycinek o rozmiarze 224×224. Od każdego piksela obrazu odejmowana jest średnia wartość koloru dla danego kanału. Są to wartości [131,09, 103,88, 91,49] odpowiednie dla kanałów RGB. Zbiór obrazu jest sztucznie modyfikowany przez dodanie obrazów czarno-białych z prawdopodobieństwem pojawienia się równym 20% oraz obrazów stanowiących odbicie lustrzane. Sieć trenowana była z użyciem optymalizatora SGD (algorytm stochastycznego spadku gradientu). Próbki podawane były w porcjach po 256 ze strategią wyrównania liczności względem klas (balansowanie próbek). Współczynnik szybkości uczenia *learning-rate* ustawiony był początkowo na wartość 0.05 dla modeli trenowanych od początku i zmniejszany 10-krotnie po 12 epokach. Wybór współczynnika uczenia wynika z analizy procesu treningu dla różnych konfiguracji. Przebiegi dla różnych wartości współczynnika *learning-rate* przedstawiono na wykresach 2.3. Na wykresach widać proces uczenia dla 30 epok. Docelowy model był trenowany przez 120 epok.



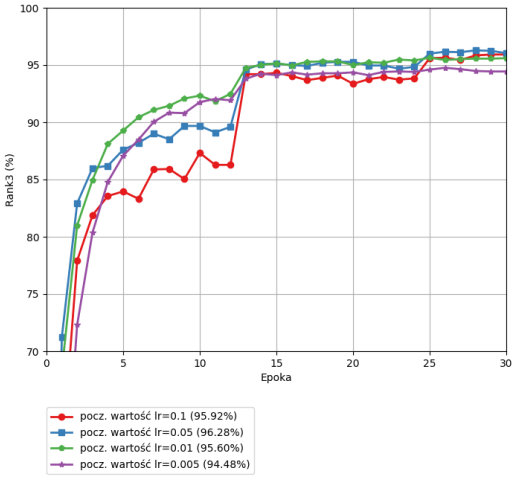
(a) Dokładność na zbiorze trenującym



(b) Zmiany wartości parametru szybkości uczenie *learning-rate*



(c) Współczynnik dokładności identyfikacji Rank-1 na zbiorze walidacyjnym



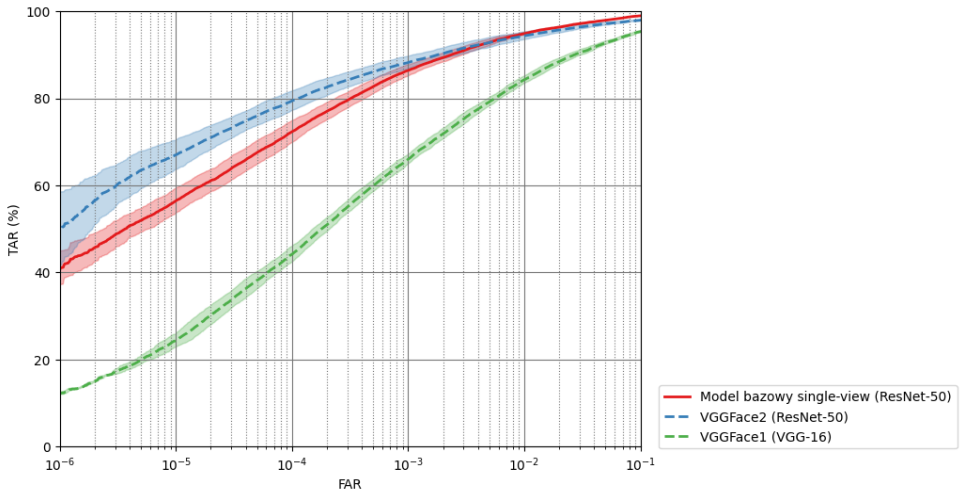
(d) Współczynnik dokładności identyfikacji Rank-3 na zbiorze walidacyjnym.

Rysunek 2.3: Obserwacja zmiany wskaźników w procesie uczenia na zbiorze trenującym 2.3a i walidacyjnym 2.3c, 2.3d. Początkowe wartości parametru szybkości uczenia **learning-rate** zostały ustawione na 0.1, 0.05, 0.01, 0.005. Zmiana parametru następowała po dwuanstu epokach, a jego wartość była zmniejszana dziesięciokrotnie.

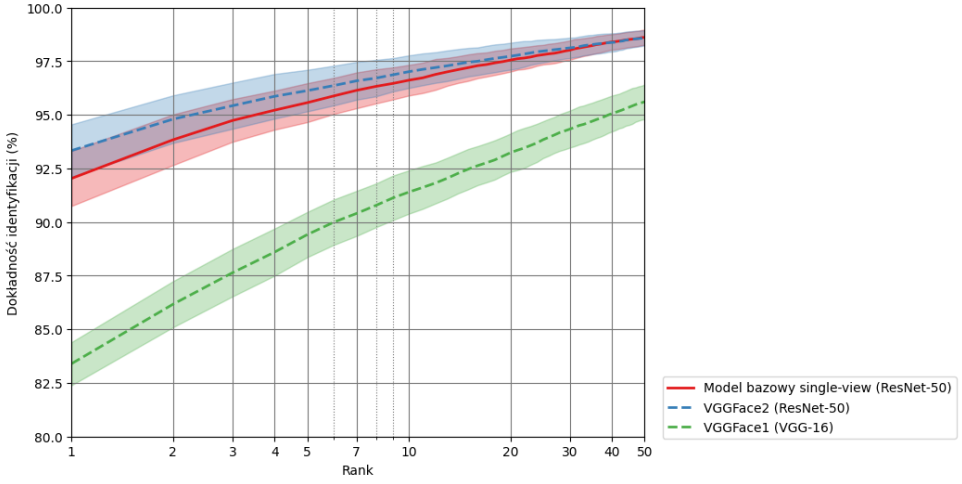
2.2.3 Ewaluacja na zbiorze IJB-C

Po wytrenowaniu modelu **bazowy-SV** na zbiorze VGGFace2 przetestowaliśmy go z użyciem benchmarku IARPA Janus Benchmark–C (IJB-C) [5] (więcej informacji o zawartości zbioru ewaluacyjnego znajduje się w 1.7 „Bazy danych i scenariusze testowe") Do badań użyto dwóch zestawów galeryjnych G1 i G2 i wyników porównań ze zbiorem próbek-zapytań (scenariusz identyfikacja 1:N na materiałach mieszanych). Galerie składają się odpowiednio z 5588 próbek należących do 1772 klas (galeria G1) oraz 6011 próbek należących do 1759 klas (galeria G2), a lista próbek zapytań ma długość 19593 wzorców dla ponad 140 tysięcy próbek. Test wymaga milionów porównań między utworzonymi wzorcami. Uzyskanie wyników dla tego benchmarku jest dosyć czasochłonne, dlatego został on wykonany jedynie dla wybranych modeli. Wyniki pokazane na wykresie 2.4 dotyczą porównań dla listy G1 i G2 (wyniki uśrednione).

Dla tak wykonanych porównań wyliczyliśmy współczynniki jakości w sposób sugerowany przez twórców benchmarku [5]. Wyniki przedstawiono w formie krzywej oceny weryfikacji ROC (ang. *Receiver Operating Characteristic*) - wykres 2.4 oraz krzywej dokładności identyfikacji CMC (ang. *Cumulative Match Curve*) - wykres 2.5. Wyniki dla modelu bazowego SV zestawiono z modelami wytrenowanymi przez twórców baz VGGFace2 i VGGFace - modele wraz z wagami zostały pobrane ze strony z materiałami udostępnionymi przez Samuela Albanie'go [27]. Architektura modelu VGGFace2 w modelach do porównania również jest oparta na architekturze ResNet-50, aczkolwiek różni się detalami implementacyjnymi względem biblioteki PyTorch. Model VGGFace2 został dodatkowo wytrenowany wstępnie na bazie MS-Celeb-1M [53]. Baza MS-Celeb-1M to bardzo duża baza danych (składa się z 10 milionów zdjęć), która jest mocno zaszumiona - posiada wiele niewłaściwych etykiet i redundantnych plików. Z tego powodu jest rzadko stosowana w docelowym procesie nauki bez dodatkowego czyszczenia. W przypadku omawianego modelu właściwa procedura uczenia na zbiorze VGGFace2 była zbliżona do procedury uczenia modelu bazowego SV, ponieważ zostały oparte na tych samych wytycznych z artykułu [4]. Modele wytrenowane na bazie VGGFace2, tj. model **bazowy-SV** i model oznaczony jako VGGFace2 dały znacznie lepsze wyniki identyfikacji w testach IJB-C niż model wytrenowany na starszej wersji zbioru - bazie VGGFace. Warto zwrócić na to uwagę w kontekście, jak duże znaczenie ma wybór i wielkość zbioru trenującego na działanie modeli.



Rysunek 2.4: Krzywe ROC (ang. *Receiver Operating Characteristic*) wyliczone na podstawie testu Identyfikacja 1:N z benchmarku IJB-C. Pokazana wartość jest uśredniona dla porównań względem galerii G1 i galerii G2. Szerokość cieniowania zależy od odchylenia standardowego wyników.



Rysunek 2.5: Krzywe dokładności identyfikacji CMC (ang. *Cumulative Match Curve*) wyliczone na podstawie testu Identyfikacja 1:N z benchmarku IJB-C. Pokazana wartość jest uśredniona dla porównań względem galerii G1 i galerii G2. Szerokość cieniowania zależy od odchylenia standardowego wyników.

Model	Architektura bazowa	Zbiór trenujący	Weryfikacja (TAR)		Identyfikacja (TPIR)	
			FAR=1e-3	FAR=1e-2	Rank 1	Rank 3
Model **bazowy-SV**	ResNet-50	VGGFace2	86.39%	94.87%	92.03%	94.73%
VGGFace2	ResNet-50	VGGFace2 + pretrening na MS-Celeb-1M	88.23%	94.36%	93.33%	95.42%
VGGFace1	VGG-16	VGGFace	66.02%	84.33%	83.39%	87.63%

Tabela 2.2: Wyniki wybranych modeli sieci do rozpoznawania twarzy na podstawie pojedynczych obrazów na podstawie protokołu testowego do identyfikacji 1:N z benchmarku IJB-C

Zaproponowany model SV nie dał wyższych wyników niż model VGGFace2 udostępniony przez autorów bazy, chociaż można je uznać za porównywalne. Oba modele bazują na tej samej architekturze ResNet-50. Jedna z różnic pomiędzy nimi leży w fazie treningu wstępnego. Jakkolwiek nie należy wykluczyć innych powodów, jak lepszy wybór parametrów procesu uczenia, to pretrening jest jedną z przyczyn wyższych współczynników dokładności modelu VGGFace2. Jako, że głównym celem pracy jest optymalizacja zadania pod kątem klasyfikacji wieloujęciowej, model bazowy został uznany za wystarczająco dobry do dalszego rozwoju. Model ten będzie bazą do budowy bardziej złożonych sieci. Większość z zaproponowanych metod ma konstrukcję modularną i moduł SV może zostać wymieniony zgodnie z decyzjami projektowymi.

Wyniki działania enkodera bazowego zostać poprawione za pomocą modyfikacji procesu treningu opisanych we wcześniej wspomnianych artykułach o najlepiej działających sieciach na benchmarku IJB-C. Wśród kierunków rozwoju, które mogłyby poprawić działanie modelu do kodowania twarzy, należy wymienić:

1. Modyfikacja architektury na większą sieć/bardziej dostosowaną do zadania - należy jednak wziąć pod uwagę ograniczenia sprzętowe i obliczeniowe, bardziej skomplikowane architektury mogą mieć większe wymagania obliczeniowe i uzyskana dokładność nie zawsze uzasadni zwiększone koszty. Dla rodziny sieci rezydualnych występują głębsze modele, takie jak ResNet-100 i ResNet-150.

2. Poszerzenie zbioru trenującego. Kilka z metod obecnych na liście liderów dla testu IJB-C opiera się na propozycji nowego czystszego lub większego zbioru obrazów twarzy[43],[44]. Na ile kluczowe są dane trenujące można zobaczyć analizując różnice w wynikach dla sieci wyuczonych na dwóch wersjach bazy VGGFace (wersja pierwsza i VGGFace2) przez ten sam ośrodek badawczy.
3. Modyfikacja procedury uczącej. Omawiane sieci trenowane były z użyciem najprostszej procedury uczącej, tj. klasyfikacja za pomocą entropii skrośnej i warstwy klasyfikacyjnej softmax. Wartymi uwagi są bardziej zaawansowane funkcje celu, które w bardziej kontrolowany sposób wpływają na odległość deskryptorów w przestrzeni reprezentacji. Mogą to być funkcję uczące odległością, jak uczenie trójkami (triplet-loss) [36] oraz sieci syjamskie z funkcjami kontrastującymi (ang. *contrastive loss*) [37]. Istnieje też wiele modyfikacji funkcji softmax, której celem jest zwiększenie marginesu między próbkami różnych klasy. Przykładami takich funkcji jest *A-softmax*[54] lub *angular margin loss* [55].

2.2.4 Poza jako czynnik wpływający na dokładność identyfikacji twarzy

Istnieje wiele czynników, które mają wpływ na obniżenie wskaźników jakości identyfikacji twarzy, są to m.in. rozdzielczość obrazu, zmienność mimiki i wieku osoby widocznej na zdjęciu, częściowe zasłonięcie obrazu oraz obecność okularów i zarostu, a także poza głowy. Wszystkie z tych cech mogą pojawić się na zdjęciach, które powstają bez kooperacji z fotografowaną osobą, np. na nagraniach z monitoringu wizyjnego. Zróżnicowana poza głowy jest czynnikiem, na który postanowiono położyć akcent w poniższej pracy, ponieważ zmienność pozy jest jest charakterystyczna dla zbiorów wieloujęciowych. Obecność ujęć skrajnych wynika z definicji tego typu danych. Od pozy głowy zależy, jaki fragment twarzy jest uchwycony na fotografii. Na zdjęciach frontalnych widoczne są oboje oczu oraz zakres twarzy leżący symetrycznie w obie strony względem osi nosa. Zdjęcia pełnego profilu obejmują lewą lub prawą stronę twarzy od osi nosa do płatka ucha. Różnice w liczbie elementów twarzy widocznych na dwóch porównywanych ze sobą fotografiach może wpłynąć na obniżenie jakości porównania.

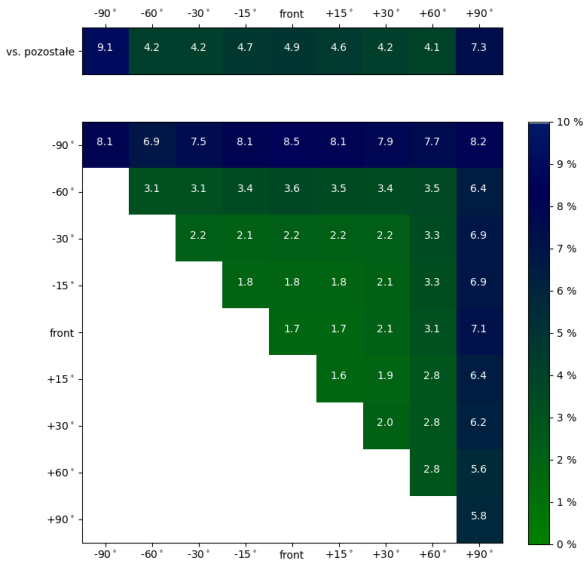
W opisanym poniżej badaniu postawiono przyjrzeć się bliżej wpływowi pozy na wskaźniki jakości porównań wzorców biometrycznych. W tym celu przeprowadzono badania na bazie MV-Multi-PIE, która jest zmodyfikowaną wersją bazy CMU Multi-PIE (została opisana w roz-

dziale 1.7.2 „MV-Multi-PIE"). Baza ta składa się ze zdjęć wykonanych w kontrolowanych warunkach, dzięki czemu łatwiej zbadać ich wpływ na algorytmy. Poszczególne obrazy zostały szczegółowo opisane pod względem występujących czynników zmienności takich jak poza, mimika i oświetlenie, przy zachowaniu stałych pozostałych warunków wykonania zdjęcia, np. tło i urządzenie rejestrujące. Ze zbioru zdjęć wybraliśmy obrazy, w taki sposób, żeby zawierały 9 zestawów danych z określonymi pozami głowy: zdjęcie frontalne ($0°$), lewe $-10°$, lewe $-30°$, lewy $-60°$, lewy pełny profil $-90°$, prawe $+10°$, prawe $+30°$, prawy $+60°$, prawy pełny profil $+90°$.

Rysunek 2.6: Obrazy z bazy testowej CMU Multi-PIE. Do badań wstępnych wybrano 9 kątów, pod którymi widoczna jest twarz w obrazie: zdjęcie frontalne ($0°$), lewe $-10°$, lewe $-30°$, lewy $-60°$, lewy pełny profil $-90°$, prawe $+10°$, prawe $+30°$, prawy $+60°$, prawy pełny profil $+90°$

Przed zakodowaniem zdjęcia zostały przycięte automatycznym detektorem twarzy z pakietu dlib [23]. Do wytworzenia wzorców twarzy użyto modelu **bazowy-SV** opisanego w poprzednim rozdziale. Porównanie zdjęć wykonano licząc odległość kosinusową między deskryptorami wytworzonymi przez model. Błędy zrównoważone wyliczono dla porównań z uwzględnieniem podziału zbioru testowego na kąty (wykres 2.7). Dodatkowo wyliczono również wskaźniki błędów przy porównaniu danego zestawu (pozy głowy) ze wszystkimi pozostałymi danymi, a więc nie należącymi do wybranego zestawu (na wykresach jest to linijka „vs. pozostałe"). Porównanie to miało na celu symulację porównania zdjęć z zestawu z próbkami z baz zdjęć niekontrolowanych (o zróżnicowanej pozie). Wyniki eksperymentu zostały przedstawione na rysunkach: 2.7, 2.8 i 2.9. Pierwszy z wykresów (2.7) pokazuje wartości wskaźnika EER dla porównań, pozostałe pokazują średnie wartości porównań pozytywnych, a więc próbek tej samej klasy o określonym kącie (rys. 2.8) oraz porównań negatywnych, czyli próbek różnych klas o określonym kącie (rys. 2.9).

Na grafice można zaobserwować, że akceptowalna wartość błędu $EER \leq 5\%$ jest zachowana dla porównań próbek, których odchylenie leży w granicach $\pm 60°$. Powyżej tej granicy następuje znacząca degradacja wskaźników błędów. Porównanie ze zdjęciami w pełnym profilu daje

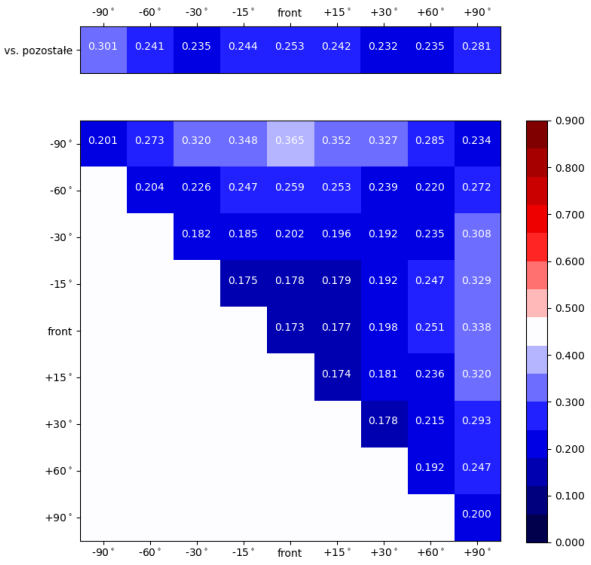


Rysunek 2.7: Wartości błędów zrównoważonych EER dla porównań obrazów twarzy z uwzględnieniem pozy widocznej na zdjęciu. W pierwszym wierszu pokazano porównanie próbek z wybranego podzbioru ujęć głowy ze wszystkimi pozostałymi podzbiorami.

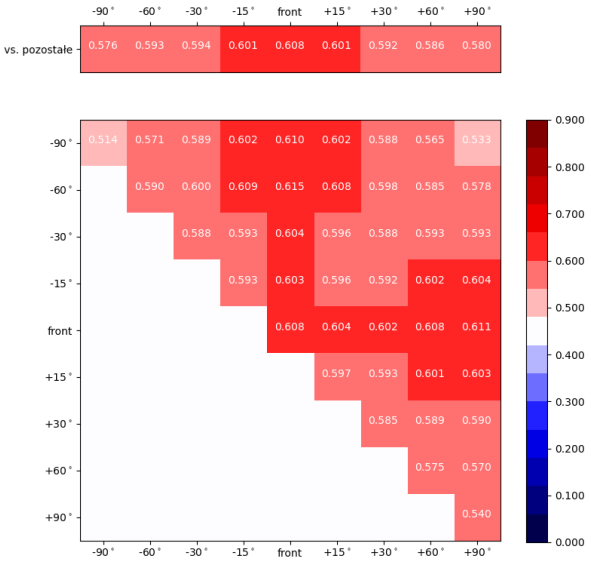
wynik $EER > 6\%$. Można również zauważyć, że istotny jest nie tylko kąt odchylenia, ale również różnica między kątami w dwóch porównywanych obrazach. Najwyższe błędy występują, kiedy następuje konieczność porównania zdjęcia frontalnego ze zdjęciami profilowymi.

Wyniki eksperymentu uzasadniają potrzebę dalszego rozwoju modeli, szczególnie w przypadku, kiedy w systemie występują obrazy zawierających pełne profile (tak jak ma to miejsce w policyjnych bazach sygnalitycznych). W ogólności przy porównaniu wybranego zestawu kątów ze wszystkimi pozostałymi (opcja „vs. pozostałe" na wykresach) wyniki są rzędu EER$> 4\%$. Tego typu porównania miały odwzorować porównania konieczne w nienadzorowanych systemach identyfikacji. Co ciekawe w tym przypadku najniższy wynik błędu nie jest dla zdjęć frontalnych, ale dla kątów $\pm 30°$ i $\pm 60°$. Być może łatwiej jest je zestawić ze zdjęciami w profilu, bo zakres widoczności cech pokrywa się w nich w większym zakresie, aczkolwiek ta teza wymagałaby dokładniejszego sprawdzenia na bardziej reprezentatywnej bazie danych. Na grafikach pokazujących uśrednione wartości miar podobieństwa dla obrazów w wybranych grupach 2.8 i 2.9 można dokonać kolejnych obserwacji w związku z obniżeniem jakości identyfikacji spowodowanym pozą głowy. Na grafice 2.8 pokazano miary dla par obrazów

należących do tej samej osoby, a na grafice 2.9 dla par obrazów pokazujących osoby obce. Im mniejsze wartości odległości, tym próbki uznaje się za podobniejsze do siebie. Co oczywiste wartości pary próbek pozytywnych w ramach tej samej pozy są sobie bliższe niż próbki o różnych kątach, a wartości te rosną wraz ze wzrostem kąta. Ciekawą rzecz można zauważyć dla porównań próbek -90° i $+90^\circ$. Tutaj odległość między próbkami pozytywnymi jest na podobnym poziomie co dla innych podzbiorów, przy stosunkowo wysokim wskaźniku błędu. Ma na niego wpływ niska średnia odległość między próbkami negatywnymi (rys. 2.9). Można to zinterpretować w ten sposób, że dla używanego w teście klasyfikatora twarze widziane z profilu są uznawana za podobne do siebie z racji samej pozy i w pewien sposób wpływa to na pomijanie cech dystynktywnych wynikających z tożsamości osoby. Zdjęcia profilowe przy porównywaniu ich ze wszystkim pozostałymi zdjęciami uzyskują EER równy 9.1% i $7,3\%$, błąd jest dosyć wysoki. Wyższe błędy dla obrazów profilowych mogą mieć wiele przyczyn. Pierwszą z nich może być mniejsza liczność tego typu typu danych w zbiorach trenujących. Nie bez wpływu na jakość rozpoznania pozostaje również działanie detektorów twarzy, które służą do wycięcia ich obszaru z całego obrazu. Detektory tak jak klasyfikatory są lepiej dostosowane do zdjęć frontalnych i gorzej radzą sobie z profilowymi.



Rysunek 2.8: Średnie wartości porównań pozytywnych dla próbek należących do tej samej klasy z uwzględnieniem podziału na pozy głowy w zbiorze. Miarą podobieństwa próbek jest odległość kosinusowa.



Rysunek 2.9: Średnie wartości porównań negatywnych dla próbek należących do tej samej klasy z uwzględnieniem podziału na pozy głowy w zbiorze. Miarą podobieństwa próbek jest odległość kosinusowa.

Rozdział 3

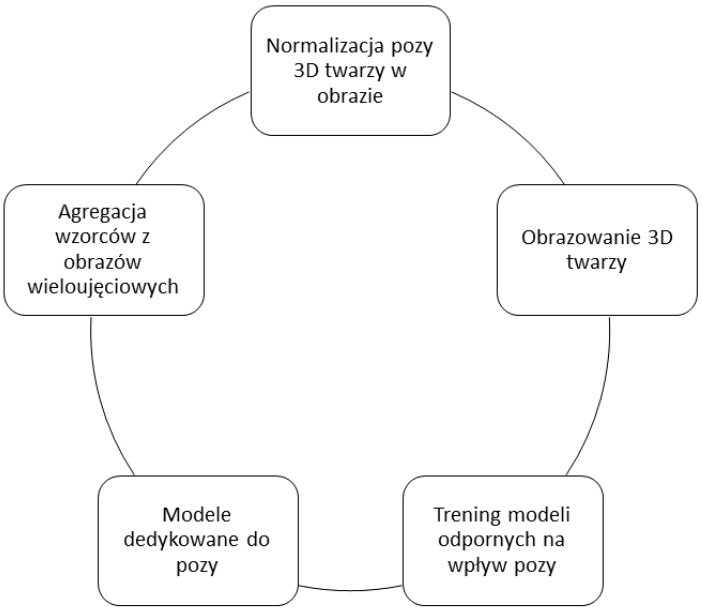
Metody stosowane w identyfikacji obrazów wieloujęciowych twarzy

3.1 Wybrane podejścia

W poprzednim rozdziale wykazano znaczący wpływ zmienności pozy głowy widocznej na zdjęciu na wyniki identyfikacji oraz obniżenie dokładności podpowiedzi zwracanych przez system rozpoznania. Przy porównywaniu zdjęć frontalnych z profilowymi dokładność wyrażona przez wartość błędu zrównoważonego EER spada ponad trzykrotnie w stosunku do porównań wykonywanych pomiędzy zdjęciami frontalnymi (na podstawie wyliczeń przedstawionych w tabeli 2.7). Jednocześnie zasygnalizowane zostało, że porównanie z użyciem zdjęć tego typu może być konieczne w systemach operujących na zbiorach danych zawierających obrazy wieloujęciowe.

W literaturze zaproponowano wiele rozwiązań problemu zmienności pozy w biometrii twarzy, które można zaaplikować między innymi: na poziomie modyfikacji danych wejściowych, dostosowania struktury enkodera wzorców lub też agregacji wzorców pojedynczych w celu stworzenia silniejszych wzorców zbiorczych. Należy zwrócić uwagę, że metody te różnią się pod względem zasobów niezbędnych do realizacji zmodyfikowanego systemu. Często wymagane jest zebranie nowego nietypowego zbioru uczącego w celu wyuczenia modelu, a złożone wzorce mogą różnić się wymaganiami jeśli chodzi o pojemność zajmowaną na nośniku danych,

a zazwyczaj są większe niż wzorce proste. Na podstawie przeglądu literatury oraz dostępnych rozwiązań biometrycznych wyróżniono kilka podejść do przetwarzania zdjęć o wysokim stopniu zróżnicowania pozy w systemach identyfikacji twarzy - metody te zostały zestawione na rysunku 3.1.



Rysunek 3.1: Wyróżnione podejścia do przetwarzania zdjęć niefrontalnych w systemach identyfikacji twarzy

1. **Normalizacja pozy 3D** - jako że największa degradacja jakości identyfikacji porównywania wynika nie z samej pozy, ale z różnicy w pozach porównywanych twarzy naturalnym pomysłem wydaje się korekta kąta twarzy przed zakodowaniem obrazu. Celem korekty jest sprowadzenie porównywanych obrazów do jednej wspólnej pozy - może to być portret frontalny - mówimy wtedy o **frontalizacji** twarzy lub też inny z góry założony kąt widzenia. Normalizacja pozy 3D w odróżnieniu od normalizacji pozy 2D, która wykorzystuje operacje, takie jak przesunięcie, skalowanie i rotacja, wymaga nieliniowej transformacji obrazu oraz estymacji struktury 3D twarzy. Wiąże się to z błędami rekonstrukcji, szczególnie przy dużych zmianach kąta. Część informacji o twarzy niewidocznej w nowej pozycji jest tracona, z kolei część twarzy niewidoczna na obrazie źródłowym musi zostać

uzupełniona. Wiele z metod normalizacji ukierunkowana jest na wyniki akceptowalne wizualnie, co nie zawsze jest tożsame z zachowaniem cech potrzebnych do identyfikacji.

2. **Obrazowanie 3D** - wzorzec twarzy utworzony na podstawie pojedynczego obrazu 2D z zasady może przechowywać tylko informacje widoczne na ujęciu źródłowym. Rozwiązaniem tego problemu mogą być różnego rodzaju struktury dedykowane do przechowywania informacji o obiektach trójwymiarowych. Reprezentacjami 3D twarzy mogą być typowe struktury danych używane w grafice 3D, takie jak chmury punktów i siatki wielokątów lub też struktury dedykowane do biometrii, takie jak postać kanoniczna zaproponowana w artykule [56]. Ich cechą wspólną jest kodowanie własności przestrzennych twarzy. Reprezentacje 3D mogą być rejestrowane za pomocą specjalizowanych urządzeń takich jak skanery 3D lub mogą być rekonstruowane na podstawie obrazów 2D (jednego lub wielu obrazów tej samej osoby). Typowe struktury danych 3D są cięższe od obrazów 2D - potrzebują więc więcej zasobów do przechowywania informacji.
3. **Modele zagregowane** - złożenie informacji 3D może nastąpić nie tylko na poziomie danych wejściowych, tak ma to miejsce w przypadku obrazów 3D, może również zostać wykonane na poziomie wzorców lub ekstrakcji cech obrazu. Modele składane z zestawu zdjęć powinny tworzyć silniejszy i mniej zaszumiony wzorzec niż wzorce pojedyncze. W literaturze rozważane są różne podejścia do tematu agregacji, wśród nich jest złożenie deskryptorów twarzy metodami statystycznymi lub struktury sieci realizujące agregację - zwracają zagregowamy deskryptor na podstawie zadanej sekwencji obrazów.
4. **Modele dedykowane do pozy** - ekstraktory cech zoptymalizowane ze względu na typ ujęcia w obrazie mogą lepiej sobie radzić niż ekstraktory wytrenowane na danych ogólnych (dowolne ujęcie). Powielenie modeli zwiększa jednak złożoność systemu oraz wymaga zaproponowania metody scalania decyzji z różnych modeli i/lub estymacji kąta głowy na zdjęciu w celu właściwego ich zastosowania. Problemem może być również mniejsza dostępność danych uczących (zbiorów, w których oprócz zdjęć frontalnych jest dostatecznie dużo przedstawicieli innych widoków).
5. **Modele niezależne od pozy** - istnieje wiele propozycji metod, których celem jest likwidacja wpływu zmienności pozy w obecnie funkcjonujących modelach. Jednym z podstawowych podejść jest odpowiednie zbalansowanie danych treningowych, które zlikwiduje tendencje systemu do gorszego radzenia sobie z danymi, które są w mniejszości. Zaproponowano również wiele procedur uczenia sieci, które mają na celu uwzględnienie

zróżnicowania pozy, jak np. metoda treningu z użyciem punktów pośredniczących [39], która została zastosowana do treningu systemu identyfikacji przez autorkę w metodzie opisanej w publikacji [41].

6. **Metody mieszane** - metody wymienione powyżej mogą być łączone i składane w jeden system identyfikacji. Przy łączeniu istnieje ryzyko nałożenia na siebie dodatkowych błędów, z którymi związana jest każda z tych metod oraz nadmiarowego zwiększenia złożoności całego systemu i należy wziąć to pod uwagę przy projektowaniu

Decyzja o wyborze podejścia do przetwarzania zdjęć niefrontalnych twarzy jest nieunikniona, jeśli chcemy osiągnąć dobre wyniki identyfikacji na zbiorach, które mogą zawierać zróżnicowane ujęcia głowy, szczególnie ujęcia skrajne, takie jak zdjęcia profilowe.

3.2 Automatyczna korekta pozy

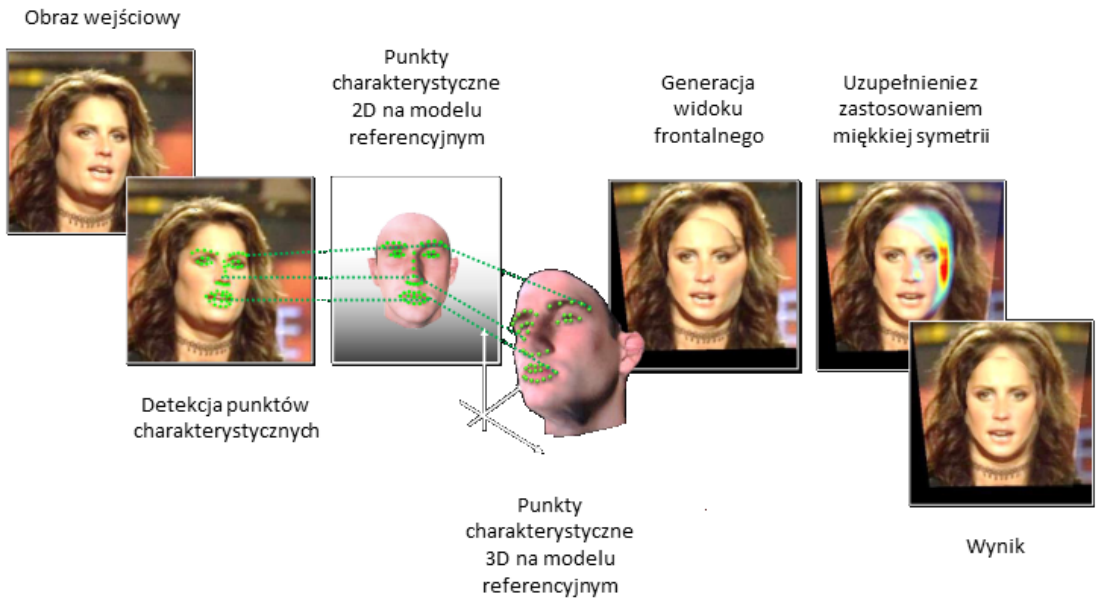
Celem normalizacji pozy jest ujednolicenie położenia, pozy oraz wielkości twarzy w obrazie, a co za tym idzie zmniejszenie jej wpływu na wyniki porównania dwóch próbek ze sobą. Przez normalizację pozy 3D rozumiemy przetworzenie obrazu w taki sposób, żeby głowa widoczna była pod pewnym ustalonym kątem biorąc pod uwagę wszystkie 3 możliwe osie obrotu. W klasycznych metodach identyfikacji biometrycznej, takich jak eigen-face [8] i twarze Fishera [9], wymaganym jest, żeby twarz często twarz była widoczna w pozie frontalnej. W przypadku sprowadzania wizerunku do pozy frontalnej sformułowanie normalizacja może być stosowane wymiennie z określeniem frontalizacja. Definicja twarzy frontalnej pozornie może wydawać się oczywista, mimo to została dokładnie określona w normach dotyczących biometrii twarzy ISO/IEC 19795-1 [14]. Portret frontalny jest zdefiniowana przez kierunek linii oczu oraz tzw. linię Frankfurt Horizon (przechodzącą przez przód prawego ucha i oko) w odniesieniu do osi pionowej. Wprowadzenia głębokich sieci splotowych do klasyfikacji twarzy zwiększyło odporność modeli na przesunięcia położenia obiektu w obrazie, z mniejszym jednak stopniu zmniejszyło wrażliwość na zmianę pozy. O ile w normalizacji 2D wykorzystywane są jedynie afiniczne przekształcenia obrazu, tj. skalowanie, obrót i translacja, o tyle w normalizacji 3D niezbędne są transformacje nieliniowe. W przypadku zdjęć twarzy, w których występują znaczne odchylenia, operacja normalizacji może się przyczynić do dużych zniekształceń, utraty części danych (część twarzy widoczna tylko w ujęciu źródłowym) oraz konieczności sztucznego uzu-

pełnienia danych nieobecnych (cześć twarzy niewidoczna na ujęciu źródłowym). Normalizacja 3D jest użyteczną techniką, ale powoduje narzut na obliczenia potrzebne przy przetwarzaniu obrazów oraz generuje błędy wynikające z niedokładnej rekonstrukcji.

3.2.1 Przegląd metod korekty pozy zaproponowanych w literaturze

Typowa procedura korekty pozy składa się z trzech kroków. Pierwszym z nich jest rekonstrukcja modelu 3D twarzy na podstawie obrazu 2D. Wśród metod rekonstrukcji modelu 3D można wyróżnić te oparte o położenie punktów charakterystycznych twarzy oraz te które nie wymagają tej informacji. W przypadku użycia metod opartych o wykrywanie punktów, takich jaką jest metoda frontalizacji Hassnera [57] znajdowane są punkty korespondujące między obrazem wejściowym, a modelem referencyjnym. Model referencyjny służy do projekcji punktów na nową płaszczyznę obrazu o zadanym kącie względem głowy, np. na portret frontalny. Zrzutowany obraz poddawany jest dodatkowym operacjom, które mają zapewnić bardziej naturalny wygląd. Widok jest wygładzany, a fragmenty twarzy niewidoczne są uzupełniane na podstawie przyjętych założeń, np. związanych z symetrią ludzkiej twarzy. W metodach korekty pozy zaproponowanych w nowszych publikacjach procedura często nie jest podzielona na etapy i jest w całości realizowana przez sieci neuronowe, np. przez sieci generatywne [58], [59], [60].

Ciekawe studium na temat frontalizacji i konieczności jej stosowania w systemach biometrycznych można znaleźć w artykule [10]. Autorzy próbują w nim odpowiedzieć na pytanie, czy w systemach opartych na głębokich sieciach splotowych korekta pozy twarzy w ogóle jest konieczna oraz do jakiego stopnia korekty jest ona możliwa. W swoich badaniach stosują metodę frontalizacji Hassnera [57] oraz metodę własną. Jak już zostało wspomniane metoda Hassnera opiera się na zastosowaniu referencyjnego modelu 3D głowy złożonego z punktów charakterystycznych. Analogiczne punkty charakterystyczne znajdowane są w przetwarzanym obrazie twarzy i z ich użyciem piksele są mapowane do modelu 3D. Brakująca informacja jest uzupełniana z założeniem symetrii twarzy. Metoda własna autorów [10] różni się sposobem rekonstrukcji. Obraz wejściowy twarzy dopasowywany jest do referencyjnej mapy głębi, a nie do referencyjnego modelu 3D. Punkty twarzy są rzutowane z użyciem analogicznej mapy głębi dedykowanej do rzutu frontalnego, co pozwala na zachowanie większej szczegółowości na poziomie pikseli przy projekcji. Następnie autorzy przeprowadzają szereg eksperymentów na



Rysunek 3.2: Procedura frontalizacji w metodzie Hassnera [57]. Grafika pochodzi z materiałów konferencyjnych autora ze strony `https://talhassner.github.io/home/`. Na rysunku pokazano kolejne kroki metody, w tym: lokalizację punktów charakterystycznych, generację widoku frontalnego i przetwarzanie końcowe.

podstawie obrazów twarzy poddanych korekcie. Wnioski, jakie zostały wysunięte były takie, że frontalizacja poprawia wyniki pod warunkiem, że zarówno zbiór trenujący jak i testowy przetwarzane są w identyczny sposób. Jednak poprawa ta nie zawsze uzasadnia dodanie narzutu obliczeniowego związanego z modułem korygującym pozę. Co więcej frontalizacja nie zakładająca symetrii twarzy daje nieco lepsze wyniki od rekonstrukcji symetrycznej.

Przy eksperymentach przeprowadzonych w tej pracy zdecydowano się zastosować metodę *unsup3D* zaproponowaną w artykule [60], która jest nowsza od omówionych metod i nie wymaga detekcji punktów charakterystycznych. Metoda odtwarza model 3D twarzy lub innego obiektu przy założeniu jego przybliżonej symetrii na podstawie pojedynczego obrazu. Konstrukcja algorytmu opiera się na wielomodalnej sieci enkoderów i dekoderów, które są trenowane jednocześnie w trybie uczenia nienadzorowanego. Wymienione moduły to:

- sieć typu enkoder-dekoder do generacji tekstury (albedo)
- sieć typu enkoder-dekoder do generacji mapy głębi (depth)
- sieć typu enkoder-dekoder do określania stopnia pewności (confidence)

- sieć typu enkoder do odtwarzania informacji o kącie widzenia
- sieć typu enkoder do odtwarzania informacji o cieniowaniu

Wewnątrz sieci tworzona jest reprezentacja 3D, którą można zrzutować na płaszczyznę o zadanym nachyleniu. Przy generacji widoku nakładana jest na siebie informacja o teksturze i cieniowaniu właściwa dla wybranego kąta.

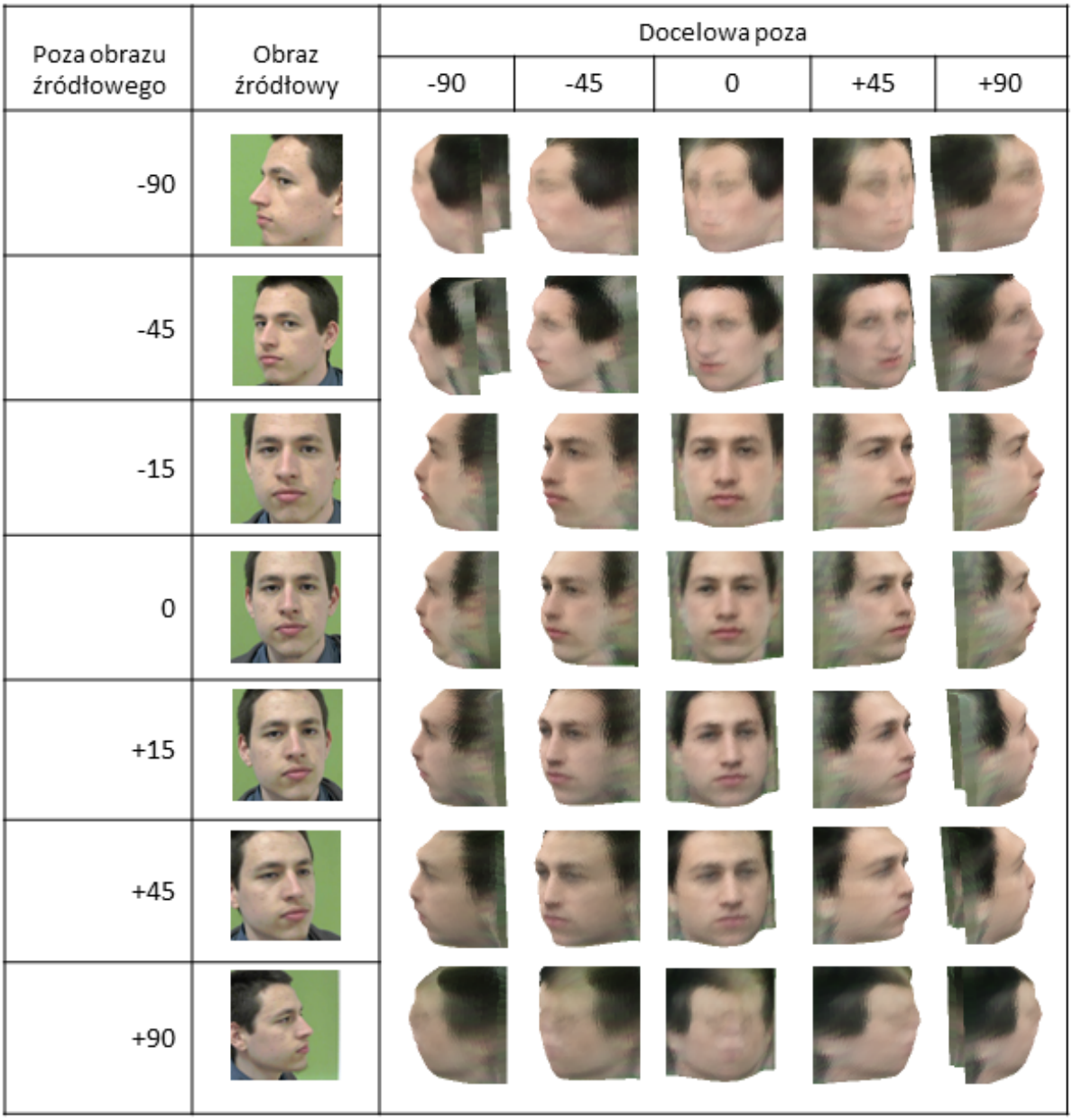
3.2.2 Opis eksperymentów i wyniki

Za pomocą biblioteki udostępnionej przez twórców metody metody unsup3D został wygenerowany zestaw sztucznych widoków na podstawie zbioru MV-Quiscampi opisanego w rozdziale 1.7.1. Do badań wykorzystano dwa podzbiory bazy MV-Quiscampi, tj. Galeria i Galeria-z-Wideo, ze względu na to, że część ta posiada etykiety ujęć i pełny zestaw kątów dla każdej osoby w bazie. Na podstawie każdego zdjęcia zostało wytworzonych 5 nowych sztucznych widoków: $-90°$, $-45°$, $0°$, $+45°$, $+90°$. Tak więc przykładowo dla każdej osoby w zbiorze w której bazowo jest siedem widoków powstało 35 sztucznych widoków. Efekty można zobaczyć na rysunku. 3.3. Jak widać metoda nie jest idealna i najlepiej radzi sobie ze zmianą na portret frontalny przy niewielkiej różnicy względem ujęcia źródłowego. Obrazy zostały przeskalowane do rozmiaru 64×64 pikseli, na którym działają użyte modele.

W pierwszym kroku testów na właściwym systemie identyfikacji porównano dokładność wyszukiwania obrazów frontalnych w bazie złożonej z obrazów frontalnych lub sztucznych widoków frontalnych wygenerowanych na podstawie obrazów o innym kącie źródłowym, tj. pełnym lewym profilu ($-90°$), pełnym prawym profilu ($+90°$), lewym półprofilu ($-45°$) oraz prawym półprofilu ($+45°$). Przeprowadzono szereg eksperymentów, w tym:

- porównanie obrazów naturalnych z obrazami skorygowanymi ze zbioru galeryjnego
- porównanie obrazów skorygowanych z obrazami skorygowanymi ze zbioru galeryjnego.

Wyniki identyfikacji w formie krzywych CMC przedstawione są na rysunku 3.4. Zastosowanie widoków sztucznych w połączeniu z oryginalnymi frontalnymi drastycznie pogorszyło wyniki wyszukiwania.



Rysunek 3.3: Wyniki korekty pozy za pomocą biblioteki **unsup3d** dla obrazów o zróżnicowanej pozie źródłowej. Obrazy wygenerowano na podstawie zbioru MV-Quiscampi. Zestaw póz źródłowych składał się z kątów: $-90°$, $-45°$ $-15°$, $0°$, $+15°$, $+45°$, $+90°$, gdzie $0°$ to zdjęcie frontalne, a wartości dodatnie związane są z prawym profilem. Zestaw wygenerowanych kątów składał się z: $-90°$, $-45°$, $0°$, $+45°$, $+90°$.

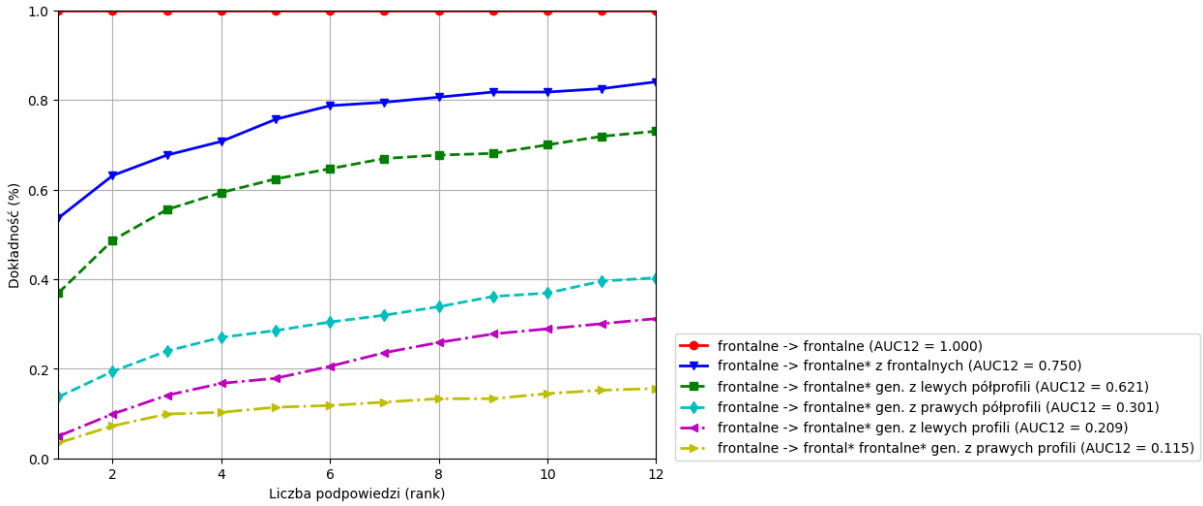
Kolejnym możliwym zastosowaniem korekty pozy jest sprowadzanie do wspólnego widoku obrazów o zróżnicowanej pozie przy wyszukiwaniu. W drugim scenariuszu założono, że chcemy wyszukać zdjęcie profilowe w bazie złożonej ze zdjęć frontalnych. Ze wcześniejszych testów wiemy, że powoduje to obniżenie dokładności względem porównań zdjęcie frontalne do frontalnego. W naszym eksperymencie, w którym przetestowano wyszukiwanie, identyfikacja za pomocą oryginalnych obrazów profilowych (lewego lub prawego) wśród zdjęć frontalnych była zestawiona z:

- wyszukiwaniem sztucznym zdjęciem frontalnym wygenerowanym na podstawie profilowego wśród zdjęć profilowych
- wyszukiwaniem oryginalnym zdjęciem profilowym wśród sztucznych zdjęć profilowych wygenerowanych na podstawie zdjęć frontalnych

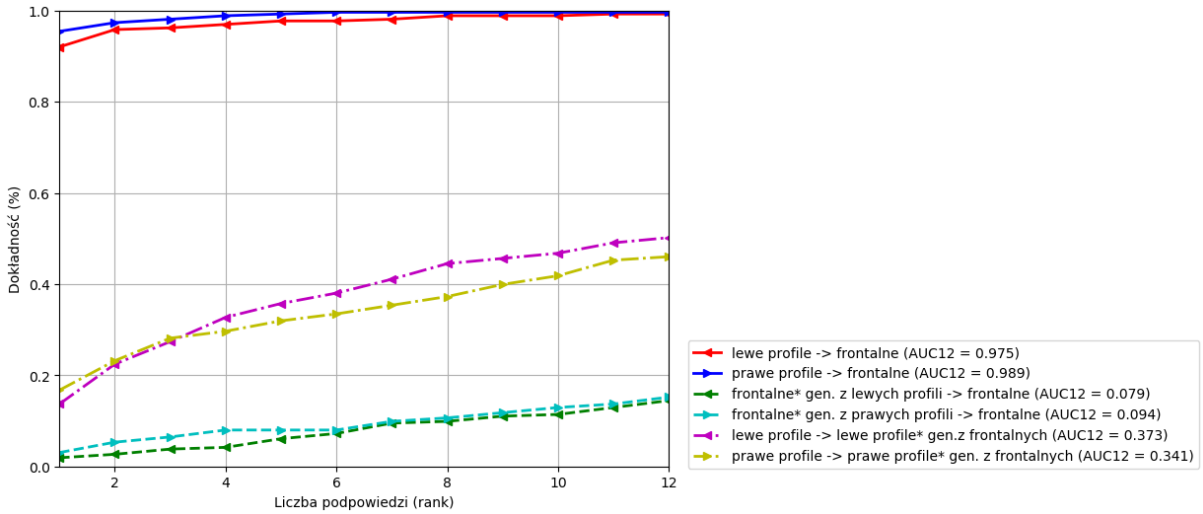
Wskaźniki identyfikacji dla eksperymentów na skorygowanych obrazach przedstawione zostały na rysunku 3.5 i są nieporównywalnie gorsze od operacji na obrazach oryginalnych. Uzyskane wyniki nie uzasadniają użycia metod korekty pozy w niefrontalnych bazach identyfikacyjnych, choć mamy świadomość, że przed całkowitym ich wykluczeniem należałoby sprawdzić więcej metod z tej klasy. Metoda korekty pozy zastosowana do generacji naszego zbioru (unsup3d) również mogłaby zostać ulepszona poprzez retrening modeli i zwiększenie ich rozdzielczości. Zadanie było trudne, ponieważ korekta była wykonana przy znaczących różnicach w pozie. Warto jednak zauważyć, że metody tego typu powodują duży narzut na obliczenia i tworzenie bazy wzorców. Wiele z nich jest zoptymalizowanych pod kątem wizualnie akceptowalnych wyników, a nie przenoszenia cech potrzebnych do identyfikacji - następuje wygładzenie obrazu w czasie projekcji i jak zauważają autorzy [10] utrata istotnych cech o wysokiej częstotliwości.

Podsumowanie eksperymentów z metodami automatycznej korekty pozy

Metody automatycznej korekty pozy nie zostały użyte finalnie w pracy, jakkolwiek przedstawienie wyników związanych z zastosowaniem metody wydawało się autorce istotnym punktem badania dostępnych podejść do analizy zdjęć niefrontalnych twarzy. Sztuczna zmiana pozy głowy w obrazie jest jedną z bardziej oczywistych odpowiedzi na problemy z porównywaniem zdjęć w różniących się ujęciach. Warto jednak zwrócić uwagę na błędy rekonstrukcji, które występują w wynikach oraz ich znaczący wpływ na dokładność identyfikacji.



Rysunek 3.4: Identyfikacja na podstawie obrazów frontalnych w bazie złożonej z widoków frontalizowanych metodą unsup3d na podstawie różnych widoków źródłowych. Metoda została zestawiona z identyfikacją za pomocą obrazów oryginalnych. Oznaczenie (*) obok nazwy widoku w opisie oznacza sztucznie wygenerowane ujęcie



Rysunek 3.5: Identyfikacja na podstawie obrazów profilowych w bazie obrazów frontalnych z zastosowaniem korekty pozy metodą unsup3d. W pierwszym scenariuszu obrazy zostały skorygowane po stronie wyszukiwania, a w drugim po stronie bazy. Oznaczenie (*) obok nazwy widoku w opisie oznacza sztucznie wygenerowane ujęcie.

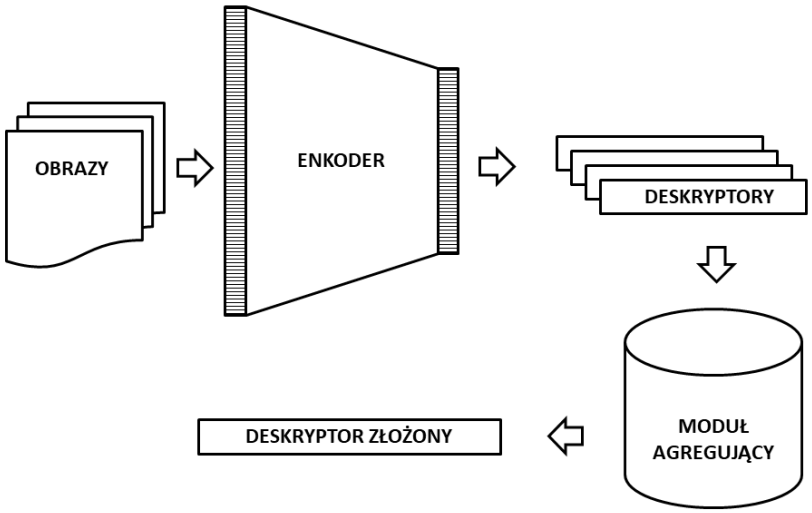
Wiele z metod korekty jest optymalizowana pod kątem „atrakcyjności dla ludzkiego oka", a nie przenoszenia cech potrzebnych do identyfikacji biometrycznej, stąd też wyraźny spadek dokładności zwracanych wyników.

3.3 Agregacja modeli

Agregacja modeli jest kolejnym z zaproponowanych i szeroko stosowanych podejść do analizowania wieloujęciowych próbek twarzy. Model zagregowany zbiera informacje ze zbioru zdjęć tej samej osoby i na ich podstawie tworzy jeden zbiorczy deskryptor twarzy. Intuicja, która stoi za tym podejściem stanowi, że cechy biometryczne zebrane z wielu próbek, w tym potencjalnie z próbek zawierających widok głowy z różnych kątów widzenia, dostarczy więcej informacji do systemu niż dane pochodzące z próbki pojedynczej. Koncepcja ta obejmuje swoim zakresem wiele rodzajów metod. Podstawowym ich wyróżnikiem jest miejsce scalenia informacji z poszczególnych obrazów wejściowych (podział na podstawie artykułu [61]). Może to nastąpić na wejściu modelu i wtedy mówimy o **wczesnym scaleniu** (*ang. early fusion*) lub na samym jego końcu przy zastosowaniu późnego scalenia (*ang. late fusion*). Drugie podejście wymaga stosunkowo mało modyfikacji względem architektur sieci stosowanych w kodowaniu jednoujęciowym - najczęściej polega na dodaniu na końcu modelu jednoujęciowego modułu, który będzie wykonywał scalenie. Punkt scalenia można zaprojektować praktycznie na każdym poziomie wielowarstwowej sieci neuronowej. W przypadku sieci, które mają taki punkt wewnątrz swojej struktury mówimy o **scaleniu pośrednim** (*ang. middle fusion*). Zaletą późnego scalenia jest możliwość wykorzystania wewnątrz struktury modeli jednoujeciowych. Dane do trenowania tego typu modeli są łatwiej dostępne i bardziej powszechne w przeciwieństwie do zbiorów wieloujęciowych.

Schemat systemu agregacji opartego o późne scalenie przedstawiony jest na rysunku 3.6. Na wejściu do systemu podawany jest N-elementowy zbiór obrazów przedstawiających jedną osobę $\{X_i\}$. Każdy z obrazów przetwarzany jest niezależnie przez enkoder E (model jednoujęciowy) w celu wytworzenia deskryptora prostego $y_i = E(x_i)$. Deskryptor złożony, który będzie reprezentował cały zbiór składany jest z deskryptorów prostych za pomocą modułu agregacji $A\ z = A(y_i)$. Składane deskryptory mogą pochodzić z jednego typu enkodera (generalny ekstraktor cech twarzy) lub też z enkoderów dedykowanych do ekstrakcji cech z okre-

ślonych ujęć (np. ekstraktor zdjęć profilowych i ekstraktor zdjęć frontalnych). Moduł agregacji może być zrealizowany za pomocą metod statystycznych lub za pomocą dodatkowych warstw w strukturze sieci.



Rysunek 3.6: Schemat rejestracji zbioru lub sekwencji obrazów twarzy należących do jednej osoby. Moduł agregujący generuje deskryptor złożony na podstawie zestawu deskryptorów prostych. Jest to jedno z proponowanych podejść do agregacji obrazów wieloujęciowych.

Agregacja deskryptorów metodami statystycznymi

Agregacja metodami statystycznymi jest jednym z typowo używanych podejść do składania zbiorów próbek twarzy. Jej zaletą jest brak konieczności dotrenowania i zmian w strukturze sieci. Mając gotowy model, który działa dla pojedynczych próbek możemy dokonać złożenia poprzez agregację deskryptorów (*pooling*) lub agregację wartości porównań (*distance aggregation*) między listą deskryptorów, a próbką-zapytaniem. Jest wiele metod, które mogą posłużyć do składania deskryptorów lub porównań przy wyborze warto się jednak kierować prostotą rozwiązania, szczególnie przy składaniu porównań, tak żeby nie wydłużyć czasu działania całego systemu identyfikacji. Przy wyborze zestawu metod inspiracja została zaczerpnięta z artykułu [62]. Artykuł dotyczy agregacji wycinków twarzy z materiałów wideo i jest kompleksowym przeglądem metod, które mogą mieć zastosowanie w tym przypadku. W trakcie przeglądu nie były jednak brane pod uwagę pozy głowy w obrazie, ten czynnik zostanie wzięty

pod uwagę w eksperymentach autorki. W pierwszym z wymienionych podejść do agregacji (*pooling*) wzorzec złożony jest obliczany na podstawie zestawu deskryptorów należących do tej samej osoby. Miara podobieństwa stosowana do zbudowania listy wynikowej w systemie identyfikacji jest liczona względem wzorca złożonego. W badaniach nad identyfikacją stosuje się podział zbioru danych na dwa podzbiory: zbiór galeryjny G oraz zapytania Q. Warto zwrócić uwagę, że złożenie jest możliwe dla próbek pochodzących z obu tych podzbiorów. Dyskusja nad składaniem wzorców rozpocznie się od metod składania próbek należących do galerii, bo tak najczęściej wygląda scenariusz identyfikacji osoby w bazie. Wśród metod agregacji deskryptorów próbek pochodzących od tej samej osoby (*pooling*) wyróżniliśmy następujące metody:

1. średnia arytmetyczna deskryptorów (określana w skrócie jako **Średnie**)
2. mediana deskryptorów (określana w skrócie jako **Mediany**)

Miarę niepodobieństwa między próbką zapytaniem x_q, a zbiorem próbek galeryjnych X_g, który został zagregowany za pomocą średniej arytmetycznej można zapisać w następujący sposób:

$$\text{Średnia} : d(X_g, x_q) = d(\frac{1}{N} \sum_{x_i \in X_g} x_i, x_q)$$

gdzie d jest przyjętą miarą metryką odległości (w pracy jest to odległość kosinusowa), a N liczbą elementów zbioru X_g. W analogiczny sposób liczona jest mediana z próbek galeryjnych.

W przypadku drugiego z wymienionych podejść (*distance aggregation*) składane są miary, a nie wartości deskryptorów. W ramach tego podejścia przetestowane zostały cztery rodzaje złożeń:

1. średnia odległość od próbek (**Avg-dist**)
2. najmniejsza odległość od próbek (**Min-dist**)
3. największa odległość od próbek (**Max-dist**)
4. odległość *softmin* (**SoftMin-dist**)

Uśrednioną odległość między zbiorem próbek galeryjnych X_g, a próbką zapytaniem x_q (**Avg-dist**) można opisać w następujący sposób:

$$\text{Avg-dist} : d(X_g, x_q) = \frac{1}{N} \sum_{x_i \in X_g} d(x_i, x_q)$$

Dla pozostałych funkcji agregujących wyprowadzenie analogicznych wzorów wydaje się trywialne. Spośród wymienionych najmniej oczywista jest funkcja *softmin* zaproponowana w artykule [63]. Jako że w cytowanym wcześniej przeglądzie [62] osiągnęła wyższe wyniki niż pozostałe metody (znacząco wyższe w przypadku weryfikacji obrazów i porównywalne w przypadku identyfikacji) pojawi się również w poniższym eksperymencie. Funkcja *softmin* dla odległości między próbkami ma następującą postać:

$$\text{SoftMin-dist} : d(X_g, x_q) = \frac{1}{B}\sum_{\beta=0}^{B} s_\beta(X_g, x_q)$$

$$s_\beta(X_g, x_q) = \frac{\sum_{x_i \in X_g} d(x_i, x_q)\epsilon^{-\beta d(x_i, x_q)}}{\sum_{x_i \in X_g} \epsilon^{-\beta d(x_i, x_q)}}$$

Sugerowany w artykule zakres współczynnika β to $[0 - 20]$. Po serii testów wstępnych w naszym eksperymencie zastosowano zakres $\beta = [0 - 5]$ ze względu na mniejszą liczbę potrzebnych obliczeń, a jednocześnie brak wpływu na wyniki końcowe. Testy opisane przez autorów zestawienia [62] zostały przeprowadzone na bazie IJB-A i wskazały, że najlepszymi podejściami są: uśrednienie próbki (**Średnia**) oraz średnia odległość (**Avg-dist**) i odległość *softmin* (**SoftMin-dist**) od próbek.

W opisanym poniżej eksperymencie sprawdzono metody składania wzorców z zastosowaniem wytrenowanej wcześniej sieci neuronowej dla widoków pojedynczych **bazowy-SV**. Test przeprowadzono na bazie MV-Quiscampi (opis bazy w załączniku 1.7 „Bazy danych i scenariusze testowe"). Zbiór galeryjny w MV-Quiscampi to 269 klas, w każdej klasie występują dokładnie trzy obrazy pokazujące głowę w innej pozycji (jedno zdjęcie frontalne i dwa pełne profile). Zbiór próbek-zapytań to 190 klas, zawiera 2907 zdjęć, przy czym występuje w nich zróżnicowana liczba próbek w ramach klasy. Sieć **bazowy-SV** generuje na wyjściu wektor cech o długości 2048 elementów. Wektory porównywane są ze sobą używając odległości kosinusowej. Jako punkt odniesienia policzono miary podobieństwa próbek-zapytań do wybranej próbki twarzy, tj. jedna próbka-zapytanie względem pojedynczej próbki galeryjnej frontalnej, jedna próbka-zapytanie względem pojedynczego prawego profilu oraz jedna próbka-zapytanie względem pojedynczego lewego profilu.

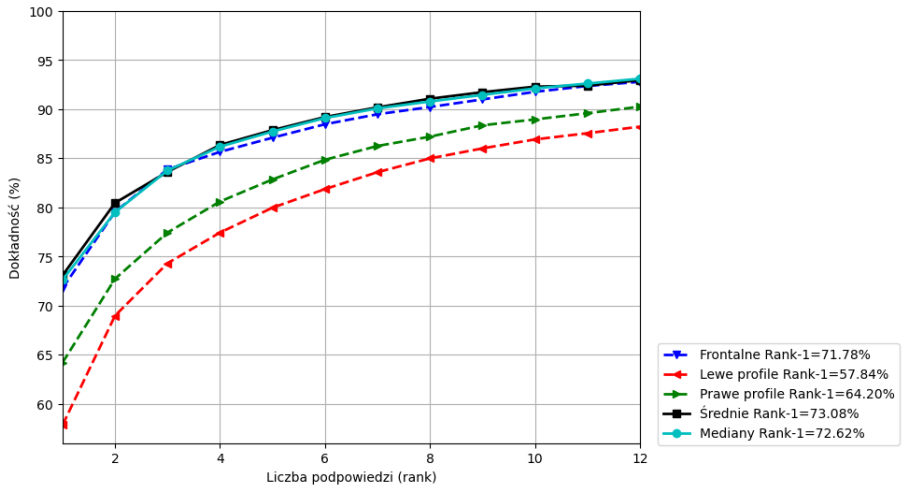
W badaniu przeprowadzonym na zbiorze galeryjnym MV-Quiscampi (wykres 3.7) agregacja deskryptorów pomogła uzyskać lepsze wyniki identyfikacji w stosunku do systemu opartego o

zdjęcia frontalne. Najlepsze wyniki, przyjmując za kryterium wyboru wartość współczynnika Rank-1, otrzymano dla galerii stworzonej z deskryptorów średnich - Rank-1 wyniósł 73,08% oraz dla median Rank-1 wyniósł 72,62%. Przy czym wybór średniej arytmetycznej deskryptorów wydaje się być właściwszy, jako że sama operacja jest mniej wymagająca obliczeniowo niż obliczenie mediany. Uzyskane wyniki były minimalnie lepsze niż współczynnik Rank-1 dla systemu opartego o zdjęcia frontalne. Dla bazy złożonej z prawych pełnych profili o lewych pełnych profili wyniki były znacząco gorsze. Były to odpowiednio Rank-1 równe 64,20% oraz 57,84%. Z tego badania również wynika, że zdjęcia profilowe stanowią duże wyzwanie dla systemów identyfikacji twarzy.

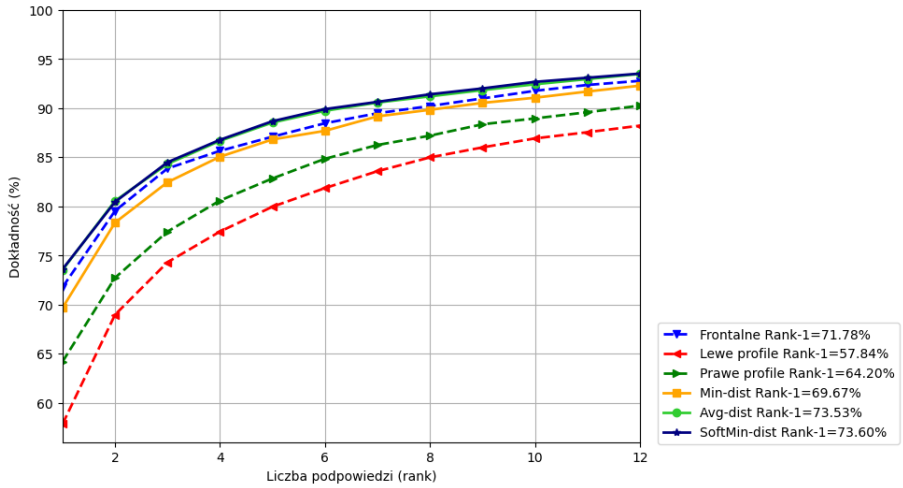
W drugim teście sprawdzono metody składania miar odległości (wykres 3.8). Najlepsze wyniki uzyskano przy złożeniu **SoftMin-dist** (funkcja *softmin*) i **Avg-dist** (średniej). Są to odpowiednio wartości Rank-1 równy 73,60% dla SoftMin-dist oraz 73,53% dla Avg-dist. Wyniki te są minimalne wyższe niż dla średnich deskryptorów z poprzedniego badania (Rank-1 równy 73,08%). Warto jednak podkreślić, że takie rozwiązanie wymaga przechowywania w bazie deskryptorów składowych i ma większą złożoność obliczeniową, ponieważ każdorazowo musimy policzyć odległości do wszystkich deskryptorów należących do zestawu. W powyższym badaniu zestawy galeryjne składały się ze zdjęcia frontalnego i dwóch profili, a więc zawierały dwa trudne ujęcia. W takiej konfiguracji agregacja deskryptorów nie podwyższa znacząco współczynników jakości identyfikacji w stosunku do systemy opartego o zdjęcia frontalne.

Agregacja metodami statystycznymi po stronie próbek-zapytań

W poprzednim rozdziale rozważane były sposoby składania wzorców utworzonych na podstawie obrazów należących do zbioru galeryjnego. Obrazy te mają etykiety, które określają kąt widzenia głowy. Złożenie tego typu obrazów może zostać zinterpretowane jako próba generalizacji twarzy niezależnie od kąta widzenia. Obrazy ze zbioru zapytań nie mają takich etykiet, aczkolwiek ich składanie również może mieć wpływ na jakość wykonanego porównania. W zbiorze zapytań wydzielonym w ramach bazy MV-Quiscampi występuje więcej próbek w ramach jednej klasy, średnio jest to 15 obrazów, a maksymalnie 73.

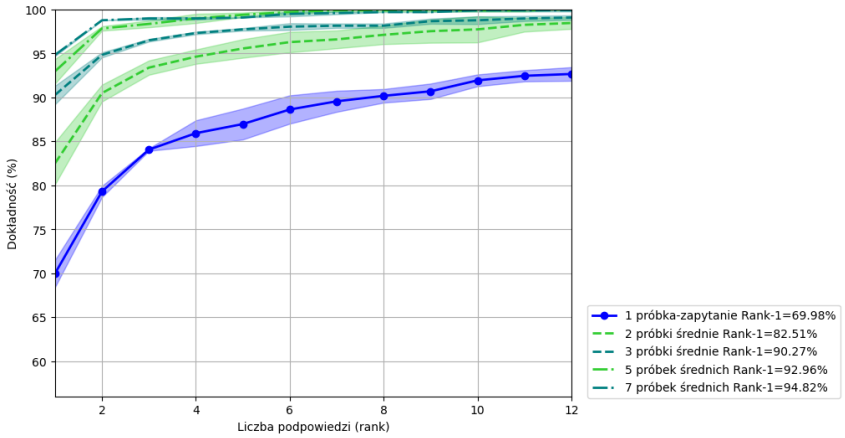


Rysunek 3.7: Agregacja prosta deskryptorów. Identyfikacja twarzy na zbiorze MV-Quiscampi. Na bazę referencyjną składają się odpowiednio: same zdjęcia frontalne, same zdjęcie profilowe lewe, same zdjęcie profilowe prawe lub wzorce złożone ze średniej arytmetycznej trzech deskryptorów i mediany.



Rysunek 3.8: Agregacja prosta miar podobieństwa deskryptorów. Identyfikacja twarzy na zbiorze MV-Quiscampi. Na bazę referencyjną składają się odpowiednio: same zdjęcia frontalne, same zdjęcie profilowe lewe, same zdjęcie profilowe prawe lub wzorce złożone ze średniej arytmetycznej trzech deskryptorów i mediany

W kolejnym eksperymencie sprawdzono agregację dla różnej liczby próbek użytych do złożenia po stronie zapytań. Przyjęto zestawy o następującej liczności: dwie, trzy, pięć i siedem próbek. W teście galeria złożona była ze zdjęć frontalnych, a deskryptory po stronie próbek-zapytań były uśredniane. Na wykresie 3.9 można zobaczyć, że dokładność systemu rosła wraz ze wzrostem liczby próbek w zestawie, co jest zgodne z intuicją.



Rysunek 3.9: Agregacja prosta wyszukiwanych próbek (zbiór Q) na zbiorze MV-Quiscampi. W bazie referencyjnej G użyto zdjęć frontalnych. Po stronie zapytań (zbiór Q) zastosowania deskryptory (**Średnie**) z zestawów o różnej liczności. Zestawy złożone były z losowo wybranych dwie, trzy, pięć i siedem próbek należących do tej samej klasy.

Agregacja deskryptorów - podsumowanie

Intuicja stojąca za pomysłem składania wzorców twarzy przy wyszukiwaniu w pewnym zbiorze referencyjnym nazywanym również galerią była oparta o założenie, że większa liczba próbek źródłowych ułatwi generalizację cech twarzy należących do konkretnego użytkownika. Po przeprowadzeniu testu okazało się, że jest ona tylko częściowo słuszna. Agregacja próbek twarzy po stronie zbioru zapytań rzeczywiście poprawiła wyniki. Była to poprawa o odpowiednio 22,4% dla dwóch próbek składanych, 33,8% dla trzech próbek składanych i 43,5% dla pięciu. Budowa wzorca zbiorczego na podstawie zdjęć profilowych pogorszyła wyniki identyfikacji. Takie zjawisko może mieć różne przyczyny:

1. Zdjęcia pełnego profilu zawierają mało informacji o strukturze twarzy i nie są tak przydatne w procesie rozpoznania, jak zdjęcia frontalne

2. Klasyfikatory twarzy (oraz detektory twarzy) nie są dobrze dostosowane do przetwarzania obrazów profilowych (w bazach do trenowania przeważają zdjęcia frontalne lub bliskie do frontalnych) i nie potrafią dokonać właściwej ekstrakcji cech dla tego typu danych.
3. Wśród obrazów-zapytań dominują obrazy pokazujące twarz w pozycji bliskiej do frontalnej
4. Wybór zbioru galeryjnego składającego się ze zdjęcia frontalnego i dwóch pełnych profili może nie być najlepszym decyzją o strukturze bazy referencyjnej. Profile przestawiają pozycję skrajną - być może zastosowanie zdjęć półprofilowych byłoby lepszym rozwiązaniem

3.4 Modele agregujące

W poprzednim rozdziale omówiliśmy identyfikację twarzy na podstawie zestawu widoków 2D wykonanych pod różnymi kątami z zastosowaniem agregacji deskryptorów. Deskryptory złożone zostały wskazane jako właściwy kierunek rozwoju dla tego typu systemu - pokazaliśmy, że ich użycie w systemie wyszukiwania poprawia wyniki i dokładność podpowiedzi. Agregacja była dotychczas wykonywana metodami statystycznymi (najlepsze wyniki stały za wyborem średniej arytmetycznej deskryptorów) - nic jednak nie stoi na przeszkodzie, żeby moduł agregujący był zintegrowany ze strukturą sieci. Agregacja może być realizowana przez jedną lub więcej warstw wewnątrz wielowarstwowej sieci neuronowej. W poniższym rozdziale zajmiemy się implementacją bardziej skomplikowanych struktur i modułów do agregacji.

Sieć agregująca otrzymuje na wejściu modelu zestaw obrazów należących do tej samej osoby $\{x^{<i>}\}$ składający się z V widoków. Na wyjściu wytwarzany jest deskryptor złożony $y^{<MV>}$ i/lub identyfikator wskazanej osoby. Do wyuczenia sieci potrzebne są zbiory danych zorganizowane w odpowiedni sposób, tzn. takie w których istnieje możliwość pogrupowania obrazów jednej osoby w podzbiory lub sekwencje o liczności V. W przeciwieństwie do wcześniejszych eksperymentów określona struktura zbioru jest potrzebna nie tylko na etapie ewaluacji, ale też do wytrenowania nowych modułów sieci neuronowej, stąd istotne jest aby poszczególne osoby mogły być reprezentowane przez wiele zestawów ujęć. Na potrzeby scenariusza testowego wystarczą dwa zestawy - jeden do galerii i jeden w charakterze zapytania.

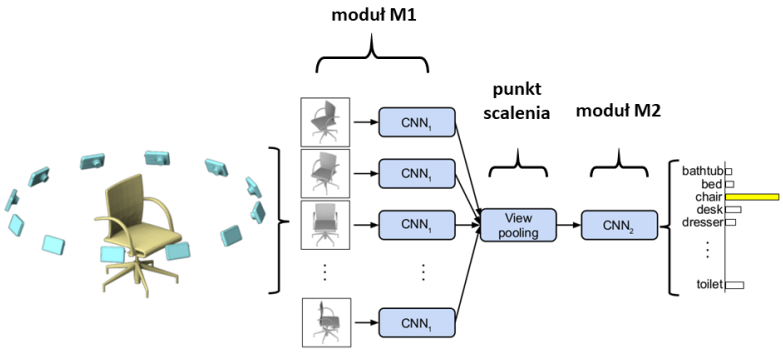
Cechą modeli agregujących jest, że wszystkie niezbędne do scalenia operacje są wykonywane wewnątrz struktury sieci. Pomimo to modele te mogą być budowane w sposób modułowy, tzn. że da się wyodrębnić w nich część, w której następuje ekstrakcja cech pojedynczego obrazu, punkt scalenia informacji oraz część w której następuje predykcja dla zbioru. Taka budowa gwarantuje możliwość częściowego transferu wag z wcześniej wyuczonych modeli jednoujęciowych (SV). Jest to o tyle ważne, że ze względu na wymagania stawiane zbiorom treningowym, modele SV mogą zostać wytrenowane wstępnie na zbiorach o znacząco większej liczności niż modele MV.

Metody oparte o sieci agregujące zostały zaproponowane do klasyfikacji obiektów 3D w oparciu o widoki 2D wytworzone przez zrzutowanie obiektu 3D na płaszczyznę. Jedną z pierwszych tego typu metod były wieloujęciowe sieci splotowe MVCNN (ang. *multi-view cnns*) [1]. Metoda jest na tyle popularna, że doczekała się wielu modyfikacji, jak LatentMVCNN [64] i AutoMVCNN [65]. Sieci MVCNN stały się również integralną częścią innych bardziej rozbudowanych metod [39] do klasyfikacji obiektów 3D . Kolejną siecią, która również służy do agregacji widoków, ale ma nieco inny pomysł na ich połączenie jest RotationNet [2], która dodatkowo w sposób nienadzorowany uczy się kodu widoku źródłowego.

3.4.1 Sieci MVCNN

Sieci typu MVCNN (ang. *multi-view cnns*), czyli wieloujęciowe sieci splotowe zostały wprowadzone w artykule „Multi-view Convolutional Neural Networks for 3D Shape Recognition" [1]. Autorzy zaproponowali to podejście jako rozwiązanie problemu klasyfikacji obiektów 3D. Obiekty 3D pochodzące z bazy ModelNet40 [20] na potrzeby eksperymentów zostały przekształcone do obrazów 2D, które pokazywały obiekt 3D pod różnymi kątami. Następnie obrazy 2D zestawione w zbiory służyły jako materiał do klasyfikacji wykonywanej przez sieć wieloujęciową. Architektura sieci MVCNN opiera się na koncepcji, którą można zastosować do szerokiej gamy modeli sieci splotowych, takich jak ResNet-50 oraz VGG-16. W artykule źródłowym stosowana była sieć VGG-M. Bazowy model jednoujęciowy (SV) dzielony jest na dwa moduły: M_1 i M_2. Pomiędzy częściami znajduje się punkt scalenia widoków przedstawiających jeden obiekt realizowany za pomocą warstwy scalającej. W pierwszej części M_1 poszczególne obrazy przetwarzane są niezależnie przez moduł sieci o współdzielonych wagach. W drugiej

części M_2 przetwarzane są dane scalone. Na wyjściu sieci znajduje się warstwa predykcyjna, która zwraca wynik dla każdego zestawu widoków.



Rysunek 3.10: Architektura sieci MVCNN. Bazą grafiki jest ilustracja z artykułu „Multi-view Convolutional Neural Networks for 3D Shape Recognition" [1], do której dodano podpisy w języku polskim spójne z wyjaśnieniami w tekście. Metoda MVCNN stosowana jest do klasyfikacji obiektów trójwymiarowych pochodzących z bazy ModelNet-40

Dane muszą być zorganizowane w V-elementowe zestawy widoków przy założeniu, że każdy z obiektów jest reprezentowany przez wszystkie V widoków. Kolejność podawania widoków do sieci nie ma znaczenia dla treningu i ewaluacji. W cytowanym artykule używano zestawów o liczności 12 i 80, a więc dużo większych niż liczba widoków przyjęta w zestawach pokazujących ludzką twarz. Punkt scalenia w architekturze zaproponowanej przez autorów był realizowany za pomocą warstwy typu max-pool (operacja typu maximum względem wektorów cech poszczególnych widoków). Autorzy odrzucili operację uśredniania, jako dającą gorsze wyniki końcowe. Różne mogą być również decyzje projektowe dotyczące umiejscowienia punktu scalenia. Autorzy MVCNN wskazali warstwę poprzedzającą warstwę predykcyjną (przedostatnia warstwa modelu) jako najlepsze miejsce. Przy wyborze istotny jest fakt, że im dalej w sieci umieścimy punkt scalenia, tym więcej wag możemy przetransferować z modelu pojedynczego *single-view*. W artykule uzyskano lepsze wyniki dla sieci MVCNN niż dla sieci bazowej SV (sieć pojedyncza o tej samej architekturze), w której uśredniano deskryptory dla poszczególnych widoków.

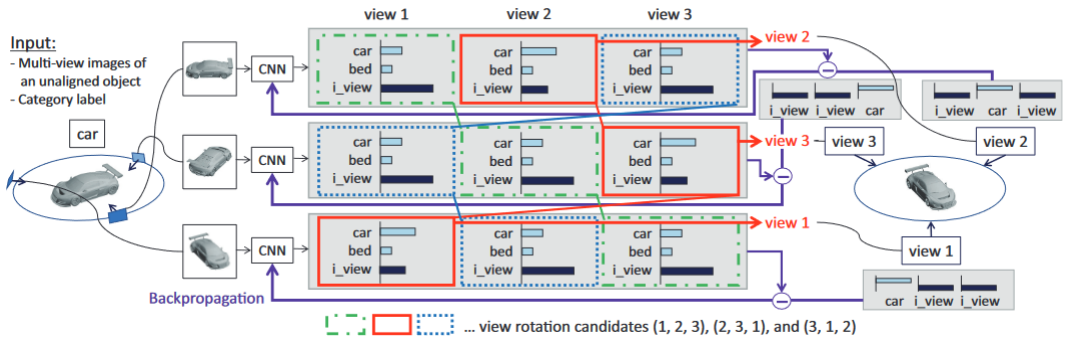
Koncepcja sieci MVCNN jest stosunkowo prosta w implementacji. Wykorzystuje wiele elementów z architektur jednoujęciowych, dlatego trening można rozpocząć od modelu wytrenowa-

nego dla pojedynczych widoków. Wymagania obliczeniowe i poziom trudności związane z wytrenowaniem sieci są porównywalne z tymi samymi sieciami w trybie *single-view*. Należy zwrócić uwagę na pewne ograniczenia związane z przyjmowaniem na wejściu zbiorów ujęć 2D o sztywno określonej liczności (wszystkie zestawy widoków muszą mieć długość V), co może być czasem problematyczne w praktycznych zastosowaniach. Z prostej budowy sieci wynika też wiele innych konsekwencji - łatwo możemy wyodrębnić z niej dane z wcześniejszych warstw i mogą one zostać użyte w systemie zgodnie z interpretacją. Poza wektorem predykcji modelu zagregowanego możemy łatwo uzyskać deskryptory złożone oraz deskryptory pojedynczych widoków.

3.4.2 Sieci typu RotationNet

Sieć o nazwie RotationNet została zaproponowana w artykule „RotationNet for Joint Object Categorization and Unsupervised Pose Estimation from Multi-View Images" [2]. Sieć jest bardziej rozbudowana niż metoda MVCNN. Główny problem, który ma rozwiązać metoda jest identyczny z tym, do którego stosowana jest sieć MVCNN. Metoda RotationNet ma za zadanie dokonać klasyfikacji obiektów 3D (również na zbiorze ModelNet-40) opierając się na projekcjach 2D tych obiektów. Poza zadaniem klasyfikacji sieć estymuje również kąt ujęcia dla poszczególnych projekcji bez konieczności podawania explicite etykiet widoku. W modelu RotationNet również można zastosować różne architektury w charakterze enkodera. Autorzy stosują w badaniach: AlexNet, VGG-M i ResNet-50.

Konstrukcja rozwiązania opiera się na założeniu, że sieć ma za zadanie klasyfikację obiektu reprezentowanego przez V-elementowy zbiór widoków. Etykiety poszczególnych widoków nie muszą być znane ani na poziomie treningu, ani na poziomie ewaluacji. Etykieta widoku przypisywana jest automatycznie w procesie uczenia. Na etapie składania wektora predykcji zmienne kodujące widok są wykorzystywane do odpowiedniego złożenia wyników pośrednich, w tym celu następuje przypisanie sekwencji kodów widoków, która jest optymalna dla zadanego wejścia. Żeby ograniczyć liczbę kombinacji możliwych widoków w artykule przyjęte są pewne założenia związane z ich kolejnością. Jednym z założeń jest, że zbiór widoków jest sekwencją o ustalonym porządku ze zmianą co ustalony kąt, ale bez założonego widoku startowego. Odpowiada to scenariuszowi fotografowania obiektu 3D na stoliku obrotowym. Model



Rysunek 3.11: Struktura sieci RotationNet. Ilustracja pochodzi z artykułu „RotationNet for Joint Object Categorization and Unsupervised Pose Estimation from Multi-View Images" [2]. Dla każdego z pojedynczych obrazów wytwarzany jest zestaw wektorów predykcji wynikający z przyjętego zestawu widoków. Przykładowo dla modelu zakładającego 3 widoki dla każdego z obrazów wejściowych generowane są 3 wektory predykcji.

RotationNet ma specyficzną budowę warstwy predykcyjnej. W metodzie przyjęto, że możliwy zbiór widoków jest V-elementowy. Dla każdego obrazu wejściowego x_i (pojedynczego widoku) wytwarzany jest zestaw wektorów $\{y^{v_a,i}\}$ do klasyfikacji odpowiednich do każdego z przyjętych widoków $\{\hat{v}_i\}_{i=1}^{V}$. Sieć wytwarza oddzielne wektory predykcji dla każdego możliwego ujęcia v_a, do którego może należeć obraz. Poszczególne wektory predykcji mają tak jak w klasycznych metodach długość N, która jest adekwatna do liczby klas. Dodatkowo każdy wektor rozszerzany jest o neuron $IV_{v_a,i}$ kodujący dane ujęcie i określany w artykule jako neuron typu *incorrect view*. Wektor predykcji podlega normalizacji softmax razem z neuronem $IV_{v_a,i}$, który pokazuje dzięki temu stopień niewłaściwości ujęcia v_a względem próbki x_i. Im mniejsza wartość neuronu $IV_{v_a,i}$, tym większe prawdopodobieństwo, że zestaw danych jest właściwy dla obrazu wejściowego. Po uwzględnieniu wszystkich widoków dla każdego pojedynczego obrazu wejściowego x_i powstaje wektor predykcji $\hat{y}_i$ o wielkości V*(N+1), gdzie V to liczba widoków, a N to liczba klas. Dla każdego V-elementowego zestawu widoków $\{x_i\}$ przedstawiających ten sam obiekt wytwarzana jest macierz wektorów o rozmiarze $V \times V(N+1)$. Na końcu modelu RotationNet następuje scalenie wyników z poszczególnych widoków tego samego obiektu biorąc pod uwagę stopnie niepewności wektorów cech. Równanie 3.1 opisuje sposób scalenia pośrednich wektorów predykcji. W ten sposób powstaje zagregowany wektor predykcji $\hat{z}$, który pozwala na obliczenie identyfikatora obiektu. Na podstawie wartości neuronów $IV_{v_a,i}$ można z modelu wyodrębnić informacje o kodach widoków, które zostały

przypisane do poszczególnych ujęć wchodzących w skład zestawu $\{x_i\}$.

$$\left\{\hat{z}, \{\hat{v}_i\}_{i=1}^{V}\right\} = argmax_{\hat{y}_i, \{v_i\}_{i=1}^{V}} \prod_{i=1}^{V} \frac{y_{v_a,i}}{IV_{v_a,i}} \tag{3.1}$$

Zbiór widoków $\{v_a\}$ jest również wybierany w procesie optymalizacji, która w bardziej szczegółowy sposób została opisana w artykule [2]. Zbiór widoków kandydujących spośród których następuje wybór może być ograniczony w celu zmniejszenia liczby potrzebnych obliczeń.

Propozycja rozwiązania jest ciekawa, jako że uwzględnia w procesie klasyfikacji rodzaj ujęcia źródłowego, a jednocześnie nie wymaga automatycznego lub ręcznego określania tej informacji. Co więcej estymacja widoku jest zintegrowana z modelem, co może być przydatne w niektórych scenariuszach zastosowania. W procesie ewaluacji nie ma konieczności podawania wszystkich V ujęć należących do zbioru przyjętego na etapie uczenia - możliwa jest klasyfikacja częściowej sekwencji. Z drugiej strony w zaproponowanym schemacie uwzględniono tylko uczenie klasyfikacyjne, to znaczy, że zadanie jest rozwiązywane dla zamkniętego zestawu klas. Nie jest oczywiste, jak przełożyć metodę na działanie na zbiorze otwartym - nie ma możliwości uzyskania deskryptora złożonego, który pozwoliłby na porównanie obiektów ze sobą. Dodatkowo konstrukcja sieci sprawia, że na wyjściu mogą powstawać wektory o znacznej długości (szczególnie dla dużej liczby klas K i wielu widoków V).

3.4.3 Trening własnych modeli agregujących do identyfikacji twarzy

Modele do klasyfikacji wieloujęciowej, takie jak wymienione w poprzednich rozdziałach: MVCNN i RotationNet zostały zaprojektowane do rozpoznawania obiektów 3D z pewnego zamkniętego słownika obiektów. Podstawowy scenariusz wykorzystania metod zakłada klasyfikację na podstawie V-elementowych zbiorów projekcji tych obiektów, gdzie obiekt każdorazowo jest reprezentowany przez wszystkie V elementów. W przypadku obydwu metod wyniki obliczane były na bazie ModelNet-40 [20], która zawiera 40 klas trójwymiarowych obiektów typu: krzesło, samolot, wanna, roślina. Natywnie w bazie obiekty są reprezentowane przez modele CAD. W artykule dotyczącym metody MVCNN wykorzystywano wersje bazy, w której każda z 40 klas składa się z 80 przykładów obiektów w części trenującej i 20 w testowej. Każdy obiekt 3D jest przekształcony do 81 obrazów 2D za pomocą narzędzia do renderingu.

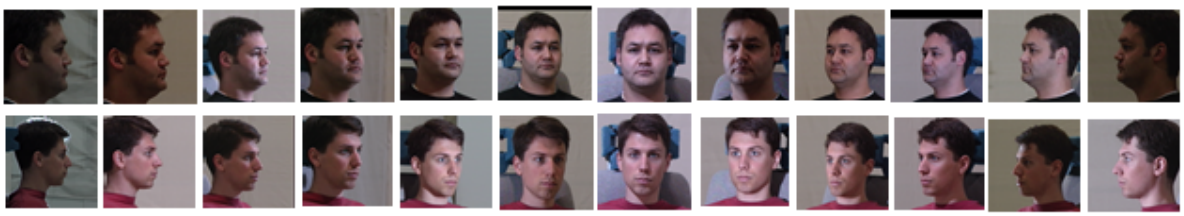
Projekcje nie mają określonego kodu widoku, ponieważ nie zawsze można w ich przypadku w jednoznaczny sposób przypisać układ współrzędnych - niektóre obiekty, takie jak butelka są symetryczne względem jednej osi obrotu. W przypadku twarzy określenie kąta ujęcia jest znacznie prostsze i zdefiniowane w ogólnie przyjętych normach takich jak norma ISO/IEC 19794-5:2005 [14].

Można zauważyć, że baza ModelNet-40 posiada stosunkowo dużo przykładów obiektów dla jednej kategorii (100), a na podstawie obiektów 3D można wytworzyć niemal dowolną liczbę widoków. Dość trudno jest zebrać bazę o podobnej charakterystyce dla twarzy, tj. posiadającą wiele zestawów składających się z N określonych ujęć dla każdej osoby. W procesie uczenia ważne jest, żeby każdy zestaw widoków był pełen - każdy zestaw powinien posiadać wszystkie V obrazów. Na tym etapie badań trudno jest oszacować wpływ braku zbalansowania danych na wyniki końcowe. W celu spełnienia powyższych warunków dla treningu identyfikacji twarzy konieczne było stworzenie nowego zestawu danych, który przypominałby swoją charakterystyką bazę ModelNet-40 dla obiektów.

W tym celu została utworzona baza MV-Multi-PIE opisana w rozdziale 1.7.2 na podstawie zbioru zdjęć twarzy CMU Multi-PIE [24]. Zbiór MV-MultiPIE wykorzystuje obrazy ze zbioru zdjęć twarzy pobranych w kontrolowanych warunkach - CMU Multi-PIE. Do zbioru MV-Multi-PIE wybrano zestawy obrazów dla 220 różnych osób, z czego 120 dostępnych jest w zbiorze trenującym (po 8 zestawów widoków) i walidacyjnym (po 3 zestawy widoków), a 100 pozostałych w zbiorze testowym. Kluczem wyboru była obecność osób na wszystkich czterech sesjach rejestracji obrazu. Z bazowego zestawu ujęć wybrano jedynie te w osi głowy - w ten sposób mamy do dyspozycji maksymalnie 13 różnych ujęć w zakresie od -90° do 90° ze skokiem co 15°. Każdy zestaw widoków MV-Multi-PIE zawiera wszystkie 15 obrazów, a etykieta ujęcia jest dostępna na etapie treningu. Dla każdego z wybranych ujęć ze zbioru źródłowego wylosowano jedno ustawienie oświetlenia na zdjęciu.

Ewaluacja modeli agregujących na zbiorach wieloujęciowych

Opisane wyżej metody klasyfikacji wieloujęciowej zostały zaproponowane do rozpoznawania określonego zbioru obiektów 3D. Każdy z modeli zawiera w swojej strukturze warstwę predykcyjną, która zwraca identyfikator klasy docelowej dla danego zestawu ujęć $\{x^{<i>}\}$. W



Rysunek 3.12: Na rysunku pokazano dwa zestawy zdjęć twarzy ze zbioru MV-Multi-PIE. W każdym z zestawów głowa osoby jest widoczna pod różnymi kątami w zakresie od $-90°$ do $90°$.

dziedzinie rozpoznawania twarzy można to przełożyć na zadanie identyfikacji osób na zbiorze zamkniętym. W literaturze dotyczącej rozpoznawania twarzy oczekiwania wobec rozwiązań zakładają możliwość zastosowania metod również na zbiorach otwartych. W systemie identyfikacji biometrycznej chcemy, żeby nie trzeba było przeliczać modelu przy każdorazowej rejestracji nowej osoby. Z tego powodu przy przeprowadzaniu badań własnych nad modelami agregującymi zaproponowaliśmy wskaźniki ewaluacji uwzględniające test na zbiorze otwartym. Dodatkowo chcieliśmy również pokazać, jak w procesie uczenia zmieniają się właściwości deskryptorów zarówno w wersji single-view (SV) oraz multi-view (MV), o ile była możliwość ekstrakcji z badanego modelu. Przy badaniu procesu uczenia modeli agregujących przyjęto następujące wskaźniki:

1. **Dokładność klasyfikacji MV na zbiorze trenującym** - jest to liczba poprawnie zwróconych identyfikatorów dla zestawów widoków przez model agregujący w stosunku do wszystkich przetworzonych zestawów widoków na danych dostępnych dla modelu w procesie uczenia.
2. **Dokładność klasyfikacji MV na zbiorze walidacyjnym** - jest to liczba poprawnie zwróconych identyfikatorów przez model agregujący w stosunku do wszystkich przetworzonych zestawów widoków, zbiór walidacyjny jest rozłączny względem zbioru trenującego, aczkolwiek musi zawierać reprezentantów tego samego zestawu klas
3. **Dokładność klasyfikacji na podstawie deskryptorów SV na zbiorze walidacyjnym** - w niektórych modelach istnieje możliwość modyfikacji przepływu danych w taki sposób, żeby warstwa predykcyjna otrzymywała na wejściu dane z pojedynczego widoku. Dokładność SV jest liczbą poprawnie zwróconych identyfikatorów przez model dla pojedynczych widoków w stosunku do wszystkich obrazów w zbiorze walidacyjnym

4. **Dokładność identyfikacji w scenariuszu open-set na zbiorze testowym w trybie SV-MV**. Scenariusz zakłada zbiór testowy, który podzielony jest na galerię i próbki-zapytania. W części galeryjnej obrazy pogrupowane są w v-elementowe zestawy widoków. Zbiór testowy jest rozłączny z trenującym pod względem obecnych w nim obrazów oraz klas do których należą. Z części galeryjnej tworzony jest rejestr złożony z deskryptorów zagregowanych. Na podstawie próbek-zapytań wytwarzane są deskryptory pojedyncze, które porównywane są z deskryptorami zagregowanymi z galerii. Wyliczana jest miara identyfikacji Rank-1 (stosunek poprawnie zwróconych podpowiedzi na 1 miejscu listy podpowiedzi w stosunku do wszystkich wzorców w galerii). Mimo, że wszystkie modele dokonują predykcji MV, nie dla wszystkich można uzyskać deskryptor zagregowany. W takim przypadku obliczany on będzie jako średnia arytmetyczna z deskryptorów pojedynczych z obrazów należących do tego samego zestawu widoków. Sposób przeprowadzenia obliczeń będzie wtedy opisany przy omawianiu wyników.

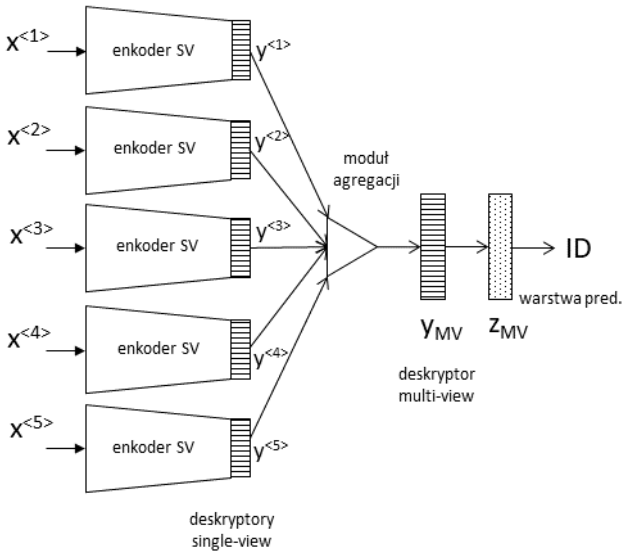
Rozdział 4

Modele własne do agregacji obrazów wielowejściowych twarzy

4.1 MVCNN-Sygnalityka

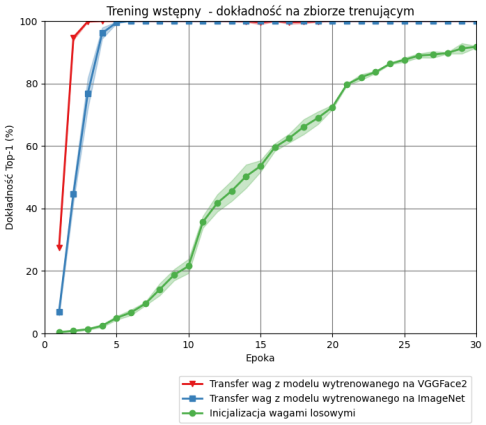
Model MVCNN-Sygnalityka wykorzystuje do agregacji koncepcje przedstawione w artykule wprowadzającym sieci MVCNN [1]. Model ma modularną strukturę i składa się z enkodera dla widoków pojedynczych (single-view), modułu agregującego oraz warstw końcowych. Wyjścia enkodera single-view są składane przez moduł agregujący. Moduł agregujący jest warstwą sieci neuronowej, która może realizować operację maksimum (*max-pooling*) lub operację uśredniania (*avg-pooling*). W części końcowej model MVCNN-Sygnalityka posiada na wyjściu warstwę predykcyjną o rozmiarze 120, który jest odpowiedni do liczby klas w bazie MV-Multi-PIE, która służy do wytrenowania sieci. W odróżnieniu od modelu MVCNN, który wykorzystywał model VGG-M, enkoder single-view w sieci MVCNN-Sygnalityka opiera się na architekturze ResNet-50 [3]. Specyfika sieci ResNet-50 implikuje pewne właściwości całego modelu. Na wejściu do sieci MVCNN-Sygnalityka podawane są obrazy o rozmiarze 224×224 z 3 kanałami barwnymi, a na wyjściu dla każdego pojedynczego widoku $x^{<i>}$ wytwarzany jest deskryptor pojedynczy $y_{<i>}$ o rozmiarze 2048. Po przetworzeniu przez moduł agregujący deskryptor zagregowany y_{MV} ma również rozmiar 2048. Enkoder single-view może zostać zainicjalizowany wagami losowymi lub można dokonać transferu wag z innej sieci o tej samej architekturze. Ze względu na budowę identyczną do zaproponowanego wcześniej modelu **bazowy-SV** do

identyfikacji twarzy na bazie VGGFace2 możemy dokonać transferu z tego modelu. Dla porównania zbadano również, jak będzie zachowywać się sieć zainicjalizowana wagami modelu wyuczonego do klasyfikacji ogólnej obrazów naturalnych z bazy ImageNet [66] udostępnionego w ramach pakietu torchvision [47]. Eksperymenty zostały przeprowadzone na zbiorze MV-Multi-PIE.

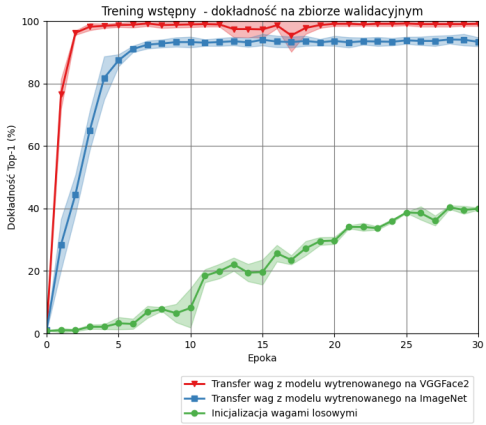


Rysunek 4.1: Schemat działania sieci agregującej MVCNN-Sygnalityka. Enkodery SV są zaimplementowane za pomocą architektury ResNet-50 i mogą zostać zainicjalizowane wagami modelu wyuczonego na bazie VGGFace2.

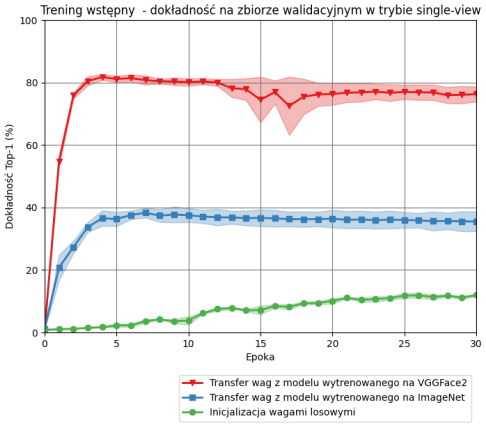
W sieci MVCNN-Sygnalityka należy przyjąć na początku wielkość zestawów danych wejściowych - ile ujęć musi się znaleźć w zbiorze na podstawie którego będzie wytworzony deskryptor $y^{<MV>}$. W przypadku korzystania z bazy MV-Multi-PIE możliwe jest wybranie trzech konfiguracji, tzn. 3, 5 i 13 widoków. Na etapie badań wstępnych, które dotyczyły sposobu inicjalizacji sieci przyjęto pięcioelementowe zestawy, które składają się ze zdjęcia frontalnego, dwóch półprofilowych i dwóch pełnych profili. W następnych badaniach będą również sprawdzane inne układy.



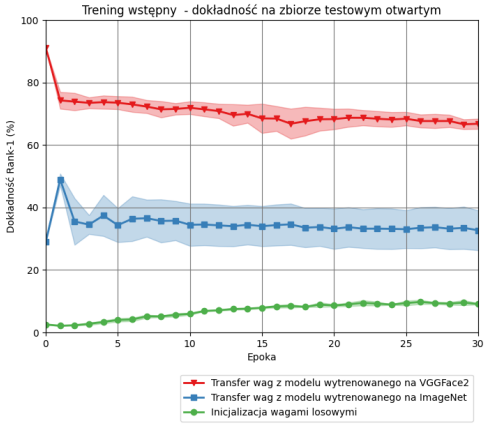
(a) Dokładność klasyfikacji MV na zbiorze trenującym



(b) Dokładność klasyfikacji MV na zbiorze walidacyjnym



(c) Dokładność klasyfikacji na podstawie deskryptorów SV na zbiorze walidacyjnym



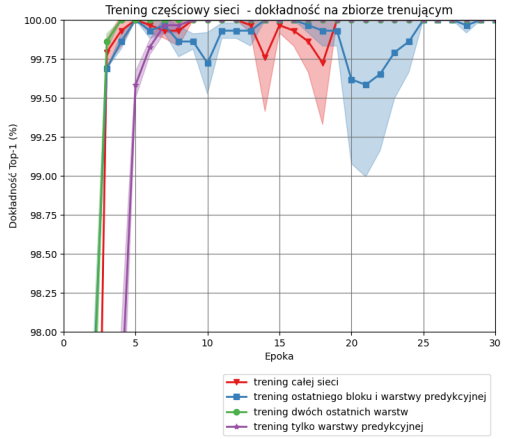
(d) Dokładność identyfikacji w scenariuszu open-set na zbiorze testowym w trybie SV-MV

Rysunek 4.2: Metoda inicjalizacji wag modelu. Trening modelu MVCNN-Sygnalityka na zestawach pięcioujęciowych twarzy ze zbiorze MV-Multi-PIE. W badaniu porównano różne sposoby inicjalizacji modelu agregującego: start z wagami losowymi, inicjalizację wagami modelu wyuczonego na zbiorze VGGFace2 (**bazowy-SV**) oraz wagami modelu wyuczonego na zbiorze ImageNet.

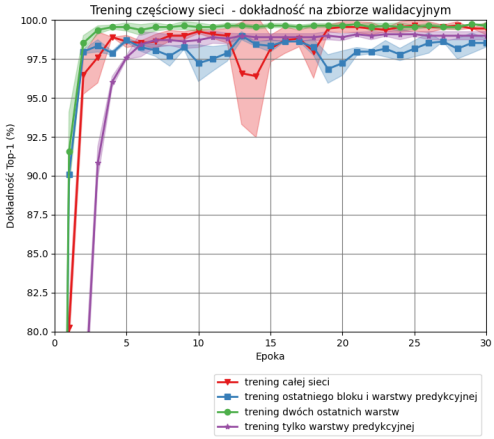
Wyniki uczenia pokazano korzystając z opisanych wcześniej scenariuszy ewaluacji - obserwowana była dokładność na zbiorze trenującym, dokładność dla deskryptorów zagregowanych na zbiorze walidacyjnym (MV), dokładność dla deskryptorów pojedynczych na zbiorze walidacyjnym (SV) oraz współczynnik Rank-1 na zbiorze testowym otwartym (inne klasy niż w zbiorze walidacyjnym). Wyniki treningu sieci dla różnych punktów startowych zostały przedstawione na wykresach 4.2. Prezentowane wyniki dotyczą 3 powtórzeń tego samego eksperymentu i są to uśrednione wartości (cienie na rysunkach pokazują odchylenie standardowe wyników). Sieć MVCNN-Sygnalityka uczona jest z zastosowaniem funkcji SGD przy przyjętej prędkości uczenia (*ang. learning rate*) która wynosi 1e-4 oraz dla korekty wagi (*ang. weight decay*) równej 1e-5. Sposób inicjalizacji sieci ma bardzo duży wpływ na wyniki identyfikacji. Można uznać, że trening wstępny enkodera single-view jest niezbędny do poprawnego działania sieci agregującej. Najniższą dokładność uzyskano dla modelu inicjalizowanego wagami losowymi, kolejny był model inicjalizowany wagami wytrenowany na bazie ImageNet, a najlepsze dla modelu inicjalizowanego wagami modelu **bazowy-SV**. Przy inicjalizacji wagami modelu **bazowy-SV** sieć uzyskuje wysokie wyniki klasyfikacji na zbiorze zamkniętym, ale widać spadek współczynnika Rank-1 na zbiorze otwartym. Z tego powodu kolejne eksperymenty będą dotyczył częściowego zamrażania wag w sieci w trakcie jej uczenia. Dzięki temu wymusimy na sieci zachowanie właściwości modelu single-view, który był trenowany na znacząco większym zbiorze danych (VGGFace2).

Trening częściowy sieci

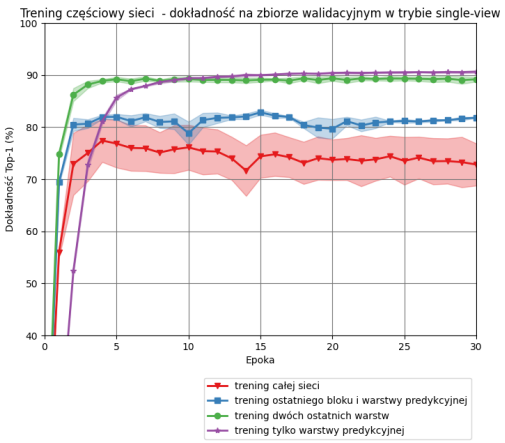
Trening z wyłączeniem części sieci jest szeroko stosowaną procedurą przy transferze wag z innego modelu, szczególnie jeżeli model źródłowy trenowany był na bardziej reprezentatywnym zbiorze danych. Zazwyczaj zamraża się pewną liczbę początkowych warstw sieci. Przez zamrażanie rozumiemy wyłączenie tych warstw z procesu uczenia albo przez brak aktualizacji wag w tych warstwach albo przez obniżenie współczynnika *learning-rate* dla tych warstw. W naszym przypadku enkoder single-view wbudowany w model MVCNN-Sygnalityka z 49 warstw (warstwy sieci ResNet-50 po odcięciu warstwy predykcyjnej).



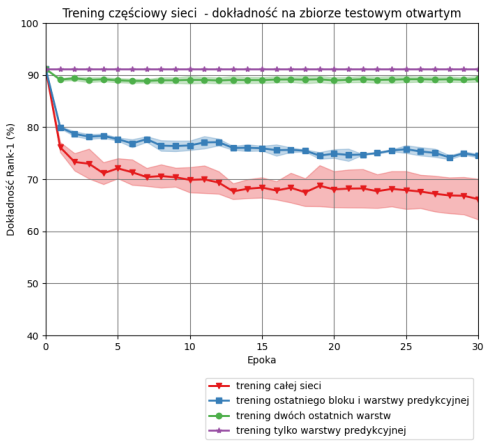
(a) Dokładność klasyfikacji MV na zbiorze trenującym



(b) Dokładność klasyfikacji MV na zbiorze walidacyjnym



(c) Dokładność klasyfikacji na podstawie deskryptorów SV na zbiorze walidacyjnym



(d) Dokładność identyfikacji w scenariuszu open-set na zbiorze testowym w trybie SV-MV

Rysunek 4.3: Częściowe zamrażanie wag sieci w procesie uczenia. Trening modelu MVCNN-Sygnalityka na zestawach pięcioujęciowych twarzy ze zbiorze MV-Multi-PIE. W badaniu sprawdzono, jak przebiega proces uczenia przy częściowym zamrażania wag sieci.

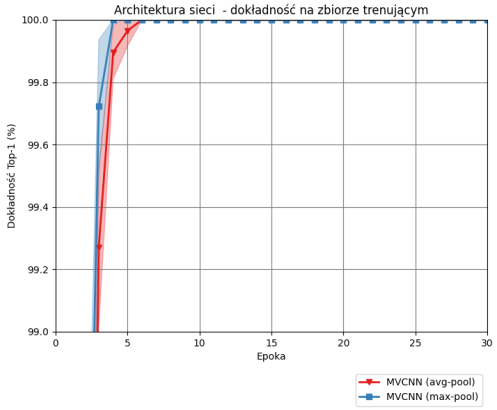
W eksperymentach omówionych w poprzedniej sekcji uczone były wszystkie 49 warstwy enkodera single-view oraz warstwa predykcyjna modułu multi-view. W kolejnych testach rozważać będziemy:

- trening ostatniego bloku i warstwy predykcyjnej (dla przyjętego modelu ResNet-50 oznacza to zamrożenie pierwszych 40 warstw)
- trening ostatniej warstwy splotowej i warstwy predykcyjnej (dla przyjętego modelu ResNet-50 oznacza to zamrożenie pierwszych 48 warstw)
- trening warstwy predykcyjnej (oznacza to zamrożenie całego enkodera SV z przetransferowanymi wagami)
- trening całej sieci

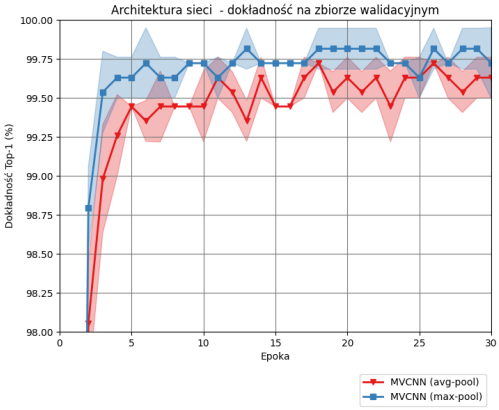
Przebieg uczenia przy częściowym zamrażaniu wag w sieci pokazano na wykresach 4.3. Najwyższą dokładność klasyfikacji MV na zbiorze walidacyjnym uzyskano dla sieci w której uczona była tylko ostatnia warstwa splotowa i warstwa predykcyjna. Najgorsze wyniki uzyskano w przypadku trenowania wszystkich wag modelu. Na zbiorze open-set jedynie przy zamrożeniu całego enkodera współczynnik Rank-1 nie uległ zmianie. Kolejne eksperymenty przeprowadzono w wersji, w której uczona jest tylko ostatnia warstwa splotowa i warstwa predykcyjna.

4.1.1 Moduł agregacji

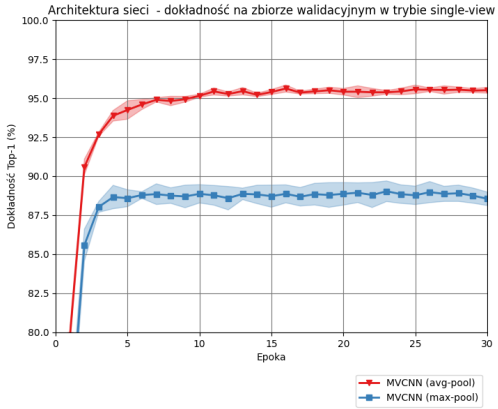
W opisie modelu na początku zostało wspomniane, że model MVCNN-Sygnalityka składa się z modułu agregującego, który zbudowany jest z warstwy typu *average-pooling*, czyli z warstwy wykonującej operację uśredniania względem kolejnych pozycji deskryptora. Architektura modelu różni się od metody źródłowej, czyli modelu MVCNN, który wykorzystywał warstwę *max-pooling*. W badaniu, którego wyniki pokazane są na wykresach 4.4, zestawiono ze sobą oba podejścia do agregacji. Należy pamiętać, że metoda źródłowa była tylko badana pod względem dokładności klasyfikacji na zbiorze zamkniętym, a więc pokrywa się z wynikami na grafice 4.4b. Biorąc pod uwagę jedynie klasyfikację MV metoda *max-pooling* rzeczywiście sprawdza się lepiej. My jednak wykorzystujemy w systemie nie tylko identyfikator klasy, ale też porównujemy ze sobą deskryptory. W takim zastosowaniu warstwa *average-pooling* jest właściwszym wyborem.



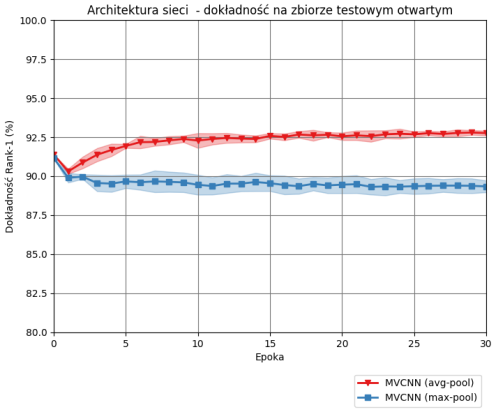
(a) Dokładność klasyfikacji MV na zbiorze trenującym



(b) Dokładność klasyfikacji MV na zbiorze walidacyjnym



(c) Dokładność klasyfikacji na podstawie deskryptorów SV na zbiorze walidacyjnym



(d) Dokładność identyfikacji w scenariuszu open-set na zbiorze testowym w trybie SV-MV.

Rysunek 4.4: Trening modelu MVCNN dla dwóch przyjętych architektur sieci na bazie MV-Multi-PIE. W badaniu porównano model sieci z warstwą uśredniającą *average-pooling* i maksimum *max-pooling*

Przy okazji warto zwrócić uwagę, że budowa modułu agregującego ma potencjał na dalszy rozwój. Punkt scalenia danych z poszczególnych enkoderów single-view może być realizowany innymi warstwami i strukturami niż warstwy *average-pooling* i *max-pooling*. Pomysł ten zostanie rozwinięty w rozdziale dotyczącym sieci SygnaT, który to model można uznać do pewnego stopnia za modyfikację sieci MVCNN-Sygnalityka. W sieci SygnaT został zaproponowany moduł agregujący wykorzystujący mechanizm atencji.

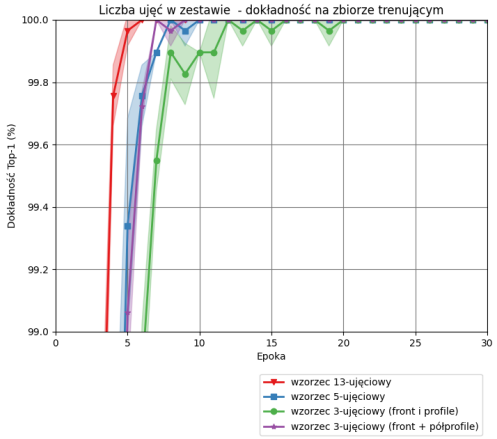
Konfiguracja zbiorów ujęć

Jak dotąd wszystkie badania przeprowadzono na zestawach widoków o arbitralnie przyjętej wielkości 5. Dyskusja o tym, jakiej wielkości powinny być zestawy widoków zależy przede wszystkim od danych na których operujemy i tego ile widoków jest możliwych do użycia w systemie. Decyzja o konstrukcji zestawów może też zależeć od uwarunkowań prawnych, ograniczeń technicznych i preferencji osób, które będą korzystać docelowo z systemu. Mimo to warto sprawdzić, jak będzie zachowywać się sieć dla różnych liczności zestawów .

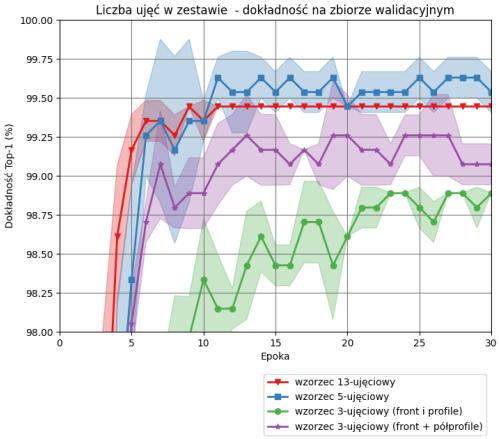
W opisanym poniżej badaniu przyjęto następujące konfiguracje:

- 13 ujęć - wszystkie ujęcia ze zbioru MV-MultiPIE, a więc: frontalne, $\pm 15°$, $\pm 30°$, $\pm 45°$, $\pm 60°$, $\pm 75°$, $\pm 90°$
- 5 ujęć - wybrane ujęcia ze zbioru MV-MultiPIE: frontalne, półprofile ($\pm 45°$) i pełne profile ($\pm 90°$)
- 3 ujęcia - wybrane ujęcia ze zbioru MV-MultiPIE frontalne i pełne profile ($\pm 90°$)
- 3 ujęcia - wybrane ujęcia ze zbioru MV-MultiPIE frontalne i półprofile ($\pm 45°$)

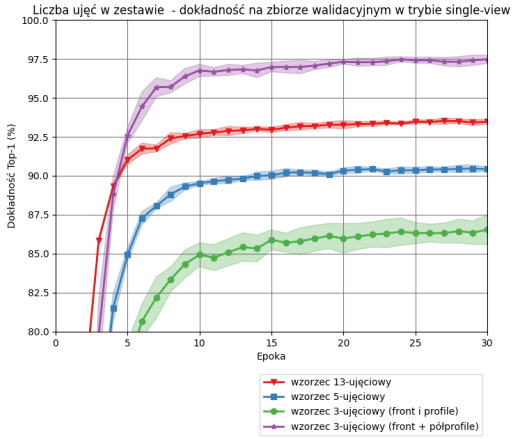
Wyniki treningu dla przyjętych konfiguracji można zobaczyć na wykresach 4.5. Wyraźnie lepsze wyniki uzyskano dla większych zestawów, tj. dla zestawów o rozmiarze 5 i 13. Warto jednak zwrócić uwagę, że przypadku tego eksperymentu wyniki nie są porównywalne. Walidacja wykonywana jest na innym zbiorze danych, a dokładnie na podzbiorze zawierającym wybrane ujęcia. Ma to miejsce również w przypadku testu otwartego - galeria dla liczby kątów mniejszej od 13 jest złożona z odpowiednio mniejszej liczby obrazów. Najgorsze wyniki dotyczą zestawu trzyujęciowego, co wiąże się ze wspomnianym ograniczeniem zbioru do budowania galerii, ale może też wynikać z przewagi zdjęć profilowych do bliższych portretowi frontalnemu w



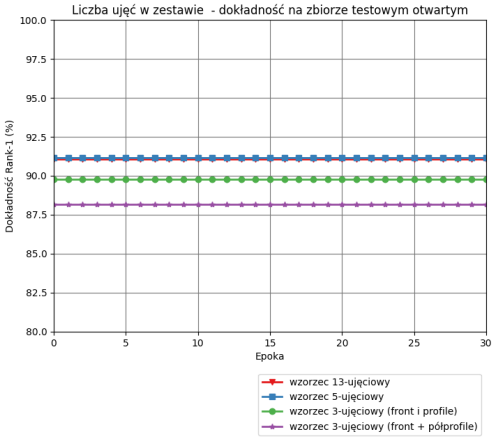
(a) Dokładność klasyfikacji MV na zbiorze trenującym



(b) Dokładność klasyfikacji MV na zbiorze walidacyjnym



(c) Dokładność klasyfikacji na podstawie deskryptorów SV na zbiorze walidacyjnym



(d) Dokładność identyfikacji w scenariuszu open-set na zbiorze testowym w trybie SV-MV.

Rysunek 4.5: Trening modelu MVCNN dla różnych konfiguracji zbiorów uwzględnianych przy tworzeniu wzorca na bazie MV-Multi-PIE. W porównaniu zbadano trzy-, pięcio- i trzynastoujęciowe zbiory. Dodatkowo dla zestawu trzyujęciowego zastąpiono profile ($\pm 90°$) półprofilami ($\pm 45°$).

zestawie, które jak wykazały wcześniejsze wyniki, utrudniają identyfikację twarzy.

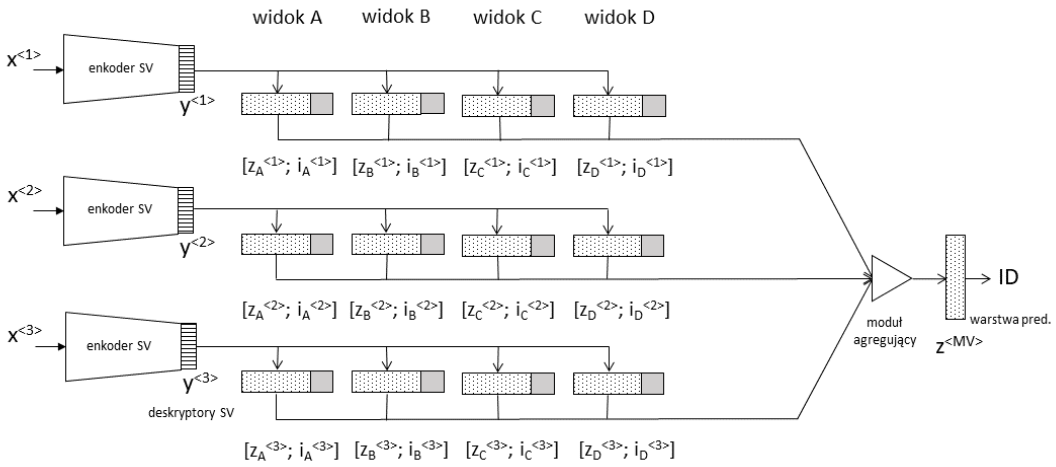
4.1.2 Wnioski

Sieć MVCNN-Sygnalityka ma stosunkowo prostą do implementacji i zrozumienia strukturę. Większość wag w modelu może i co więcej powinna być przetransferowana z modelu single-view wyuczonego na pojedynczych ujęciach. Konieczność wyboru liczności zbioru ujęć stanowi pewne ograniczenie przy wykorzystaniu możliwości zbiorów trenujących i walidacyjnych.

4.2 RotationNet-Sygnalityka

Model RotationNet-Sygnalityka wykorzystuje architekturę RotationNet, która została zaproponowana w artykule [2], ale dostosowuje jej założenia do specyfiki danych wieloujęciowych twarzy. W sieci typu RotationNet możemy wyróżnić trzech części. Pierwsza to enkoder *single-view*, który jest zaimplementowany w taki sam sposób jak ma to miejsce w modelu MVCNN-Sygnalityka. Enkoder *single-view* jest siecią ResNet-50 i może zostać zainicjalizowany wagami modelu **bazowego-SV**. Enkoder na wejściu przyjmuje obrazy o rozmiarze 224×224 z 3 kanałami barwnymi, a na wyjściu wytwarza deskryptor pojedynczy $y^{<SV>}$ o rozmiarze 2048. Druga część modelu to zestaw warstw predykcyjnych. Każda z warstw koduje identyfikator klasy oraz dodatkowo prawdopodobieństwo jednego z kątów widoku próbki. Kąty widoków należą do przyjętego V-elementowego słownika widoków. Przykładowo dla słownika złożonego z widoków: lewy profil, lewy półprofil, zdjęcie frontalne, prawy półprofil, prawy profil, wielkość V wynosi 5. Tak więc po przetworzeniu każdej z próbek na wyjściu otrzymujemy V warstw predykcyjnych o rozmiarze $N+1$, gdzie N jest liczbą klas, która w przypadku bazy MV-Multi-PIE wynosi 120. Na końcu modelu RotationNet-Sygnalityka predykcje pojedynczych widoków są agregowane i zapisywane w ostatniej warstwie, która koduje wynik dla całego zestawu próbek wejściowych. Wielkość warstwy końcowej jest równa N, tak jak ma to miejsce w typowym modelu do klasyfikacji. Funkcja, która służy do złożenia predykcji cząstkowych została przedstawiona przy okazji omawiania metody RotationNet 3.1. Jako że wyniki modeli cząstkowych są składane na poziomie predykcji w sieci RotationNet-Sygnalityka nie ma

możliwości uzyskania deskryptora złożonego $y^{<MV>}$, który mógłby zostać użyty w systemie otwartym. W części eksperymentów, która tego wymaga stosowana jest średnia arytmetyczna z deskryptorów $y^{<SV>}$ dla poszczególnych widoków.



Rysunek 4.6: Schemat działania sieci agregującej RotationNet-Sygnalityka. Enkodery SV są zaimplementowane za pomocą sieci ResNet-50. Liczba widoków jest przyjęta arbitralnie.

Kluczową dla obliczeń częścią metody RotationNet jest generacja widoków kandydujących. Jest to list możliwych widoków, które są dopuszczone w zestawie. W metodzie źródłowej wiele założeń wynikało z właściwości bazy ModelNet-40 wytworzonej ze skanów obiektów. W bazie ModelNet-40 zazwyczaj trudno wyznaczyć jednoznaczny punkt odniesienia, ale można założyć pewną kolejność w sekwencji. W praktyce może to być użyteczne, np. w przypadku skanowanie obiektów na stoliku obrotowym - nie jest wtedy ważne w jaki sposób położy się przedmiot na początku skanowania, ale wiadomo, że kąt zmienia się o pewną przyjętą liczbę stopni. Nie jest to właściwe i praktyczne dla zestawów złożonych ze zdjęć twarzy, dlatego w proponowanej metodzie przyjęto dwa warianty: w pierwszym kolejność obrazów jest znana z góry, a w drugim sprawdzane są wszystkie możliwe kombinacje. Drugi wariant jest możliwy tylko dla zestawów próbek o wielkości mniejszej niż 5. Dla zestawów o wielkości 13 widoków liczba kombinacji przekracza 6 miliardów. Dla zestawów o wielkości 5 sprawdzane jest 120 kombinacji, a dla zestawów o wielkości 3 sprawdzane jest 6 kombinacji.

Środowisko testowe w procesie uczenia sieci RotationNet-Sygnalityka miało identyczne ustawienia, jak w przypadku sieci MVCNN-Sygnalityka. Wykorzystano bazę MV-Multi-PIE, a poszczególne eksperymenty były powtarzane trzykrotnie w celu uśrednienia wyników. Badania

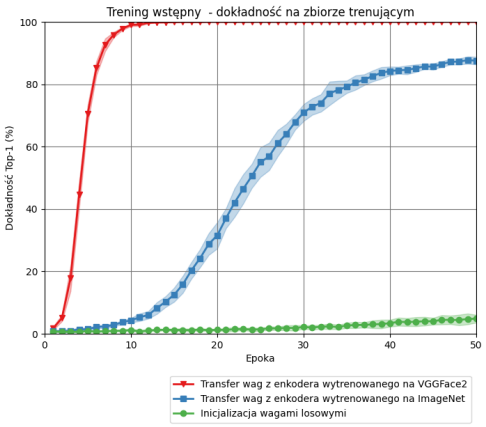
rozpoczęły się od przyjrzeniu się bliżej wpływowi inicjalizacji na przebieg procesu uczenia. Sprawdzone zostały trzy opcje startowania treningu: inicjalizację wagami losowymi, wagami z modelu wyuczonego na bazie ImageNet [66] oraz wagami modelu **bazowego-SV** wyuczonego na bazie VGGFace2.

Dla modelu RotationNet-Sygnalityka obserwacja jest podobna, co dla sieci MVNN-Sygnalityka, że pretrening modelu na obrazach twarzy w trybie single-view jest niezbędny do uzyskania wysokich wyników identyfikacji. Bez wcześniejszego wyuczenia enkodera, trening modelu jest praktycznie niemożliwy, co można zaobserwować na wykresach 4.7.

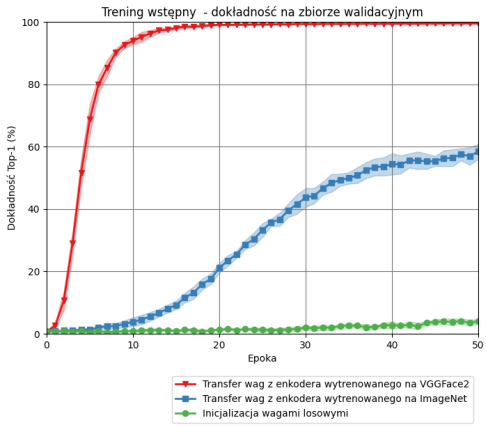
Trening częściowy sieci

W kolejnym kroku przeprowadzono test treningu sieci RotationNet-Sygnalityka z częściowym zamrażaniem wag modelu. Model był zainicjalizowany wagami z modelu wyuczonego na bazie VGGFace2 (**bazowy-SV**). W teście porównano model trenowany w całości z: modelem trenowanym od ostatniego bloku, modelem trenowanym od ostatniej warstwy splotowej oraz modelem, w którym trenowana była tylko warstwa predykcyjna. Sieć RotationNet-Sygnalityka potrzebuje więcej epok do uzyskania zbieżności niż sieć MVCNN-Sygnalityka przy takich samych ustawieniach i wielkości porcji danych wsadowych. W teście, podobnie jak w przypadku sieci MVCNN-Sygnalityka, sprawdzono cztery wersje treningu. Było to uczenie całej sieci, uczenie jedynie warstwy predykcyjnej, uczenie dwóch ostatnich warstw sieci oraz uczenie jedynie ostatniego bloku sieci wraz z warstwą predykcyjną sieci ResNet-50.

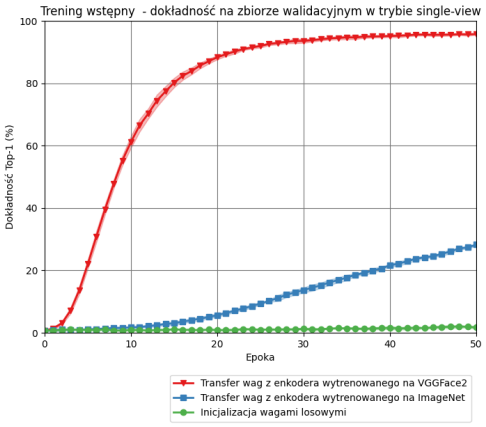
Dla wszystkich ustawień wartości dokładności identyfikacji MV na zbiorze walidacyjnym były porównywalne i wobec zakresu zmienności w ramach eksperymentu trudno wyróżnić najlepsze rozwiązanie. Warto jednak zwrócić uwagę na wyniki w trybie open-set, które w przypadku treningu ostatniej warstwy poprawiły się względem punktu początkowego.



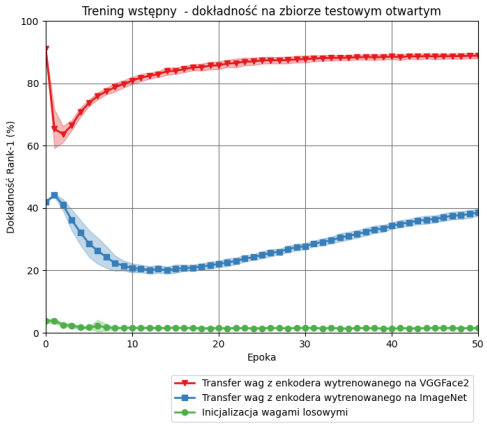
(a) Dokładność kasyfikacji MV na zbiorze trenującym



(b) Dokładność klasyfikacji MV na zbiorze walidacyjnym

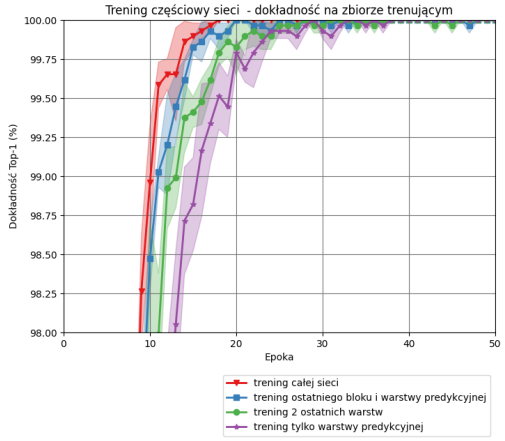


(c) Dokładność klasyfikacji na podstawie deskryptorów SV na zbiorze walidacyjnym

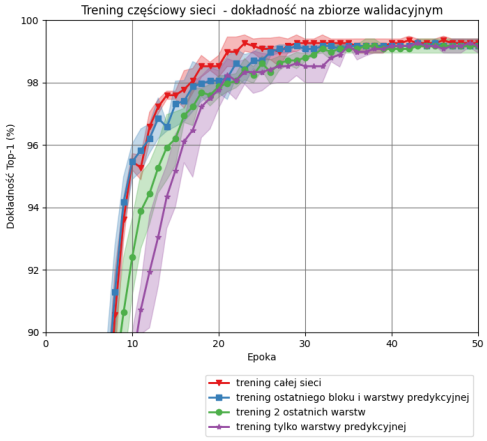


(d) Dokładność identyfikacji w scenariuszu open-set na zbiorze testowym w trybie SV-MV. Deskryptory MV są średnią arytmetyczną deskryptorów SV dla widoków z jednego zestawu.

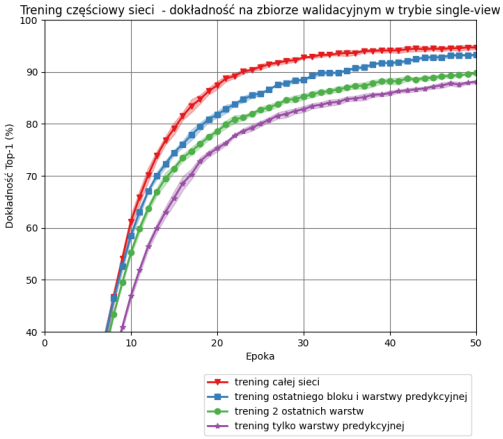
Rysunek 4.7: Trening modelu RotationNet-Sygnalityka na zestawach pięcioujęciowych ze zbiorze MV-Multi-PIE. W badaniu porównano różne sposoby inicjalizacji modelu agregującego. Enkoder SV był inicjalizowany wagami modelu wyuczonego na zbiorze VGGFace2 (**bazowy-SV**), ImageNet oraz inicjalizowany losowo.



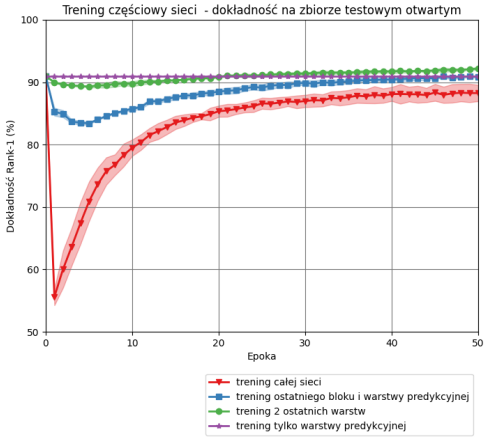
(a) Dokładność klasyfikacji MV na zbiorze trenującym



(b) Dokładność klasyfikacji MV na zbiorze walidacyjnym



(c) Dokładność klasyfikacji na podstawie deskryptorów SV na zbiorze walidacyjnym



(d) Dokładność identyfikacji w scenariuszu open-set na zbiorze testowym w trybie SV-MV. Deskryptory MV są średnią arytmetyczną deskryptorów SV dla widoków z jednego zestawu.

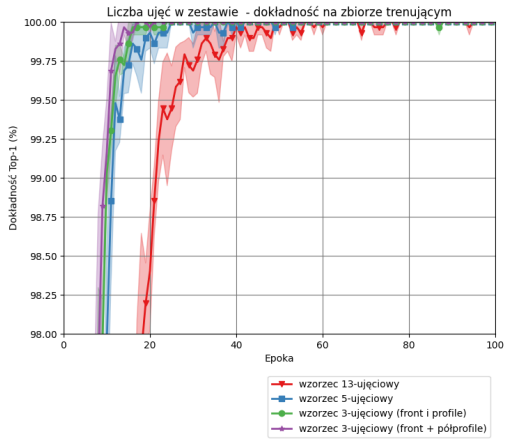
Rysunek 4.8: Trening modelu RotationNet-Sygnalityka na zestawach pięcioujęciowych twarzy ze zbioru MV-Multi-PIE. W badaniu sprawdzono, jak przebiega proces uczenia przy częściowym zamrażaniu wag sieci.

Konfiguracja zbiorów ujęć

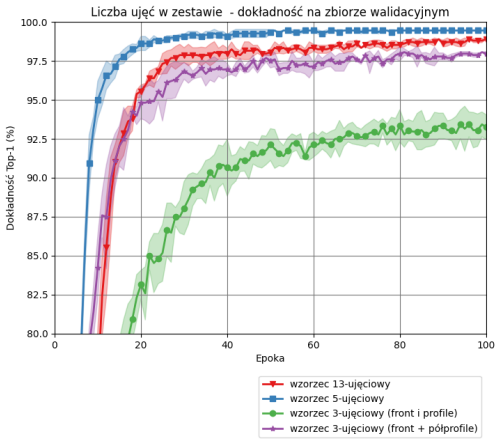
Tak jak w przypadku eksperymentów z siecią MVCNN-Sygnalityka przyjęto, że model do identyfikacji RotationNet-Sygnalityka będzie budowany na podstawie zestawów ujęć o różnej liczności. Uwzględnione zostały następujące liczności zbiorów widoków:

- 13 ujęć - wszystkie ujęcia ze zbioru MV-MultiPIE, a więc: frontalne, $\pm 15^\circ$, $\pm 30^\circ$, $\pm 45^\circ$, $\pm 60^\circ$, $\pm 75^\circ$, $\pm 90^\circ$
- 5 ujęć - wybrane ujęcia ze zbioru MV-MultiPIE: frontalne, półprofile ($\pm 45^\circ$) i pełne profile ($\pm 90^\circ$)
- 3 ujęcia - wybrane ujęcia ze zbioru MV-MultiPIE frontalne i pełne profile ($\pm 90^\circ$)
- 3 ujęcia - wybrane ujęcia ze zbioru MV-MultiPIE frontalne i półprofile ($\pm 45^\circ$)

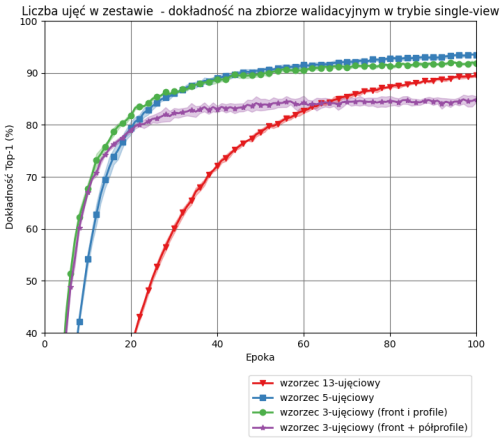
W przypadku metody RotationNet-Sygnalityka kluczowe w procesie uczenia jest generowanie widoków kandydujących. Znalezienie kodu ujęcia dla każdej próbki wejściowej jest częścią klasyfikacji zarówno na etapie uczenia sieci, jak i ewaluacji. Identyfikator zbioru wejściowego jest określany na podstawie wybranej sekwencji ujęć. Wybór sekwencji jest wynika z z minimalizacji wartości neuronów kodujących ujęcie. Sieć RotationNet-Sygnalityka estymuje jakie widoki i w jakiej kolejności otrzymała na wejściu bez potrzeby podawania tej informacji bezpośrednio. W przypadku małego zestawu widoków, możliwe jest sprawdzenie wszystkich kombinacji: dla trzech zdjęć w zestawie jest to sześciu kombinacji, a dla pięciu zdjęć w zestawie dziesięć kombinacji. Zestawy o liczności 13 przekraczają już dostępne możliwości obliczeniowe. Z tego też powodu dla zestawów trzynastoujęciowych przyjęto stałą kolejność ujęć w zestawie wejściowym. Wersja nie wykorzystuje więc pełnych możliwości modelu, a eksperyment był przeprowadzony dla spójności z badaniami nad siecią MVCNN-Sygnalityka. Najlepsze wyniki dokładności identyfikacji otrzymano dla modeli pięcioujęciowych. Najtrudniej było wyuczyć model trzyujęciowy, w którym obecne były pełne profile twarzy. Ciekawe wyniki można zaobserwować dla zbioru testowego. Tutaj udało się osiągnąć wyniki wyższe niż dla bazowego klasyfikatora SV, to znaczy, że sieć lepiej nauczyła się generalizować proces agregacji dla zdjęć twarzy. Warto jednak zauważyć, że w przypadku ekstrakcji deskryptorów zagregowanych MV, które są potrzebne do obliczeń w scenariuszu testowym, były to uśrednione deskryptory proste SV, ponieważ model RotationNet nie zakłada możliwości generacji tego typu informacji.



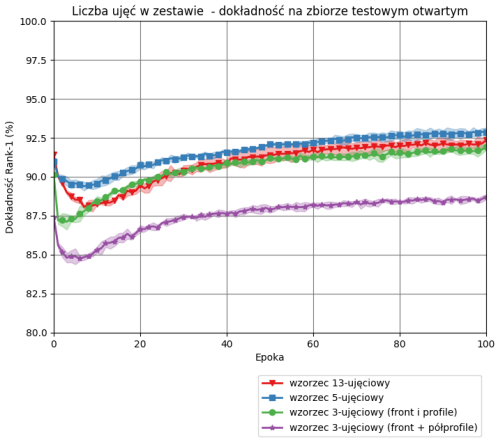
(a) Dokładność klasyfikacji MV na zbiorze trenującym



(b) Dokładność klasyfikacji MV na zbiorze walidacyjnym



(c) Dokładność klasyfikacji na podstawie deskryptorów SV na zbiorze walidacyjnym



(d) Dokładność identyfikacji w scenariuszu open-set na zbiorze testowym w trybie SV-MV. Deskryptory MV są średnią arytmetyczną deskryptorów SV dla widoków z jednego zestawu.

Rysunek 4.9: Trening modelu RotationNet dla różnych konfiguracji zbiorów uwzględnianych przy tworzeniu wzorca na bazie MV-Multi-PIE. W porównaniu zbadano trzy-, pięcio- i trzynastoujęciowe zbiory. Dodatkowo dla zestawu trzyujęciowego zastąpiono profile ($\pm 90°$) półprofilami ($\pm 45°$).

Przypisywanie kodów ujęć do próbek jest nierozłączną częścią procesu identyfikacji w modelu RotationNet-Sygnalityka oraz stanowi jej zaletę, która może być wykorzystana w systemie. W kolejnym kroku przeprowadzony został eksperyment, czy estymacja ujęć wykonywana przez model RotationNet-Sygnalityka jest wiarygodna. W tym celu policzona została macierz pomyłek dla uczenia nienadzorowanego, jakie następuje wewnątrz modelu. Jako że jest to uczenie nienadzorowane nie ma jednoznacznego przypisania widoków zdefiniowanych w sieci (oznaczonych jako widok A, B, C, D, E w tabeli 4.1) do widoków rzeczywistych, czyli ujęcia frontalnego, lewego i prawego półprofilu oraz lewego i prawego pełnego profilu. Tabela 4.1 pokazuje, jaka część widoków źródłowych została przypisana do danej reprezentacji wewnętrznej widoku. Można zobaczyć, że mimo braku mapowania, ujęcia są przypisywane do wybranych widoków z ponad 90% dokładnością. Najniższą dokładność zanotowano dla wyjść z sieci przypisanych do obrazów frontalnych (kolumna - Widok A), ale były one mylone tylko z obrazami półprofilowymi, a nie profilowymi. Ogólnie można zauważyć, że pomyłki zdarzają się częściej dla ujęć bliższych pod względem różnicy w kątach, a rzadziej między tymi o dużej różnicy kątowej.

Obraz źródłowy	Widok A	Widok B	Widok C	Widok D	Widok E
lewy profil ($-90°$)	0.00%	98.00%	0.00%	2.00%	0.00%
lewy półprofil ($-45°$)	2.67%	1.00%	0.33%	96.00%	0.00%
frontalny	91.67%	0.00%	0.33%	1.00%	7.00%
prawy półprofil ($+45°$)	1.33%	0.00%	1.67%	0.00%	97.00%
prawy profil ($+90°$)	0.00%	1.33%	94.67%	0.00%	4.00%

Tabela 4.1: Macierz pomyłek estymacji kodów widoków w sieci RotationNet-Sygnalityka dla modelu pięcioujęciowego. Liczba w tabeli jest stosunkiem liczby obrazów przypisanych do danego kodu widoku do liczby wszystkich obrazów o danym ujęciu.

4.2.1 Porównanie MVCNN-Sygnalityka i RotationNet-Sygnalityka

Model MVCNN-Sygnalityka jest prosty do implementacji i z tego też powodu jest często stosowany w innych bardziej kompleksowych rozwiązaniach dotyczących klasyfikacji wielouję-

ciowej. Ma budowę modułową, a duża część wag w układzie może zostać przetransferowana z modelu single-view o zbliżonej architekturze. Model może więc zostać wstępnie wytrenowany na bazie o większym wolumenie, w której nie ma podziału na podzbiory wieloujęciowe. Baza do trenowania sieci MVCNN-Sygnalityka na etapie treningu i ewaluacji wymaga podzbiorów o stałej przyjętej liczności. Kolejność obrazów nie ma znaczenia. Ze struktury modelu wynika, że stosunkowo łatwo można uzyskać z niego deskryptor zagregowany, deskryptory pojedyncze dla poszczególnych widoków, identyfikator klasy dla wzorca zagregowanego, a także identyfikator klasy dla wzorca pojedynczego. Poszczególne moduły i wyjścia z sieci są łatwe do opisu, a więc i interpretacji. Model MVCNN-Sygnalityka ma potencjał na rozwój poprzez modyfikację modułu agregującego z użyciem innych funkcji niż średnia i maksimum.

Model RotationNet-Sygnalityka jest bardziej skomplikowany niż model MVCNN-Sygnalityka. Przy dużej liczbie klas i ujęć ostatnia warstwa przyjmuje znaczne rozmiary, posiada również nadmiarowe wektory predykcji, które mogą być pomijane na etapie składania wyniku końcowego. Funkcja agregująca nie jest prosta w interpretacji i nie pozwala na ekstrakcji deskryptorów. Uczenie i ewaluacja modelu RotationNet-Sygnalityka wymaga dwuetapowej optymalizacji, która najpierw przypisuje kody widoków do poszczególnych obrazów, a potem na ich podstawie składa predykcję końcową. Mechanizm przypisania widoków może zależeć od założeń przyjętych na początku treningu, np. od listy dopuszczonych widoków kandydujących. Sieć uczy się dłużej niż MVCNN-Sygnalityka i na etapie ewaluacji dodaje pewien narzut na złożenie predykcji. Zaletą jest przypisywanie widoków do ujęć, które jest dodatkową informacją, która może być przydatna w systemie. Uwzględnienie widoku wydaje się też być zgodne z intuicją dla lepszej pracy systemu niż traktowanie obrazów na wejściu bez kontekstu jako zbiór danych. Metoda RotationNet-Sygnalityka została zaprojektowana do klasyfikacji na zbiorze zamkniętym, co sprawia, że bez modyfikacji trudno o zastosowanie w systemach rzeczywistych identyfikacji twarzy.

W celu dokładniejszego porównania wyników działania obydwu modeli przeprowadzony został dodatkowy test na bazie MV-MultiPIE. Test był pełniejszy niż wyniki przedstawione na wcześniejszych wykresach pokazujących przebieg treningu, ponieważ była w nim stosowana walidacja skrośna. Test dotyczył identyfikacji na zbiorze otwartym. Wyniki zostały obliczone dla dwóch wielkości wzorców złożonych: wzorców trzyjęciowych 4.2 i wzorców pięcioujęciowych 4.3.

Metoda	Rank-1	Rank-1 (tylko frontalne)	Rank-1 (tylko półprofile)	Rank-1 (tylko pełne profile)
Model bazowy single-view (avg)	90.94%	93.59%	90.84%	79.85%
MVCNN-Sygnalityka (max-pooling)	89.19%	95.79%	91.21%	74.18%
MVCNN-Sygnalityka (avg-pooling)	92.41%	94.69%	92.86%	83.88%
RotationNet-Sygnalityka	**95.68%**	**96.70%**	**95.79%**	**91.76%**

Tabela 4.2: Porównanie wyników identyfikacji z użyciem modeli agregujących na bazie MV-Multi-PIE z walidacją skrośną. W scenariuszu testowym przyjęto **trzyujęciowe** zestawy danych wejściowych przy budowaniu galerii.

Metoda	Rank-1	Rank-1 (tylko frontalne)	Rank-1 (tylko półprofile)	Rank-1 (tylko pełne profile)
Model bazowy single-view (avg)	92.18%	95.05%	92.85%	80.95%
MVCNN-Sygnalityka (max-pooling)	90.19%	95.97%	92.86%	74.18%
MVCNN-Sygnalityka (avg-pooling)	96.54%	98.17%	96.89%	89.56%
RotationNet-Sygnalityka	**96.63%**	**97.80%**	**97.07%**	**89.93%**

Tabela 4.3: Porównanie wyników identyfikacji z użyciem modeli agregujących na bazie MV-Multi-PIE z walidacją skrośną. W scenariuszu testowym przyjęto **pięcioujęciowe** zestawy danych wejściowych przy budowaniu galerii.

Wyniki dokładności identyfikacji Rank-1 zostały obliczone dla całego zbioru próbek-zapytań oraz dla wybranych ujęć pochodzących z tego samego zbioru. Wyodrębniono zdjęcia frontalne, zdjęcia półprofilowe i pełne profilowe. Po przeprowadzeniu pełnego testu można zobaczyć, że metoda RotationNet-Sygnalityka uzyskała najlepsze wyniki dla obydwu wariantów liczności ujęć we wzorcu. Metoda MVCNN-Sygnalityka była lepsza niż uśredniony wzorzec single-view, ale tylko po zastosowaniu modułu agregującego *avg-pooling*. Wersja z modułem max-pooling była gorsza od enkodera *single-view*. Obserwując wyniki widać, że metoda RotationNet-Sygnalityka uzyskała bardzo dobre wyniki dla obrazów profilowych. Jest to wynik o tyle ciekawy, że uwzględnienie widoku źródłowego albo wzajemnej relacji między obrazami należącymi do wieloujęciowego zestawu wydaje się właściwym kierunkiem rozwoju.

4.3 SygnaT (Sygnalityka-Transformer)

Z analizy modeli MVCNN-Sygnalityka oraz RotationNet-Sygnalityka wynikało, że oba te rozwiązania można z powodzeniem zastosować do zamkniętego zbioru klas obrazów ludzkich twarzy ujętych pod różnymi kątami. Wyniki dokładności identyfikacji w trybie multi-view dla przyjętego zestawu klas były stosunkowo wysokie. Żadna z sieci nie dała jednak wystarczająco zadowalających wyników w scenariuszu testowym otwartym, w którym wyszukiwano obrazów single-view w bazie referencyjnej multi-view, w którym pojawiały się obiekty nieznane wcześniej dla modelu. Różnice w dokładnościach między testem otwartym, a zamkniętym może mieć wiele przyczyn. Jednym z powodów może być zastosowanie uczenia klasyfikacyjnego, w którym funkcja celu skupia się przede wszystkim na identyfikatorze klasy. Rozwiązaniem tego problemu mogłaby być modyfikacja funkcji celu stosowanej w procedurze uczącej. Drugi powód może się wiązać z ograniczeniami samej architektury systemu, która opiera się na sztywnych założeniach, co do danych wejściowych. Do tych założeń należy między innymi określona liczba elementów potrzebnych do utworzenia wzorca lub kolejność widoków w sekwencji wejściowej. Warto zwrócić też uwagę, że obie metody zostały zaproponowane dla zbioru ModelNet-40, w których każdy obiekt był prezentowany przez dużą liczbę widoków.

Model MVCNN-Sygnalityka ma stosunkowo prostą budowę, ale nie wykorzystuje w zasadzie informacji o strukturze zbioru wejściowego. Dane wewnątrz modelu są uśredniane w taki sam sposób niezależnie od porządku i właściwości podawanych zdjęć. Metoda jest zaprojektowana

pod stałą liczbę widoków. Model RotationNet-Sygnalityka dał nieco lepsze wyniki w testach niż MVCNN-Sygnalityka, ale konstrukcja modułu agregującego jest bardzo skomplikowana. Wielkość modułu rośnie wraz ze wzrostem liczby akceptowanych widoków V. W czasie klasyfikacji istnieje możliwość utworzenia wzorca na podstawie zestawu danych mniejszego niż V, ale długość zestawu nie może być większa niż V. Dodatkowo model ten zakłada pewien porządek sekwencji, bo przy większej liczbie V sprawdzenie wszystkich możliwych kombinacji jest nieopłacalne, widoki nie mogą się też powtarzać. Model RotationNet-Sygnalityka został zaprojektowany do klasyfikacji na zamkniętym zbiorze etykiet klas, a ze względu na poziom komplikacji funkcji składającej informację z wyjść modelu, sposób ekstrakcji wzorca do zastosowania w testach otwartych nie jest oczywisty. Z dostrzeżonych w trakcie badań problemów wynikła potrzeba ulepszenia omawianych rozwiązań. W założeniu rozwiązanie powinno mieć stosunkowo prostą strukturę, taką jaką ma sieć MVCNN-Sygnalityka. We wcześniejszych analizach sygnalizowano już, że schemat budowy sieci MVCNN-Sygnalityka pozwala na wprowadzenia własnych bardziej wyszukanych modułów agregujących niż warstwa maksimum lub uśredniająca. Jednocześnie dobrze, gdyby docelowy model brał pod uwagę rodzaj ujęcia, jakie występuje w zestawie wejściowym danych. Na podstawie tych zaleceń autorka zdecydowała się na wprowadzenie własnego modelu sieci agregującej sieci, w której moduł agregujący jest zaimplementowany z zastosowaniem mechanizmu atencji. Zaproponowany model sieci agregującej będzie nazywany SygnaT, która to nazwa jest skrótem od Sygnalityka-Transformer.

Mechanizm atencji jest szeroko stosowany w sieciach neuronowych opartych o architekturę *transformers*. Architektura sieci *transformers* zyskała niezwykłą popularność ze względu na wyniki w rozwiązywaniu zagadnień języka naturalnego, takich jak modelowanie języka i tłumaczenie maszynowe. Przetwarzanie obrazów w formie sekwencji nie wydawało się na początku rozwoju algorytmów intuicyjnym rozwiązaniem. Mimo to rozwiązanie określane jako *vision transformers* również okazało się być skuteczne i dorównywało pod względem wskaźników dokładności rozwiązaniom opartym o sieci splotowe. Sieci typu *vision transformers* takie jak ViT [67] i DeiT [68] opierają się na pomyśle, w którym obraz jest przekształcany do sekwencji przez podział obrazu wejściowego na wycinki. Każdy z wycinków jest przetwarzany do postaci tokena, który koduje jedną z pozycji w sekwencji. W zaproponowanym przez nas modelu SygnaT wykorzystano wiele z koncepcji, które są podstawą działania sieci *vision transformers*, a podstawą modułu agregującego w modelu jest wspomniany wcześniej mechanizm atencji.

4.3.1 Transformers - sieci neuronowe z mechanizmem atencji

Transformers to rodzaj sieci głębokich wewnątrz których wykorzystuje się bloki z mechanizmem atencji. Mechanizm atencji w strukturze sieci został zaproponowany w artykule [69] do rozwiązania zadania tłumaczenia języka naturalnego. Na bazie pomysłu powstało wiele metod, których celem jest analiza słów i zdań, takich jak BERT [70] i GPT-2 [71]. Bazowo sieci typu *transformers* są oparte o strukturę enkoder-dekoder [69]. Zarówno enkoder, jak i dekoder składa się z bloków, które implementują mechanizm atencji, a dokładniej moduły typu *multi-head attention*. Wraz z rozwojem metod pojawiły się też metody, w których wykorzystywany był sam enkoder ze struktury sieci *transformer* - przykładem tego podejścia jest sieć BERT [70]. Modelowanie języka naturalnego jest w niej wykonywane jedynie za pomocą enkodera, który uczy się zależności między fragmentami zdania reprezentowanymi przez tokeny w sposób częściowo nienadzorowany poprzez losowe maskowanie części tokenów. W tym podejściu do uczenia maskowane jest ok. 15% tokenów w sekwencjach treningowych, a sieć miała za zadanie odtworzyć ich wartość z kontekstu. Dzięki temu zabiegowi model był zmuszony w procesie uczenia do wykrycia zależności w zdaniach. Dzięki bardzo dobrym wynikom w standardowych zadaniach z zakresu modelowania języka naturalnego, w tym w tłumaczeniu, generowaniu odpowiedzi na pytania i analizie sentymentu, sieci *transformers* zyskały szybko popularność.

Dopiero w toku rozwoju badań sieci te zostały zastosowane do analizy obrazów. W jednej z pierwszych prac [67] na ten temat wprowadzono model sieci ViT (*vision transformer*). W metodzie ViT obraz jest dzielony na wycinki o rozmiarze 16×16 i rozwijany w sekwencję. Wartości pikseli w wycinku są traktowane jako token (analogicznie do tokenów reprezentujących słowa lub fragmentu tekstu w przypadku analizy języka). Autorzy w artykule opisują także możliwość wykorzystania sieci hybrydowej, w której wycinki nie są wykorzystywane bezpośrednio jako wektor cech, ale przetwarzane przez sieć splotową. Proponowany w pracy model SygnaT można przyrównać właśnie do takiej sieci hybrydowej ViT. W architekturze sieci SygnaT można wyróżnić enkodery pojedynczych wyników, na których wyjściu powstają deskryptory single-view, które mogą być traktowane jako tokeny należące do sekwencji wejściowej. We wspomnianej już metodzie analizy obrazów za pomocą sieci typu *transformers* - metodzie ViT [67] wykorzystywane są dwa dodatkowe mechanizmy, które wywodzą się z metody

BERT [70] i które mają istotny wpływ na działanie modelu. Pierwsza to wykorzystanie dodatkowego tokenu klasy <CLS>, który dodawany jest na początku sekwencji. Jako że sieci typu *transformers* w założeniu dokonują transformacji sekwencji w sekwencji należy wprowadzić dodatkowe podejście w przypadku potrzebny wytworzenia wartości skalarnej, która opisuje całą sekwencję. Taka potrzeba pojawia się w przypadku zadania klasyfikacji i wtedy przydaje się token specjalny. Wyjście, które odpowiada tokenowi <CLS> jest wykorzystywane jako podstawie do klasyfikacji końcowej całej sekwencji. Kolejnym mechanizmem specyficznym dla sieci *transformers* jest kodowanie pozycyjne (ang. *postional encoding*). Kodowanie pozycyjne ma zapewnić przenoszenie informacji o pozycji tokenu w sekwencji. Jest to szczególnie istotne w przypadku analizy języka, żeby przekazać dalej informację o kolejność fragmentów zdania oraz w przypadku wycinków obrazu, dzięki temu można zachować dane o wzajemnym położeniu wycinków w przestrzeni. W modelach do identyfikacji obrazów twarzy nie jest to niezbędne, ponieważ kolejność tokenów nie jest istotna, a jedynie widok, który jest na nich ujęty.

4.3.2 Moduł multihead self-attention (MHA)

Sieć agregująca SygnaT zawiera w swojej strukturze moduł agregujący z mechanizmem atencji (*ang. attention*), a mówiąc bardziej precyzyjnie jest to struktura typu *multi-head self-attention*, która w dalszej części tekstu będzie nazywana w skrócie MHA. Sieci wykorzystujące MHA zostały wprowadzone w artykule [69]. Mechanizm atencji został wprowadzony jako alternatywny w stosunku do sieci rekurencyjnych sposób analizy sekwencji, w którym można wychwycić zależności pomiędzy poszczególnymi jej elementami. W artykule [69] atencja jest zdefiniowana jako funkcja, która mapuje pewną sekwencję wektorów Q oraz zestaw par wektorów kluczy K i wektorów wartości W w wektor wyjściowy. Wyjście funkcji jest średnią ważoną wektorów wartości V, przy czym wagi stosowane przy uśrednianiu nazywane również atencją są wyliczane na podstawie funkcji zgodności pomiędzy wektorami Q i wektorami K. Do wyliczenia zgodności wykorzystuje się funkcję *scaled dot-product attention* (skalowaną atencję skalarną), która ma następującą postać:

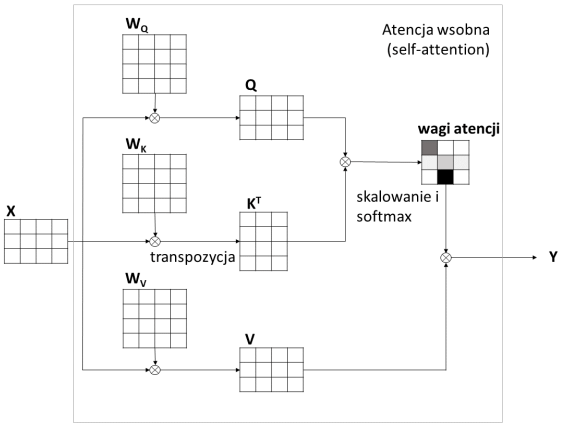
$$Att(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Wejściem funkcji jest zbiór wektorów zapytań Q (*queries*), kluczy K (*keys*) i wartości V (*values*).

Atencja jest dzielona przez współczynnik $\sqrt{d_k}$, gdzie d_k to rozmiar wektorów kluczy K. W przypadku naszego modułu stosujemy atencję wsobną (*ang. self-attention*), w której wszystkie zestawy wektorów Q, K, W są przekształceniem tej samej sekwencji wejściowej X.

$Q = XW_Q, K = XW_K, V = XW_V$

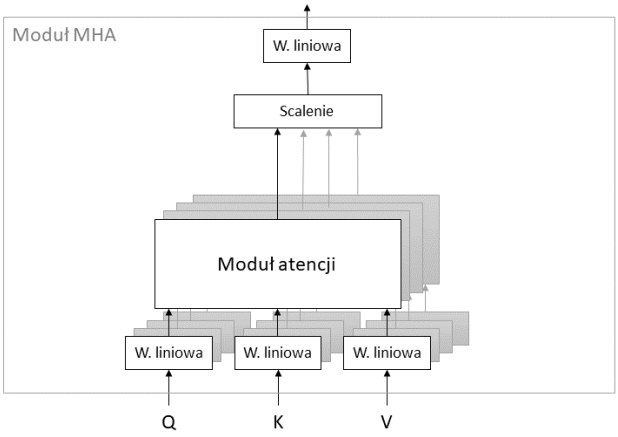
Koncepcja przepływu informacji w module wykorzystującym mechanizm atencji wsobnej jest przedstawiona na rysunku 4.10.



Rysunek 4.10: Funkcja atencji wsobnej (ang. *self-attention*). Macierz wejściowa X jest przekształcana do trzech reprezentacji wewnętrznych zapytań Q (*queries*), kluczy K (ang. *keys*), wartości V (ang. *values*). Wagi atencji wyliczane są na podstawie iloczynu macierzy Q i K^T.

W metodzie zaproponowanej w pracy sekwencja wejściowa X to zbiór wektorów reprezentujących poszczególne obrazy, w przeciwieństwie do metody VIT, gdzie są to wycinki jednego obrazu. X ma rozmiar $S \times D$, gdzie S to długość sekwencji, a D to rozmiar reprezentacji. Dla ułatwienia porównań rozmiar reprezentacji został ustalony na 2048, który jest równy wielkości deskryptorów widoków w modelach MVCNN-Sygnalityka i RotationNet-Sygnalityka. Moduł atencji zawiera wewnątrz trzy macierze wag, które są uczone w trakcie treningu. Są to macierze: W_Q, W_K, W_V i służą do wytworzenia odpowiednio trzech wewnętrznych reprezentacji Q, K i V, o których była mowa wcześniej (*queries*, *keys* i *values*). Sposób kodowania macierzy reprezentacji wewnętrznej w module atencji może wydawać się na pierwszy rzut oka nadmiernie skomplikowany, ale po analizie struktury widać, że jest to sposób na wyliczenie wag (wektorów atencji) za pomocą których możemy uśrednić dane wejściowe z uwzględnieniem

ich wzajemnych zależności. W sieciach typu *transformers* stosowany jest zrównoleglony moduł atencji nazywany *multi-head self-attention* (MHA). Sam mechanizm aplikacji funkcji skalowalnej atencji w module MHA pozostaje bez zmian, ale operacja wykonywana jest równolegle na N fragmentach wektora wejściowego w równoległych modułach nazywanych również głowami (ang. *heads*). Wektor wynikowy powstaje jako sklejenie wyników z poszczególnych modułów składowych. Schemat modułu MHA został przedstawiony na rysunku 4.11.



Rysunek 4.11: Schemat działania modułu MHA (ang. *multi-head-self-attention*). Dane wejściowe dzielone są na N części nazywanych również głowami (ang. *heads*), gdzie N to zdefiniowana liczba równoległych modułów składowych . Każda z części przetwarzana jest przez inny moduł atencji wsobnej, a potem wyniki są sklejane w jedną macierz wynikową.

4.3.3 Struktura modelu SygnaT

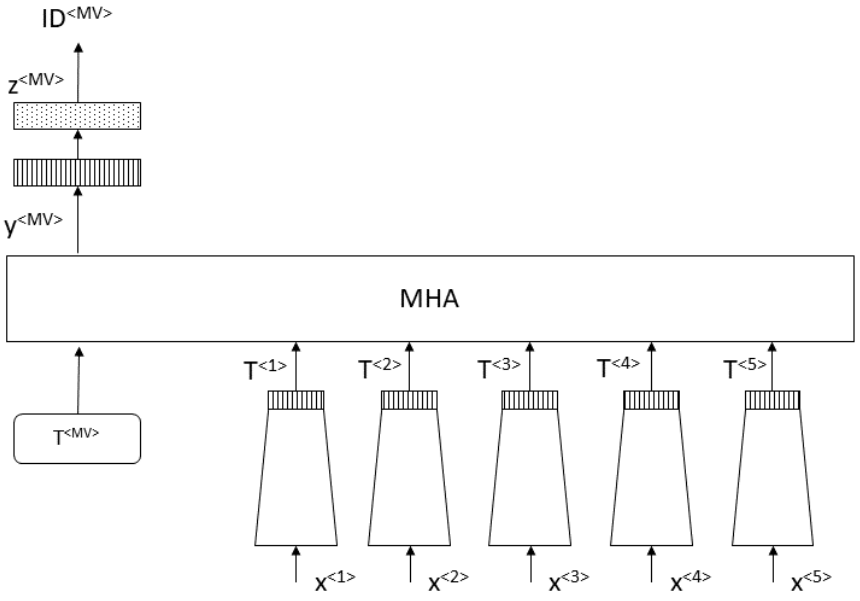
W architekturze modelu SygnaT można wyróżnić kilka części składowych. Tak jak w przypadku modeli MVCNN-Sygnalityka i RotationNet-Sygnalityka wykorzystuje się enkoder pojedynczych widoków, który przekształca pojedyncze obrazy należące do zbioru wejściowego $X^{<i>}$ do postaci deskryptorów $T^{<i>}$ o zadanej długości. Dla spójności z pozostałymi metodami opartymi o koncepcje sieci *transformers*, deskryptory te będą nazywane tokenami. Sieć, która dokonuje przekształcenia jednego obrazu wejściowego w token, to w przypadku sieci SygnaT enkoder o architekturze ResNet-50 wytrenowany wcześniej na zbiorze VGGFace2 (model **bazowy-SV**) i opisany w rozdziale 2.2. Enkoder tokenów przyjmuje na wejściu barwne

obrazy o rozmiarze 224×224, a na wyjściu zwraca deskryptor o rozmiarze 2048. Tokeny pojedynczych obrazów $T^{<i>}$ stanowią podstawę do utworzenia sekwencji wejściowej dla modułu MHA, czyli modułu agregującego z mechanizmem atencji.

Jako, że sieci typu *transformers* zaprojektowane są do przekształcania sekwencji w sekwencję stosowane są różne podejścia w przypadku konieczności określenia właściwości całego zbioru. Pierwszym podejściem jest dodawanie tokenów specjalnych do sekwencji wejściowej, które mają specjalną interpretację. W modelach takich jak BERT [70] i VIT [67] stosowany jest token specjalny <CLS>, który odpowiada za klasyfikację całego obrazu bądź całego zdania. Analogicznie do tokenu <CLS> w modelu SygnaT wprowadzono token <MV> do generacji zagregowanego deskryptora dla zestawu obrazów $X^{<MV>}$. Token specjalny <MV> jest inicjalizowany wartościami losowymi, które są aktualizowane w procesie uczenia sieci, tak samo jako wagi samej sieci. Wyjście powiązane z tokenem <MV> jest używane jak wynik końcowy agregacji. To podejście będzie określane w pracy jako **klasyfikacja tokenem** lub w skrócie **SygnaT (token)**. Drugim standardowym podejściem jest uśrednienie wyjść modułu MHA dla wszystkich elementów wejściowych sekwencji. To podejście będzie nazywane **klasyfikacją uśrednioną** i oznaczone w porównaniach jako **SygnaT (avg)**.

- W przypadku klasyfikacji tokenem (podejście oznaczane w skrócie jako **token**) deskryptor zestawu $X^{<i>}$ wyliczany jest jako wyjście powiązane z tokenem specjalnym <MV>: $\hat{y} = y^{<MV>}$
- W przypadku klasyfikacji uśrednionej (w skrócie **avg**) deskryptor zestawu $X^{<i>}$ wyliczany jest jako średnia arytmetyczna z wyjść dla wszystkich tokenów wejściowych: $\hat{y} = \frac{\sum_i^N y^{<i>}}{N}$

Struktura zaproponowanej sieci SygnaT została przedstawiona w sposób graficzny na rysunku 4.12. W eksperymentach używano modułu MHA złożonego z ośmiu równoległych modułów atencji tzw. głów , a znaczy to, że każdy z tokenów wejściowych jest cięty przed wejściem do modułu na osiem równych części. Części są przetwarzane równolegle przez głowy modułu MHA. Potem wektory wynikowe $Y^{<i>}$ powstają przez sklejanie wektorów składowych z poszczególnych głów.



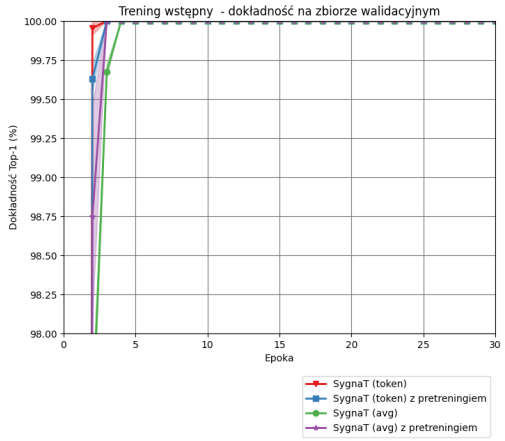
Rysunek 4.12: Schemat modelu SygnaT zaproponowanego przez autorkę pracy. Sekwencja tokenów wejściowych, która jest przetwarzana przez moduł agregujący, jest wynikiem transformacji obrazów w zestawie wejściowym za pomocą enkodera single-view. Na rysunku pokazana jest wersja sieci SygnaT z tokenem specjalnym <MV>. Moduł agregujący wykorzystuje mechanizm atencji do składania tokenów wejściowych wraz z tokenem specjalnym i wytworzenia finalnej reprezentacji.

4.3.4 Eksperymenty

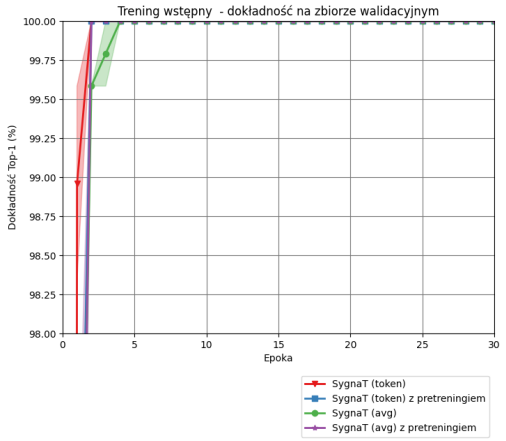
Dla zachowania zgodności z badaniami opisanymi w poprzednich rozdziałach eksperymenty przeprowadzone zostały na zbiorze MV-Multi-PIE, z walidacją na zbiorze zamkniętym i z testem w scenariuszu na zbiorze otwartym. Eksperymenty przeprowadzono dla zbiorów pięcioujęciowych. Enkoder tokenów dla pojedynczych widoków został zainicjalizowany wagami modelu **bazowy-SV**. Moduł agregujący opiera się na architekturze MHA z ośmioma głowami. Wielkość reprezentacji wewnętrznej wynosi 2048, tak więc wielkość deskryptora dla pojedynczej głowy wynosi 256. Inne wielkości modułu MHA zostały zbadane w ramach prac wstępnych nad implementacją metody, nie otrzymano jednak wyraźnego zysku ze zwiększa-

nia liczby głów w module, dlatego wybrano liczbę 8. Deskryptor końcowy, który używany jest przy wyliczaniu podobieństwa między próbkami, uzyskiwany jest na dwa sposoby: poprzez uśrednienie wyjść dla wszystkich tokenów wejściowych (opcja **avg**) oraz przez wykorzystane wyjścia tokena specjalnego <MV> (opcja **token**). W odróżnieniu od modeli MVCNN-Sygnalityka i RotationNet-Sygnalityka, omawianych w poprzednich rozdziałach, w których losowo inicjalizowana była jedynie warstwa ostatnia, model SygnaT wymagał inicjalizacji wag całej struktury MHA. Ze względu na potencjalnie niedostateczne rozmiary zbioru treningowego MV-Multi-PIE zdecydowano się dokonać pretreningu sieci na zbiorze VGGFace2. Zbiór VGGFace2 nie ma co prawda struktury odpowiedniej do założonej w badaniach - nie ma podziału na zbiory wieloujęciowe. Z tego powodu został zastosowany jedynie do treningu wstępnego. Na etapie ewaluacji i obserwacji wyników pośrednich, w której wymagana była informacja o pozie twarzy obrazie użyto etykiet wyestymowanych na podstawie położenia punktów charakterystycznych na twarzach widocznych na zdjęciach.

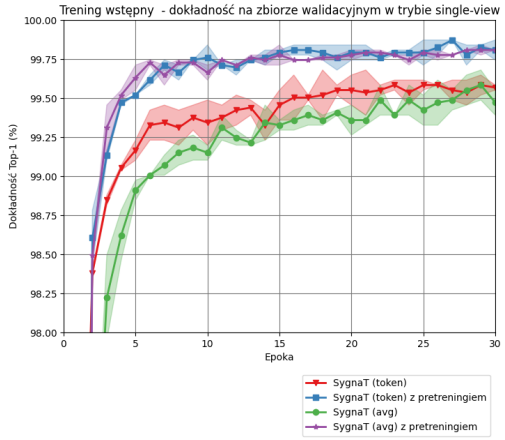
Wyniki klasyfikacji na zbiorze treningowym i walidacyjnym dla klasyfikacji MV bardzo szybko osiągnęły wartości współczynnika dokładności Top-1 bliskie 100%. Wyniki na zbiorze otwartym są wyższe niż dla analizowanych wcześniej metod MVCNN-Sygnalityka i RotationNet-Sygnalityka. Ponownie kluczowym dla całego procesu okazał się pretrening sieci na zbiorze VGGFace2 i transfer wag. Wyniki końcowe Rank-1 dla testu identyfikacji na zbiorze otwartym wyniosły 98,43% dla modelu SygnaT (avg) i 98,33% dla model SygnaT (token). Nie zanotowano znaczących różnic pomiędzy siecią SygnaT (token), a SygnaT (avg). Dla sieci SygnaT (avg) wyniki są nominalnie wyższe, ale mieściły się w zakresie zmienności. Wyniki dla sieci SygnaT są znacząco wyższe niż dla pozostałych modeli, szczególnie w zastosowaniach, w których próbka zapytanie jest tworzona na podstawie pojedynczego ujęcia. Architektura sieci SygnaT ma w dodatku zbliżoną strukturę do modelu MVCNN-Sygnalityka, przez co pozwala na transfer wag zarówno w zakresie enkodera tokenów, jak również enkodera i modułu agregującego. Sieć SygnaT jest dostosowana do zadania, nie traktuje wszystkich obrazów w zbiorze wejściowym bezkontekstowo, jak ma to miejsce w przypadku MVCNN-Sygnalityka. Moduł atencji uczy się wzajemnych zależności między kolejnymi tokenami. W przeciwieństwie do sieci RotationNet-Sygnalityka generacja deskryptora multi-view jest częścią architektury sieci SygnaT i można z niej łatwo uzyskać identyfikator klasy oraz deskryptor opisujący zbiór wieloujęciowy.



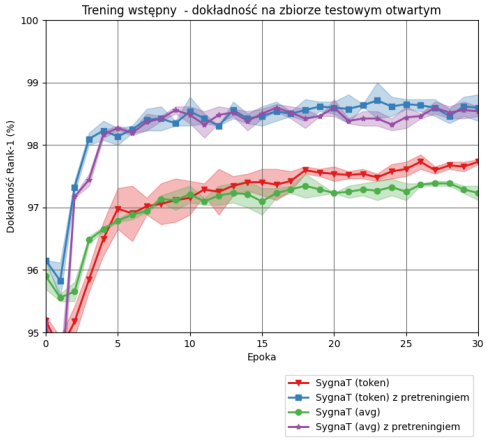
(a) Dokładność klasyfikacji MV na zbiorze trenującym



(b) Dokładność klasyfikacji MV na zbiorze walidacyjnym



(c) Dokładność klasyfikacji na podstawie deskryptorów SV na zbiorze walidacyjnym



(d) Dokładność identyfikacji w scenariuszu open-set na zbiorze testowym w trybie SV-MV.

Rysunek 4.13: Trening modelu SygnaT na bazie MV-Multi-PIE dla dwóch podejść do agregacji wyniki końcowego: deskryptor z tokena klasy (token) oraz uśrednione deskryptory (avg). W badaniu uwzględniono dwie różne metodologie treningu: tylko na zbiorze MV-Multi-PIE oraz z pretreningiem na zbiorze VGGFace2

Metoda	Rank-1	Rank-1 (tylko frontalne)	Rank-1 (tylko półprofile)	Rank-1 (tylko pełne profile)
Model bazowy single-view (avg)	92.18%	95.05%	92.86%	80.95%
MVCNN-Sygnalityka (avg-pool)	96.54%	98.17%	96.89%	89.56%
RotationNet-Sygnalityka	96.63%	97.80%	97.07%	89.93%
SygnaT (token)	98,33%	98,90%	97,43%	**97,07%**
SygnaT (avg)	**98,43%**	**99,81%**	**97,61%**	95,60%

Tabela 4.4: Porównanie wyników identyfikacji na bazie MV-Multi-PIE z walidacją skrośną dla modeli MVCNN-Sygnalityka, RotationNet-Sygnalityka oraz dla dwóch wariantów sieci SygnaT. W scenariuszu testowym przyjęto pięcioujęciowe zestawy danych wejściowych przy budowaniu galerii. Identyfikację testowano na całym zbiorze próbek-zapytań oraz na wybranych rodzajach ujęć.

4.3.5 Test na zbiorze IJBC-C

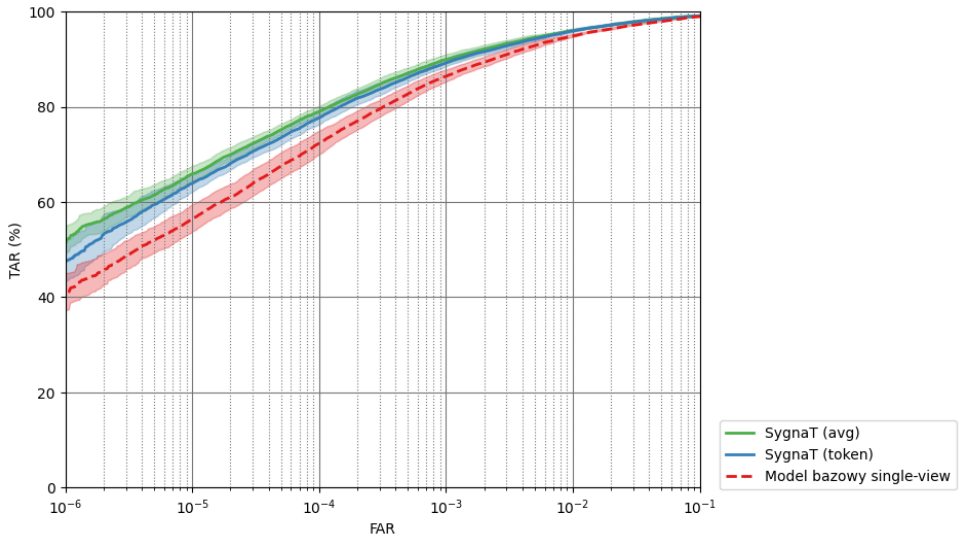
Głównym założonym celem była optymalizacja procesu rejestracji próbek dla zbiorów wieloujęciowych, tzn. składających się z określonego zestawu ujęć. Budowa sieci SygnaT pozwala na dosyć dużą elastyczność jeśli chodzi o zastosowania sieci do agregacji próbek - pozwala zestawić ze sobą dowolną liczbę próbek wejściowych. Nie zakłada również porządku sekwencji, tak jak ma to miejsce w przypadku sieci RotationNet-Sygnalityka. W całym procesie treningu przyjęto jednak określoną strukturę podzbiorów zdjęć twarzy. Mimo to elastyczność zaproponowanej architektury skłaniała do sprawdzenia działania modelu w trudniejszym scenariuszu użycia. Model SygnaT został wykorzystany do zbudowania systemu identyfikacyjnego w oparciu o benchmark testowy IJB-C, dla którego prezentowano wcześniej wyniki bazowego modelu single-view. W scenariuszu IJB-C występuje podział na podzbiory do tworzenia wzor-

ców, ale mają one bardzo zróżnicowane liczności, a wewnątrz znajduje się wiele obrazów nienadzorowanych i o różnych, ale nieokreślonych pozach. Test jest rozbudowany (pojedyncza galeria w teście składa się z prawie 20 000 wzorców i ponad 120 000 próbek). Mimo trudności testu i sporych wymagań obliczeniowych badanie wykazało, że model SygnaT może być z powodzeniem zastosowany w takim scenariuszu. Wyniki identyfikacji dla testu 1:N w formie krzywych CMC oraz krzywe ROC zostały przedstawione na wykresach 4.14 i 4.15. Model SygnaT użyty w porównaniu był wytrenowany na bazie VGGFace2. Krzywe CMC i ROC dla dwóch wersji modelu SygnaT zostały zestawione ze wcześniejszymi wynikami obliczonymi dla modelu opartego o kodowanie pojedynczych widoków - model **bazowy-SV**.

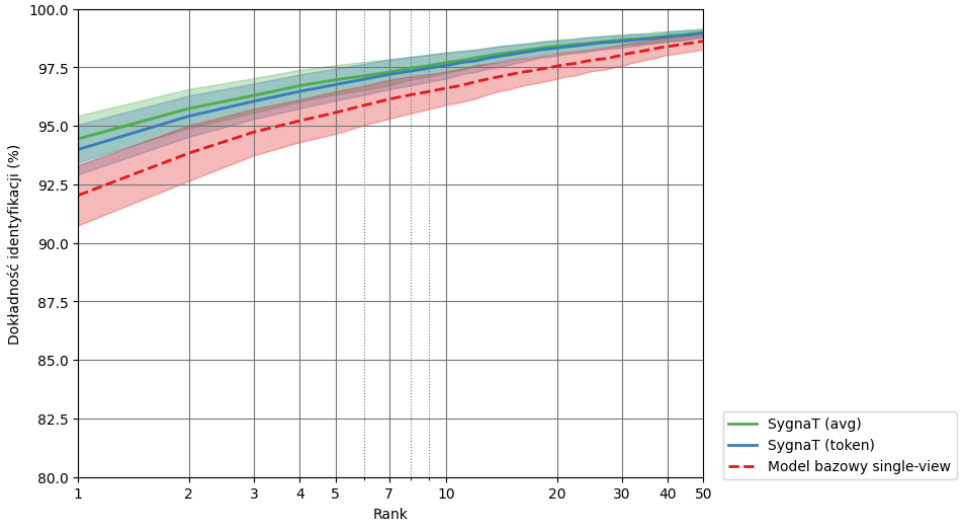
Jak widać na wykresach 4.14 i 4.15 zastosowanie sieci SygnaT pozwoliło na poprawienie wyników identyfikacji również w teście uniwersalnym (bez wyróżnienia zbiorów wieloujęciowych) zarówno dla wersji modelu z klasyfikacją uśrednioną (**avg**), jak i dla wersji z **tokenem**. W przypadku badań na zbiorze IJB-C można zaobserwować wyraźną przewagę wersji modelu SygnaT z klasyfikacją uśrednioną. Model SygnaT (avg) uzyskał wskaźnik Rank 1 równy 94.45% (przy współczynniku równym 92.03% dla modelu bazowego) oraz TAR@FAR=1e-3 równy 89.93% (przy współczynniku równym 86.39% dla modelu bazowego. Eksperyment ten pokazuje jak szerokie potencjalne zastosowanie może mieć model SygnaT przy budowaniu baz wzorców do identyfikacji twarzy.

Model	Weryfikacja (TAR)		Identyfikacja (TPIR)	
	FAR=1e-3	FAR=1e-2	Rank-1	Rank-3
Model bazowy single-view	86.39%	94.87%	92.03%	94.73%
SygnaT (token)	89.23%	95.92%	93.99%	96.05%
SygnaT (avg)	**89.93%**	**96.01%**	**94.45%**	**96.30%**

Tabela 4.5: Wyniki dla protokołu testowego do identyfikacji 1:N z benchmarku IJB-C



Rysunek 4.14: Krzywa oceny weryfikacji ROC wyliczona na podstawie testu Identyfikacja 1:N z benchmarku IJB-C. Pokazana wartość jest uśredniona dla porównań względem galerii G1 i galerii G2. Szerokość cieniowania zależy od odchylenia standardowego wyników. Na wykresie porównano modele: SygnaT z klasyfikacją uśrednioną, SygnaT z klasyfikacją tokenem oraz model **bazowy-SV**.



Rysunek 4.15: Krzywe dokładności identyfikacji CMC wyliczone na podstawie testu Identyfikacja 1:N z benchmarku IJB-C. Na wykresie porównano modele: SygnaT z klasyfikacją uśrednioną, SygnaT z klasyfikacją tokenem oraz model **bazowy-SV**.

Rozdział 5

Podsumowanie

W pracy przedstawiono metody identyfikacji twarzy na podstawie zdjęć wieloujęciowych, takich jak zdjęcia sygnalityczne, które w przyjętym scenariuszu składają się z 5-elementowego zestawu ujęć: pełny lewy profil, lewy półprofil, portret frontalny, prawy półprofil, pełny prawy profil. Modele wykorzystane w badaniach zostały zainspirowane algorytmami do klasyfikacji obiektów 3D, w tym modelem MVCNN [1] i modelem RotationNet [2]. Zaproponowano również własny model agregujący wykorzystujący mechanizm atencji [69], nazywany w pracy SygnaT.

Modele MVCNN i RotationNet w wersji zaproponowanej przez autorów artykułów [1] i [2] do klasyfikacji obiektów były trenowane i testowane na zbiorze ModelNet-40 [20], który składa się z 40 klas obiektów 3D. Obiekty 3D przed podaniem do obu rodzajów sieci były przetwarzane do sekwencji ujęć danego obiektu pod różnymi kątami. W praktyce więc modele te przetwarzały zbiory wieloujęciowe i stąd pomysł by zastosować je do wieloujęciowych zdjęć twarzy.

Używane w pracy zbiory zdjęć wieloujęciowych twarzy nie miały swojego źródła w skanach 3D, ale warto podkreślić, że istnieje taka możliwość. W takim przypadku wszystkie przedstawione metody można zakwalifikować jako metody klasyfikacji 3D twarzy. W trakcie badań naukowych wykraczających poza zakres tej pracy autorka implementowała algorytmy rozpoznawania twarzy operujące na danych 3D (wyniki opisane zostały w artykule [6]). Wnioski z testowania algorytmów operujących bezpośrednio na chmurach punktów 3D wykazały, że trudno jest uzyskać na tyle dobre wyniki, żeby uzasadniały znaczący wzrost wymagań oblicze-

niowych związanych z przetwarzaniem danych 3D. Kolejną barierą, która była ograniczeniem w uzyskaniu dokładności porównywalnej z metodami rozpoznawania twarzy 2D były stosunkowo małe i gorzej dostępne bazy danych 3D. Dlatego też często sugerowanym podejściem do analizy obiektów 3D jest właśnie przetwarzanie sekwencji ich projekcji na płaszczyznę 2D pod różnymi kątami. W artykule [1] pokazano, że takie podejście generuje lepsze wyniki klasyfikacji niż klasyczne podejścia 3D. Dodatkowo w przypadku rozpoznawania zdjęć twarzy ważnym aspektem jest możliwość transferu wag z modeli single-view (wytrenowanych dla zdjęć pojedynczych) dla których jest więcej odpowiednich do treningu baz danych. Transfer wag, jak pokazano w pracy, znacząco polepsza wyniki końcowe.

W ramach prowadzonych badań przygotowano środowisko testowe, przez które rozumiane jest zebranie zbiorów danych oraz dostosowanie scenariuszy testowych. Architektura wieloujęciowego systemu identyfikacji była związana ze specyficznymi wymaganiami odnośnie zbiorów testowych i trenujących. Budowa zbiorów testowych zakładała podział na dwa podzbiory: próbki galeryjne, które powinny zawierać kilka zestawów zdjęć twarzy dla jednej osoby i na podstawie których tworzona była baza referencyjna oraz próbki-zapytania, które były wyszukiwane w bazie. Z powodu braku dostępnych zbiorów wieloujęciowych obrazów twarzy, środowisko wymagało zebrania takich danych lub dostosowania istniejących zgodnie z wymaganiami projektu. W eksperymentach omawianych w pracy wykorzystano 4 różne zbiory danych:

- MV-MultiPIE - zbiór zdjęć wykorzystujący obrazy ze zbioru CMU MultiPIE [24], podzielony na podzbiory wieloujęciowe oraz poddany detekcji twarzy i przycięty na potrzeby zadania
- MV-Quiscampi - zbiór do testów utworzony na podstawie obrazów z bazy Quiscampi, w której występują zarówno próbki nienadzorowane z zewnętrznego monitoringu wizyjnego, jak i wieloujęciowe zdjęcia galeryjne tych samych osób, zbiór wymagał korekty i uzupełnienia współrzędnych detekcji twarzy, poprawy etykiet oraz wyodrębnienia dodatkowych obrazów z klatek wideo
- VGGFace2 - popularny zbiór obrazów twarzy do treningu sieci neuronowych. W części eksperymentów używane były automatycznie wygenerowane etykiety pozy głowy.
- IJBC-A - rozbudowany benchmark do testowania algorytmów rozpoznawania twarzy

Zbiorów danych jest kilka, ponieważ część z nich spełniała warunki stawiane bazom do treningu sieci, a część z nich była potrzebna do testowania w wybranych scenariuszach. Omówienie i cechy poszczególnych zbiorów zostało umieszczone w dodatku 1.7 „Bazy danych i scenariusze testowe".

Projekt metod identyfikacji na podstawie zdjęć wieloujęciowych wymagał przeglądu i usystematyzowania aktualnie używanych w tym celu metod. Skupiono się przede wszystkim na rozwiązaniach problemu tworzenia wzorców na podstawie zestawów zdjęć oraz problemów związanych z przetwarzaniem zdjęć niefrontalnych. Wybrane podejścia zostały zaimplementowane w formie biblioteki programistycznej w języku Python oraz przetestowane z użyciem wspomnianych wcześniej scenariuszy testowych. W badaniach wyróżniono następujące podejścia:

- normalizacja pozy 3D w obrazie
- agregacja wzorców pojedynczych
- oraz modele agregujące

Prace nad systemem identyfikacji rozpoczęły się od wytrenowania bazowego modelu single-view do identyfikacji twarzy na podstawie zdjęć pojedynczych. Model był oparty o architekturę ResNet-50 [3] i został wytrenowany na zbiorze VGGFace2 [4]. Na podstawie modelu single-view zbadano i omówiono wpływ pozy głowy na dokładność systemu (opis eksperymentów w rozdziale 2.2.4 „Poza jako czynnik wpływający na dokładność identyfikacji twarzy"). Wytrenowanie własnego modelu bazowego było też kluczowe z drugiego powodu - wagi z modelu single-view były częściowo transferowane do pozostałych omawianych modeli. Wykorzystanie tego samego rdzenia we wszystkich metodach sprawiło, że łatwiej je było ze sobą porównać. Warto przy okazji zwrócić uwagę, że moduł enkodera single-view może być wymieniony w docelowych modelach na inny wytrenowany z użyciem preferowanej architektury sieci. Dzięki takiemu podejściu, nie trzeba się skupiać na wyborze konkretnej architektury sieci głębokiej, co samo w sobie jest szerokim i dynamicznie rozwijającym się tematem badawczym. W dalszej części badań model single-view był używany pod nazwą **bazowy-SV**.

W rozdziale 3.2 „Automatyczna korekta pozy" przeprowadzono eksperymenty nad normalizacją pozy 3D z wykorzystaniem metody *unsup3D* [60]. Metoda unsup3d składa się z grupy sieci generatywnych odpowiedzialnych za wytworzenie różnych aspektów sztucznych ujęć,

takich jak tekstura, geometria, czy cieniowanie. Ze względu na niezadowalające wyniki identyfikacji z użyciem sztucznie wygenerowanych zdjęć metoda ta została odrzucona z dalszych badań.

W następnym kroku sprawdzono, czy agregacja cech biometrycznych z wielu zdjęć jest właściwym kierunkiem rozwoju systemu identyfikacji. W rozdziale (3.3 „Agregacja modeli ") pokazano, że ma to sens zarówno po stronie galerii, jak i próbek-zapytań. Dzięki temu wiemy, że zwiększanie liczby zdjęć ma pozytywny wpływ na dokładność wyników przy tworzeniu bazy identyfikacyjnej oraz przy wyszukiwaniu. Autorka pracy zauważyła niestety znaczący spadek dokładności dla zdjęć w ujęciach skrajnych, takich jak profilowe. Intuicyjnie zbiory ze zdjęciami o zróżnicowanej pozie powinny tworzyć lepsze wzorce, ponieważ pokazują większy zakres cech twarzy niż same zdjęcia frontalne. Niestety na początkowym etapie badań obecność zdjęć profilowych w zestawach danych powodowała pogorszenie jakości wzorców. Testowanie metod agregacji deskryptorów biometrycznych objęło sprawdzenie funkcji matematycznych takich średnia, mediana, minimum i funkcja softmin [62]. Najlepsze wyniki uzyskano przy uśredniania deskryptorów oraz funkcji softmin.

W dalszej części pracy opisanej w rozdziale 3.4 „Modele agregujące" rozwijano modele agregujące, w których wszystkie elementy systemu rejestracji wzorca były realizowane za pomocą jednego modułu sieci. Punktem wyjścia do rozwoju modeli była metoda MVCNN. Model MVCNN przetwarza V-elementowy zbiór obrazów pokazujących ten sam obiekt i generuje na wyjściu deskryptor zagregowany i identyfikator dla całego zbioru. W strukturze samej sieci można wyróżnić część, która jest odpowiedzialna za ekstrakcję cech pojedynczych obrazów (enkoder single-view). Wagi enkodera zostały przetransferowane z wytrenowanego wcześniej modelu single-view (**bazowy-SV**). Na bazie metody zaimplementowano własny model MVCNN-Sygnalityka do generacji wzorców twarzy. W metodzie MVCNN-Sygnalityka w odróżnieniu od metody źródłowej zastosowano warstwę skalującą z uśrednianiem (*average-pooling*) - w oryginalne była to warstwa z operacją maksimum (*max-pooling*).

W eksperymentach wykazano, że za pomocą modelu MVCNN-Sygnalityka można stworzyć skuteczny system identyfikacji twarzy. W tym podejściu zidentyfikowano jednak problem z niepełnymi zbiorami. W metodzie nie jest jasne jak przetwarzać zbiory w których brakuje ujęć z założonego zestawu, ponieważ model MVCNN został zaproponowany dla podzbiorów o

stałej liczności. Dokładny wpływ niepełnych podzbiorów i najlepszego postępowania w takim przypadku wymaga dalszego zbadania.

Metoda MVCNN nie wykorzystywała w żaden sposób informacji o ujęciach widocznych na zdjęciach. Z tego powodu skierowano zainteresowanie na model RotationNet, który oprócz identyfikacji wieloujęciowej w sposób nienadzorowany uczył się kodu widoków zdjęć w zbiorze. Model RotationNet miał o wiele bardziej rozbudowaną warstwę modułu agregacji niż model MVCNN. W module występował podział na wyjścia z sieci dedykowane do konkretnych ujęć z założonego słownika widoków. W czasie przetwarzania zbiorów przez model następuje przypisanie najbardziej prawdopodobnej sekwencji widoków, na podstawie którego wybierane są odpowiednie wyjścia z sieci. Na bazie metody zaimplementowano własny model do identyfikacji twarzy RotationNet-Sygnalityka.

Dla obydwu metod: MVCNN-Sygnalityka i RotationNet-Sygnalityka uzyskano wysokie wyniki dokładności identyfikacji na zbiorze zamkniętym: odpowiednio 99,63% i 99,44% (dla zestawów pięcioujęciowych). Zauważono jednak niewielką poprawę na zbiorze testowym otwartym w porównaniu z agregacją deskryptorów. Dla sieci MVCNN-Sygnalityka współczynnik Rank-1 na krzywej CMC wyliczonej dla testu otwartego wynosił 96,53%, a dla sieci RotationNet-Sygnalityka wynosił 96,62%. Z powodu niesatysfakcjonujących wyników w scenariuszu otwartym zaproponowano autorską sieć z modułem agregacji wykorzystującym mechanizm atencji. Zaproponowany model został nazwany SygnaT i wykorzystywał do kodowania wzorca szereg metod stosowanych w sieciach typu *transformers*. Przetestowano dwa podejścia do składania wyniku końcowego: jeden na podstawie dodatkowego tokenu, a drugi na podstawie uśrednionych wyjść modułu atencji. Uzyskano dokładność identyfikacji równą 98,33% (wersja z tokenem) i 98,43% (wersja uśredniona). Dodatkowo po podziale próbek-zapytań w zbiorze testowym względem ich ujęć zauważono, że następuje tu znacząca poprawa identyfikacji dla zdjęć profilowych, co było celem zadania.

Rozwiązanie zostało zaproponowane do kodowanie zbiorów złożonych z przyjętych a priori zestawów ujęć. W odróżnieniu od sieci MVCNN, sieć SygnaT w dosyć prosty sposób może być użyta do zestawów danych o zróżnicowanej liczności. W ramach eksperymentu sprawdzono działanie modelu w scenariuszu testowym IJBC-A, który posłużył wcześniej do badania modelu single-view. Wyniki pokazały, że mimo że dane nie mają założonej struktury - są zbiorami

wieloobrazowymi, ale nie składają się z określonych ujęć - model uzyskał wyższe współczynniki Rank-1 niż model bazowy.

W pracy pokazano, jakie podejścia można zastosować do rejestracji bazy do identyfikacji twarzy na podstawie zdjęć wieloujęciowych. Wzorce zagregowane z użyciem modeli agregujących wieloujęciowych okazały się najlepszym rozwiązaniem problemu rejestracji. Zaproponowany przez autorkę model SygnaT wykazał wyższe współczynniki dokładności niż model bazowy single-view oraz większą odporność na obecność zdjęcia w skrajnych ujęciach w zbiorze próbek-zapytań. Ponadto model jest elastyczny i umożliwia zastosowania w różnych scenariuszach testowych, zarówno dla zamkniętego zbioru etykiet, jak i na zbiorze otwartym. Modularna budowa pozwala na wymianę enkodera wbudowanego w model agregujący na inny zdefiniowany według wybranej architektury. Model akceptuje na wejściu zestawy danych o innej liczności i nie wymaga pełnej zgodności z założonym słownikiem widoków na wejściu, co może być przydatne w przypadku danych, które nie zachowują przyjętej struktury. Ponadto na danych niestrukturalizowanych również następuje poprawa wyników w stosunku do modelu bazowego, co pokazano w eksperymentach.

Materiały dodatkowe

A Projekty powiązane z tematyką pracy

Badania prowadzone w pracy doktorskiej przebiegały często równolegle z innymi badaniami naukowymi, które autorka prowadziła w instytucie NASK (Naukowa i Akademicka Sieć Komputerowa). Niektóre z nich pokrywały się tematycznie lub dotyczyły biometrii twarzy. Dlatego też wnioski lub potrzeby, które zostały zauważone w trakcie pracy przy projektach były motywacją do prowadzenia własnych badań naukowych. Miała też miejsce sytuacja odwrotna, tzn. wiedza uzyskania w trakcie własnych badań naukowych mogła zostać wykorzystana w pracy zawodowej. Z tego powodu poniżej zamieszczam listę projektów, w których brałam udział i które są tematycznie powiązane z badaniami przedstawionymi w pracy.

A.1 BIOWIZ

Celem projektu było opracowanie i stworzenie systemu mającego na celu identyfikację osób (sprawców przestępstw) na podstawie wizerunku utrwalonego na zdjęciu lub materiałach wideo. Projekt został sfinansowany z środków Narodowego Centrum Badań i Rozwoju w ramach programu Bezpieczeństwo i Obronność Państwa (DOB-BIO7/18/02/2015). Część biometryczna systemu złożona była ze zintegrowanych modułów, w tym modułu detekcji twarzy, modułu śledzenia oraz kilku silników biometrycznych wytwarzających wzorce biometryczne twarzy i sylwetki. Budowa modułowa pozwalała na sprzężenie z zewnętrznymi bazami danych, w tym z bazą sygnalityczną Krajowego Systemu Identyfikacji Policyjnej oraz innymi materiałami operacyjnymi dostarczonym przez organy ścigania. System został zaprojektowany jako narzędzie interaktywne, mogące działać w wielu scenariuszach, np. detekcja twarzy, wy-

odrębnianie klatek z obrazami twarzy i śledzenie twarzy w materiałach wideo, identyfikacja osoby zaznaczonej na zdjęciach lub materiale wideo na podstawie profilu biometrycznego. Doświadczenie pracy z bazami zdjęć sygnalitycznych twarzy było inspiracją do stworzenia metod opisanych w pracy doktorskiej. Projekt był prowadzony od grudnia 2015 do października 2017

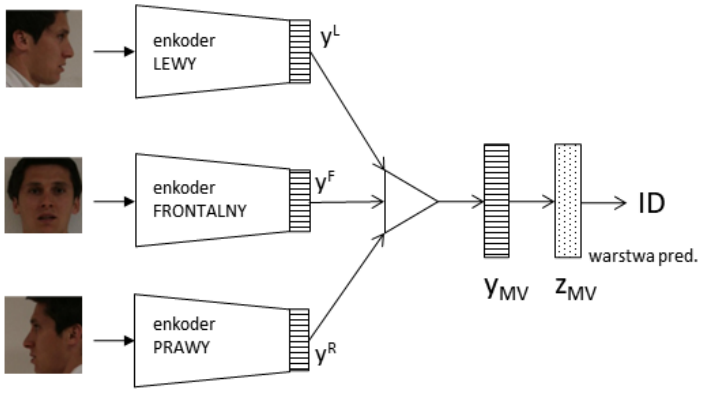
Uczestnicy: Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych (Lider), NASK PIB, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Stowarzyszenie Polska Platforma Bezpieczeństwa Wewnętrznego

Rysunek 1: Większość prac badawczych nad systemem BIOWIZ odbywała się z wykorzystaniem zdjęć związanych ze sprawami aktów chuligańskich w czasie meczów piłkarskich. Podstawowym scenariuszem zastosowania systemu była identyfikacja sprawców bądź w materiałach wideo, bądź w zdjęciach osób podejrzanych zebranych przez policję, w tym katalogach zdjęć sygnalitycznych. Materiały wideo często stanowiły wyzwanie dla systemów biometrycznych, ponieważ zawierały wiele twarzy, które mogły być osłonięte oraz nie były skierowane w stronę kamery.

A.2 Opracowanie algorytmu identyfikacji twarzy na konkurs ICB-RW 2016

Celem konkursu International Challenge on Biometric Recognition in the Wild 2016 [72] było zaproponowanie metody identyfikacji twarzy na podstawie zdjęć z monitoringu wizyjnego (statyczne klatki z umieszczonej na zewnątrz kamery) w bazie utworzonej z obrazów twarzy w trzech pozach (portret frontalny i dwa profilowe). Dane, które służyły do ewaluacji zapropo-

nowanych metody są częścią bazy Quiscampi, która została wykorzystana w pracy doktorskiej po dostosowaniu do wymagań jako MV-Quiscampi (szerzej opisana w rozdziale1.7.1). Autorka pracy zaproponowała metodę opartą na trzech modelach dedykowanych do konkretnej pozy oraz agregacji poszczególnych deskryptorów wyjściowych. Poszczególne modele bazowały na architekturze VGG-16 [25] i były dotrenowywane na wybranym zestawie ujęć. Algorytm zajął czwarte miejsce i został opisany w artykule podsumowującym wyniki porównania [73].

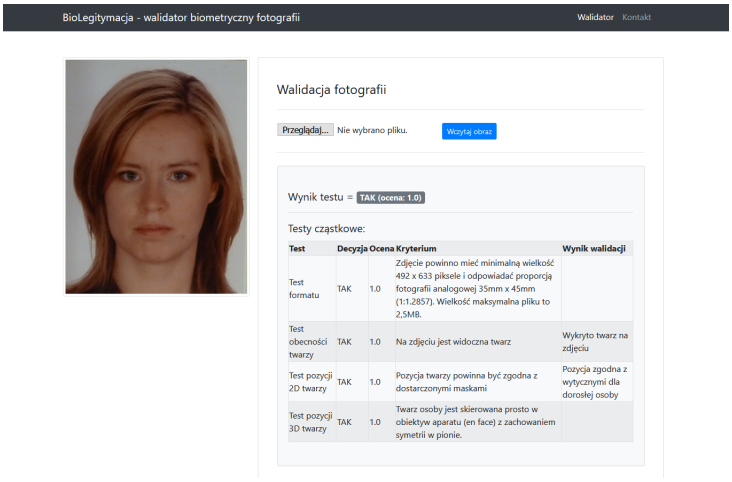


Rysunek 2: Koncepcja systemu opisanego w artykule *„Deep Neural Network Ensembles Dedicated to Different Head Poses for Face Identification"*. Zaproponowany system identyfikacji składał się z trzech modeli dedykowanych do pozy: frontalnej, lewego i prawego profilu. Wytworzone deskryptory były składane, żeby zwrócić identyfikator wyszukiwanej osoby.

A.3 BioWalidator

Projekt BioWalidator był projektem wewnętrznym prowadzonym w NASK-PIB (Naukowa i Akademicka Sieć Komputerowa) i dotyczył opracowania aplikacji, która daje możliwość sprawdzenia wymagań biometrycznych dla fotografii twarzy. Konieczność sprawdzenia jakości fotografii cyfrowej do dokumentów wynika z tego, że są one później wykorzystywane w automatycznych systemach weryfikacji tożsamości. Wyszukiwanie na podstawie obrazów o niskiej jakości lub próba ich rejestracji w bazie może znacząco obniżyć skuteczność całego systemu. Lista kryteriów została stworzona na podstawie rządowej instrukcji wykonywania zdjęć do dokumentów paszportowych oraz dowodów osobistych [11] i obejmowała wymagania takie jak obecność i położenie twarzy w obrazie, odpowiednia poza, jednolite tło oraz

wystarczająca jakość samego obrazu. Biblioteka walidująca został przygotowana w środowisku Python, a jej funkcje mogły zostać udostępnione jako mikroserwis. Docelowym sposobem użytkowania miało być sprawdzanie zdjęć osób ubiegających się o dowód osobisty poprzez formularz na stronie internetowej. Projekt był prowadzony od sierpnia 2019 do kwietnia 2020.



Rysunek 3: Zrzut ekranu z aplikacji BioWalidator. Aplikacja analizowała wgrane do niej zdjęcia - sprawdzała, czy na zdjęciu obecna jest twarz i czy spełnia wymagania dokumentów tożsamości. Projekt miał być rozwiązaniem problemów, które pojawiały się w systemie zdalnych wniosków o polski dowód tożsamości. Po wprowadzeniu obowiązku pobierania odcisków palców do dowodu nie ma już możliwości zdalnego ubiegania się o dowód, dlatego system został zamknięty.

Bibliografia

[1] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. ICCV*, 2015.

[2] A. Kanezaki, "RotationNet: Learning object classification using unsupervised viewpoint estimation," *CoRR*, vol. abs/1603.06208, 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *International Conference on Automatic Face and Gesture Recognition*, 2018.

[5] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus Benchmark - C: Face Dataset and Protocol," in *2018 International Conference on Biometrics (ICB)*, pp. 158–165, Feb 2018.

[6] W. Gutfeter and A. Pacut, "Face 3D biometrics goes mobile: Searching for applications of portable depth sensor in face recognition," in *2015 IEEE 2nd International Conference on Cybernetics (CYBCONF)*, pp. 489–494, June 2015.

[7] B. DeCann and A. Ross, "Relating ROC and CMC curves via the biometric menagerie," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–8, Sep. 2013.

[8] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–591, 1991.

[9] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[10] S. Banerjee, J. Brogan, J. Krizaj, A. Bharati, B. R. Webster, V. Struc, P. J. Flynn, and W. J. Scheirer, "To frontalize or not to frontalize: Do we really need elaborate pre-processing to improve face recognition?," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 20–29, March 2018.

[11] Ministerstwo Spraw Wewnętrznych i Administracji, "Instrukcja wykonywania zdjęć do dokumentów paszportowych oraz dowodów osobistych." http://www.gov.pl/web/gov/zdjecie-do-dowodu-lub-paszportu, May 2016.

[12] ISO/IEC 19795-1:2006(E), "Information technology — Biometric performance testing and reporting – Part 1: Principles and framework. ," standard, International Organization for Standardization, Geneva, CH, 2006.

[13] P. J. Grother and M. L. Ngan, "Face Recognition Vendor Test (FRVT) Performance of Face Identification Algorithms NIST IR 8009," tech. rep., NIST Interagency/Internal Report (NISTIR), May 2014.

[14] ISO/IEC 19794-5:2005(E), "Information technology – Biometric data interchange formats – Part 5: Face image data. Final Draft International Standard (FDIS)," standard, International Organization for Standardization, Geneva, CH, 2005.

[15] ICAO, "Biometrics Deployment of Machine Readable Travel Documents," standard, International Civil Aviation Organization, 2015.

[16] K. G. Policji, "Zarządzenie nr 64 Komendanta Głównego Policji z dnia 17 marca 2003 r. zmieniające zarządzenie w sprawie uzyskiwania, przetwarzania i wykorzystywania przez Policję informacji oraz sposobów zakładania i prowadzenia zbiorów tych informacji," in *Dziennik Urzędowy Komendy Głównej Policji*, vol. 9, (Warszawa), 2003.

[17] J. Neves, J. Moreno, and H. Proença, "QUIS-CAMPI: an annotated multi-biometrics data feed from surveillance scenarios," *IET Biometrics*, vol. 7, no. 4, pp. 371–379, 2018.

[18] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, pp. 4–12, April 2012.

[19] Occipital, "Sensor structure.io." https://structure.io/structure-sensor.

[20] Z. Wu, S. Song, A. Khosla, X. Tang, and J. Xiao, "3D ShapeNets for 2.5D object recognition and next-best-view prediction," *CoRR*, vol. abs/1406.5670, 2014.

[21] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *CoRR*, vol. abs/1411.1792, 2014.

[22] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, June 2014.

[23] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[24] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, IEEE Computer Society, September 2008.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[26] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," 2017.

[27] S. Albanie, "VGGFace2 - list of pretrained PyTorch models with weights." https://www.robots.ox.ac.uk/~albanie/research.html, 2017.

[28] N. I. of Standards and Technology, "NIST Face Challenges." https://www.nist.gov/programs-projects/face-challenges, 2015.

[29] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[30] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1939, 2015.

[31] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother, "IARPA Janus Benchmark-B Face Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 592–600, 2017.

[32] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning. Systemy uczące się.* Warszawa: Wydawnictwo Naukowe PWN, 2018.

[33] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.), vol. 28(3) of *Proceedings of Machine Learning Research*, (Atlanta, Georgia, USA), pp. 1139–1147, PMLR, 17–19 Jun 2013.

[34] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. 61, pp. 2121–2159, 2011.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.

[36] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," *CoRR*, vol. abs/1503.03832, 2015.

[37] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, (Washington, DC, USA), pp. 539–546, IEEE Computer Society, 2005.

[38] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," *CoRR*, vol. abs/1404.4661, 2014.

[39] C.-H. Ho, P. Morgado, A. Persekian, and N. Vasconcelos, "PIEs: Pose invariant embeddings," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[40] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood Components Analysis," in *Advances in Neural Information Processing Systems 17* (L. K. Saul, Y. Weiss, and L. Bottou, eds.), pp. 513–520, MIT Press, 2005.

[41] W. Gutfeter and A. Pacut, "Proxy embeddings for face identification among multi-pose templates," in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2020, Volume 5: VISAPP, Valletta, Malta, February 27-29, 2020* (G. M. Farinella, P. Radeva, and J. Braz, eds.), pp. 513–520, SCITEPRESS, 2020.

[42] Papers with code, "Face Verification on IJB-C." https://paperswithcode.com/sota/face-verification-on-ijb-c, 2022.

[43] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, and J. Zhou, "WebFace260M: A benchmark unveiling the power of million-scale deep face recognition," *CoRR*, vol. abs/2103.04098, 2021.

[44] X. An, X. Zhu, Y. Xiao, L. Wu, M. Zhang, Y. Gao, B. Qin, D. Zhang, and Y. Fu, "Partial FC: Training 10 million identities on a single machine," *CoRR*, vol. abs/2010.05222, 2020.

[45] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "ElasticFace: Elastic margin loss for deep face recognition," *CoRR*, vol. abs/2109.09416, 2021.

[46] Y. Shi, A. K. Jain, and N. D. Kalka, "Probabilistic Face Embeddings," *CoRR*, vol. abs/1904.09658, 2019.

[47] PyTorch, "An open source machine learning framework that accelerates the path from research prototyping to production deployment." https://pytorch.org, 2016.

[48] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems." https://www.tensorflow.org/, 2015. Software available from tensorflow.org.

[49] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2." https://github.com/facebookresearch/detectron2, 2019.

[50] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[51] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[52] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[53] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition," *CoRR*, vol. abs/1607.08221, 2016.

[54] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6738–6746, 2017.

[55] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *CoRR*, vol. abs/1801.07698, 2018.

[56] A. Bronstein, M. Bronstein, and R. Kimmel, "Three-dimensional face recognition," *International Journal of Computer Vision*, vol. 64, pp. 5–30, 2005.

[57] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," *CoRR*, vol. abs/1411.7964, 2014.

[58] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *In Proceeding of International Conference on Computer Vision*, (Venice, Italy), October 2017.

[59] Y. Yin, S. Jiang, J. P. Robinson, and Y. Fu, "Dual-attention GAN for large-pose face frontalization," *CoRR*, vol. abs/2002.07227, 2020.

[60] S. Wu, C. Rupprecht, and A. Vedaldi, "Unsupervised learning of probably symmetric deformable 3D objects from images in the wild," *CoRR*, vol. abs/1911.11130, 2019.

[61] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, June 2014.

[62] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," *CoRR*, vol. abs/1603.05474, 2016.

[63] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. G. Medioni, "Do we really need to collect millions of faces for effective face recognition?," *CoRR*, vol. abs/1603.07057, 2016.

[64] Q. Yu, C. Yang, H. Fan, and H. Wei, "Latent-MVCNN: 3D shape recognition using multiple views from pre-defined or random viewpoints," *Neural Processing Letters*, vol. 52, 08 2020.

[65] Z. Li, H. Wang, and J. Li, "Auto-MVCNN: Neural architecture search for multi-view 3D shape recognition," *CoRR*, vol. abs/2012.05493, 2020.

[66] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[67] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[68] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *CoRR*, vol. abs/2012.12877, 2020.

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[70] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[71] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[72] J. Neves and H. Proença, "ICB-RW 2016: International challenge on biometric recognition in the wild," in *2016 International Conference on Biometrics (ICB)*, pp. 1–6, June 2016.

[73] H. Proença, M. Nixon, M. Nappi, E. Ghaleb, G. Özbulak, H. Gao, H. K. Ekenel, K. Grm, V. Struc, H. Shi, X. Zhu, S. Liao, Z. Lei, S. Z. Li, W. Gutfeter, A. Pacut, J. Brogan, W. J. Scheirer, E. Gonzalez-Sosa, R. Vera-Rodriguez, J. Fierrez, J. Ortega-Garcia, D. Riccio, and L. De Maio, "Trends and controversies," *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 41–67, 2018.