# 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*) oraz model RotationNet. 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 i wytrenowany na bazie VGGFace2 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 and RotationNet. 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 architecture. Therefore weights can be transferred from the baseline single-view model which was trained on VGGFace2 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) 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