Jolanta CHMIELIŃSKA, Jacek JAKUBOWSKI

Wojskowa Akademia Techniczna, Wydział Elektroniki

doi:10.15199/48.2019.11.39

Rozpoznawanie twarzy z wykorzystaniem technik głębokiego uczenia i fuzji danych obrazowych

Streszczenie. *W artykule przedstawiono wyniki oryginalnych badań nad zastosowaniem sieci neuronowej wykorzystującej techniki głębokiego uczenia w zadaniu identyfikacji tożsamości na podstawie obrazów twarzy zarejestrowanych w zakresie widzialnym i w podczerwieni. W badaniach użyte zostały obrazy twarzy eksponowanych w zmiennych ale kontrolowanych warunkach. Na podstawie uzyskanych wyników można stwierdzić, że oba badane zakresy spektralne dostarczają istotnych ale różnych informacji o tożsamości danej osoby, które się wzajemnie uzupełniają.*

Abstract*. The paper presents the results of the original research on the application of a neural network using deep learning techniques in the task of identity recognition on the basis of facial images acquired in both visual and thermal radiation ranges. In the research, the database containing images acquired in various but controlled conditions was used. On the basis of the obtained results it can be established that both investigated spectral ranges provide distinctive and complementary details about the identity of an examined person. (**Face recognition based on deep learning techniques and image fusion**).*

Słowa kluczowe: rozpoznawanie twarzy, sieci konwolucyjne, fuzja danych.
Keywords: face recognition, convolutional networks, data fusion.

Wprowadzenie

Biometria jest nauką zajmującą się matematyczno-statystycznymi badaniami zmienności populacji organizmów żywych oraz pomiarami ich mierzalnych cech, która znajduje szerokie zastosowanie w problematyce komputerowego rozpoznawania tożsamości. Oprócz oczekiwanej wysokiej wiarygodności działania systemu biometrycznego bardzo istotna jest akceptowalność sposobu pobrania danych do analizy. W wielu praktycznych zastosowaniach istnieje silna potrzeba realizacji koncepcji badania przesiewowego przeprowadzanego w trybie „on-line". Najlepszym materiałem do tego celu, ze względu na brak konieczności współdziałania podmiotu poddawanego identyfikacji i coraz lepsze uzyskiwane wiarygodności, jest zarejestrowany obraz twarzy. Dlatego też cieszy się on dużym zainteresowaniem wielu ośrodków badawczych zajmujących się opracowywaniem biometrycznych systemów identyfikacji i weryfikacji. Stosowane metody rozpoznawania twarzy można podzielić na dwie grupy: metody holistyczne, czyli oparte na analizie obrazu twarzy jako całości, oraz analityczne, bazujące na mierzalnych cechach ludzkiej twarzy. W ostatniej dekadzie rozwinęły się również metody będące podstawą podjętych w niniejszej pracy badań, które wykorzystują informacje zawarte w lokalnych cechach obrazów twarzy, ekstrahowanych za pomocą masek realizujących operację splotu (konwolucji) [1]. Zastosowane narzędzie obliczeniowe realizuje proces nienadzorowanego poszukiwania dystynktywnej informacji bezpośrednio w obrazie i znajduje zastosowanie np. w aplikacjach społecznościowych jak Facebook, którego algorytm DeepFace [2], nauczony na bazie 4 milionów fotografii profilowych użytkowników portalu, oferuje błąd rozpoznawania w zadaniu weryfikacji na poziomie 3%, a więc niższym od błędu człowieka.

Problem badawczy

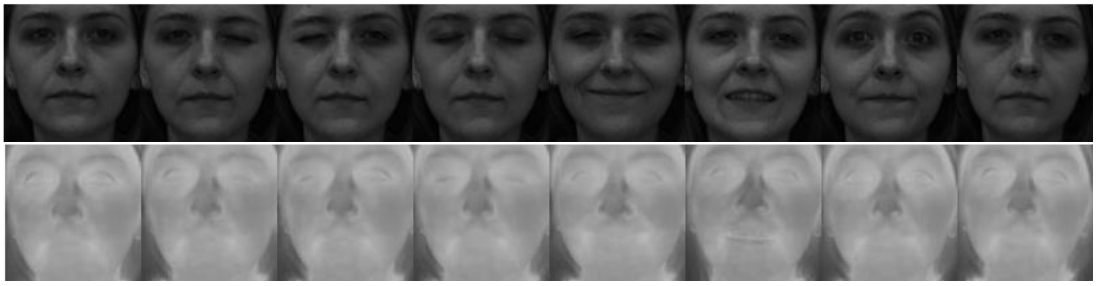
Identyfikacja osoby na podstawie twarzy jest dla ludzkiego mózgu zadaniem wręcz trywialnym (z wyjątkiem osób cierpiących na prozopagnozję), jednak automatyzacja maszynowa tego procesu jest zagadnieniem bardzo złożonym. Istniejące metody wykazują wysoką poprawność wyników przy przetwarzaniu obrazów zarejestrowanych w podobnych warunkach. Niestety, w praktycznych zastosowaniach utrzymanie stałych warunków podczas akwizycji obrazów wzorcowych i testowych jest niemożliwe, chociażby ze względu na upływ czasu, zmiany makijażu lub fryzury czy stan psychofizyczny osoby poddawanej analizie. Nawet zachowanie stałych warunków oświetleniowych, jak również jednakowych pozycji twarzy względem urządzenia rejestrującego obraz, może być trudne do realizacji, zwłaszcza w zadaniu rozpoznawania tożsamości w przypadku gdy proces identyfikacji ma przebiegać bez współpracy a może nawet wiedzy osoby badanej. Idealna metoda automatycznego rozpoznawania twarzy powinna być niezależna od wszystkich czynników, które mogą się zmienić od momentu rejestracji obrazów wzorcowych do chwili akwizycji obrazów testowych. W praktyce poszukuje się metod wykazujących dużą odporność na zmiany opisanych czynników lub części z nich, a mianowicie tych, które w konkretnym zastosowaniu nie mogą być kontrolowane. Przeprowadzone dotychczas badania wykazują, że wykorzystanie obrazu zarejestrowanego kamerą termowizyjną pozwala na znaczące ograniczenie negatywnego wpływu zmian oświetleniowych wywołanych tradycyjnymi sztucznymi źródłami światła na wynik procesu identyfikacji [3]. Obraz termowizyjny powstaje bowiem na podstawie promieniowania termalnego (pewnego jego zakresu) emitowanego przez każdy obiekt, którego temperatura jest większa od zera bezwzględnego (im cieplejszy obiekt tym krótsze fale emituje). Jednak, aby móc w pełni ocenić możliwość zastosowania tego typu obrazów jako danych wejściowych procedury rozpoznawania, konieczne jest uwzględnienie również innych czynników najczęściej występujących w trakcie akwizycji, takich jak np. wyrażane emocje badanej osoby. Nie bez znaczenia pozostaje również użyta metoda przetwarzania obrazów. Dostępne pozycje literaturowe wykazują bardzo wysoką wiarygodność systemów bazujących na rozwiązaniach z zakresu sieci konwolucyjnych w zastosowaniu do obrazów z zakresu światła widzialnego i podczerwieni [4][5][6][7][8][9]. Celem podjętej pracy stała się zatem ocena możliwości wykorzystania takich sieci w zadaniu rozpoznawania twarzy również na podstawie obrazów zintegrowanych, tj. będących efektem fuzji danych z obu zakresów spektralnych.

Materiał

Badania wykonane zostały z wykorzystaniem własnej bazy danych zawierającej obrazy twarzy 93 osób przedstawiające te same ujęcia zarejestrowane w zakresie światła widzialnego i podczerwieni przy zmianach warunków akwizycji. Rejestracje zrealizowano za pomocą autorskiego oprogramowania dedykowanego do obsługi kamer BASLER ace-1300gc oraz FLIR A35sc.

Tabela 1. Charakterystyka użytej bazy danych obrazowych

rodzaj czynnika	liczba osób	uwagi
kąt ustawienia kamer względem twarzy	59	zmiana kąta ustawienia kamery od -50° do 50° z krokiem 5° (21 obrazów); akwizycja w pomieszczeniu przy naturalnym świetle dziennym
kąt padania światła	51	zmiana kąta padania światła w zakresie od -90° do 90° z krokiem 30° (7 obrazów); wykorzystano dodatkowe, kierunkowe źródło światła żarowego o ustalonej intensywności
intensywność światła	31	wykorzystano dodatkowe źródło światła żarowego o frontalnym kącie padania (7 obrazów)
rodzaj światła	31	wykorzystano różne źródła światła (żarówkę żarową, świetlówkę kompaktową oraz żarówkę LED) o frontalnym kącie padania i ustalonej intensywności (3 obrazy)
mimika	53	zarejestrowano obrazy twarzy przy kontrolowanych zmianach mimiki (8 obrazów)

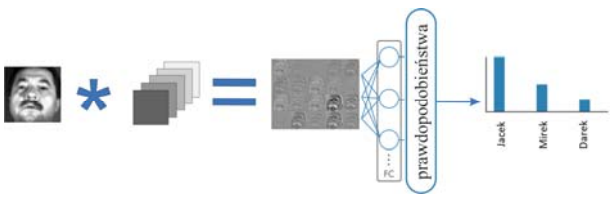


Rys. 1. Przykładowe obrazy z bazy danych zarejestrowane w obu zakresach spektralnych odpowiadające naturalnemu wyrazowi twarzy, zamknięciu oczu i symulowanym emocjom

Pierwsza z nich, pracująca w zakresie światła widzialnego, dysponowała możliwością przesyłania obrazów w skali szarości z 8-bitową rozdzielczością a druga jako kamera termalna (zakres spektralny 7,5μm÷13μm) transmitowała wyniki pomiaru temperatury z rozdzielczością 14-bitową. Wobec konieczności posługiwania się danymi obrazowymi, wymusiło to konieczność zamiany wartości temperatury na liczby reprezentujące poziomy w skali szarości. W Tabeli 1 została szczegółowo przedstawiona wykorzystana baza danych, w której znalazły się obrazy zarejestrowane z uwzględnieniem wpływu następujących czynników: kąta ustawienia kamery względem twarzy, kąta padania światła, intensywności oświetlenia, rodzaju źródła światła i wyrazu twarzy. Na rys. 1 przedstawiono przykładowe rejestracje uzyskane przy kontrolowanych zmianach mimiki.

Proponowana metoda

Konwolucyjne sieci neuronowe realizują proces nienadzorowanego poszukiwania dystynktywnej informacji bezpośrednio w obrazie z wykorzystaniem odpowiednio dużego zbioru danych. Charakteryzują się hierarchiczną strukturą wielowarstwową i wykorzystują mechanizmy tzw. głębokiego uczenia (ang. *Deep Learning*) w procesie rozkładu danych wejściowych na cechy. Podstawą jest operacja splotu (konwolucji), której efektem są maski filtrów wypracowane w procesie uczenia i traktowane jako nauczone cechy o charakterze lokalnym. Pierwsze warstwy wydobywają cechy proste, wspólne dla danych o bliskim sąsiedztwie. Następne warstwy wykorzystują je do wydobycia kolejnych, bardziej ogólnych własności danych służących w konsekwencji do klasyfikacji. Uproszczoną postać sieci konwolucyjnej przedstawia rys. 2.



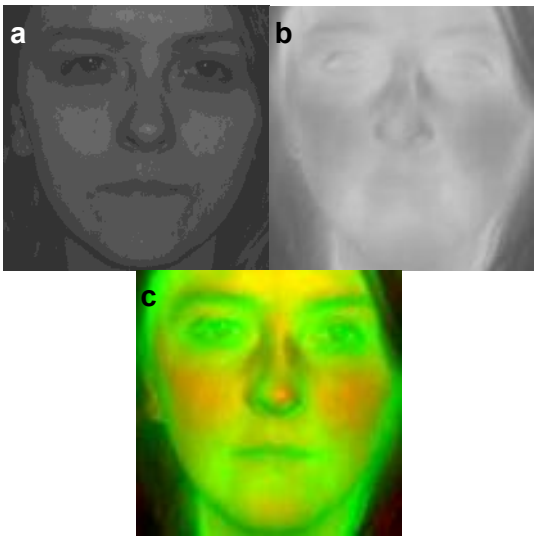
Rys. 2. Poglądowa ilustracja przetwarzania zachodzącego w strukturze neuronowej sieci konwolucyjnej

Standardowa struktura sieci konwolucyjnej, zwanej również w literaturze polskojęzycznej siecią splotową [3], składa się z następujących warstw:

- warstwy wejściowej, której wymiary determinują rozmiar przetwarzanych obrazów,
- jednej lub kilku warstw konwolucyjnych (ang. *convolutional layers*), które składają się z określonej liczby neuronów-filtrów o zadanych rozmiarach pól recepcyjnych wytwarzających tzw. mapy cech,
- jednej lub kilku warstw ReLU (ang. *Rectified Linear Unit*), które w rozwiązaniach głębokiego uczenia zastępują funkcje sigmoidalne; warstwy te eliminują wartości ujemne zwracane przez warstwę poprzedzającą poprzez zastąpienie ich zerami,
- jednej lub kilku warstw redukujących (ang. *pooling layers*), które realizują filtrację statystyczną w obrębie maski o zadanych rozmiarach wyznaczając wybraną statystykę (wartość maksymalną, minimalną lub maksymalną),
- jednej lub kilku warstw pełnego połączenia FC (ang. *fully-connected layer*), w których każdy neuron jest połączony ze wszystkimi wyjściami warstwy poprzedzającej, przy czym ostatnia warstwa pełnego połączenia posiada tyle neuronów ile klas ma rozpoznawać dana sieć,
- warstwy softmax, która wyznacza wartości prawdopodobieństwa przynależności obrazu wejściowego do poszczególnych klas i jest utożsamiana z funkcją aktywacji ostatniej warstwy pełnego połączenia.

Sieci takie można tworzyć samodzielnie dysponując dużym zbiorem danych obrazowych oraz wydajną techniką obliczeniową, ale można też dokonać adaptacji nauczonych już sieci do rozwiązania zupełnie innego problemu w ramach swoistego transferu wiedzy (ang. *Transfer Learning*) [10]. W pracy przyjęto strukturę sieci opracowanej od podstaw na potrzeby zdefiniowanego metodyką badań procesu rozpoznawania. Przyjęta metodyka badań polegała na pięciokrotnym losowym wyborze zadanej liczby osób, których obrazy były dzielone na zbiór uczący i testujący, przy czym zbiór uczący stanowił 70 % obrazów danej osoby. Następnie na wylosowanych danych z obu zakresów spektralnych oddzielnie przeprowadzane były procedury uczenia i testowania sieci o zadanej strukturze. Ponadto, jako dane wejściowe sieci wykorzystane zostały obrazy

powstałe w wyniku fuzji obrazów z obu zakresów spektralnych. W tym celu obrazy przedstawiające te same ujęcia twarzy zostały nałożone na siebie tworząc obraz RGB, w którym obraz z zakresu światła widzialnego stanowi składową R, obraz termalny składową G. Składowa B reprezentowana była przez macierz zerową. Na rys. 3 przedstawione zostały obrazy z obu zakresów spektralnych jednego ujęcia twarzy oraz wynik fuzji tych obrazów. Rozdzielczość przetwarzanych obrazów wynosiła 100×100 pikseli. Ponadto, przed podaniem na wejście sieci obrazy zostały poddane procedurze normalizacji za pomocą rozciągnięcia histogramu.



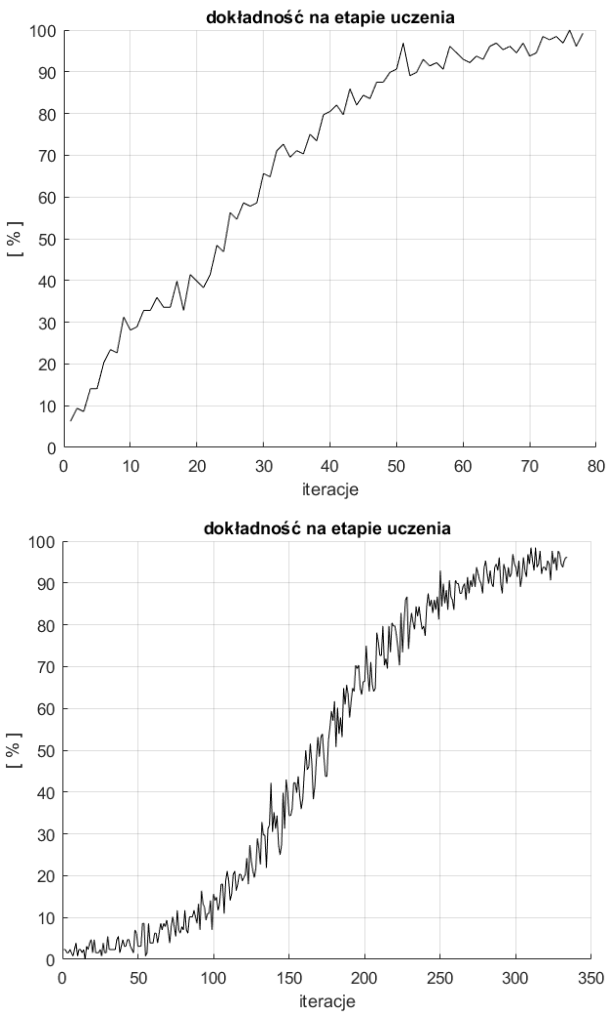
Rys. 3. Przykład obrazów z zakresu światła widzialnego (a), podczerwieni (b) przedstawiających to samo ujęcie twarzy oraz wynik ich fuzji (c).

Wybór struktury sieci przeprowadzony został w sposób eksperymentalny. Przedstawione w dalszej części opracowania wyniki otrzymano przy wykorzystaniu sieci składającej się z:

- warstwy wejściowej o wymiarze dopasowanym do wymiaru przetwarzanych obrazów,
- warstwy konwolucyjnej składającej się z 10 filtrów o wymiarach 9 x 9,
- warstwy ReLU,
- warstwy typu maxpool o rozmiarze maski 2 x 2 i kroku równym 2 w obu wymiarach,
- warstwy konwolucyjnej o strukturze takiej samej jak pierwsza warstwa konwolucyjna,
- drugiej warstwy ReLU,
- warstwy pełnego połączenia z 1000 wyjść,
- trzeciej warstwy ReLU,
- warstwy typu dropout do poprawy zdolności generalizacji sieci (losowe zerowanie pewnej liczby wejść),
- drugiej warstwy pełnego połączenia z liczbą wyjść równą liczbie rozpoznawanych osób,
- warstwy softmax.

Wyniki

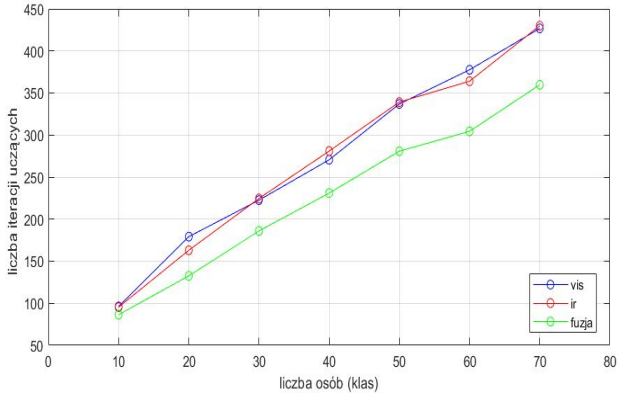
Procedura uczenia sieci polegała na doborze wartości wag sieci wg wybranego algorytmu uczącego. W każdej iteracji uczącej wagi sieci podlegają adaptacji na podstawie wyników rozpoznania uzyskanych dla pewnego podzbioru obrazów uczących. W przeprowadzonych badaniach podzbiór ten składał się ze 128 obrazów. W celu uniknięcia efektu przeuczenia sieci, proces uczenia był kończony w momencie, kiedy wartość średnia dokładności uczenia z 20 ostatnich iteracji uczących była większa niż 95%. Na rys. 4 przedstawione zostały przykładowe krzywe uczenia obrazujące zmianę dokładności rozpoznania uzyskanych dla obrazów prezentowanych w kolejnych iteracjach uczących.



Rys. 4. Krzywe uczenia sieci przy przetwarzaniu danych powstałych w wyniku fuzji obrazów z zakresu światła widzialnego i podczerwieni dla 10 klas (powyżej) i 70 klas (poniżej), czyli osób

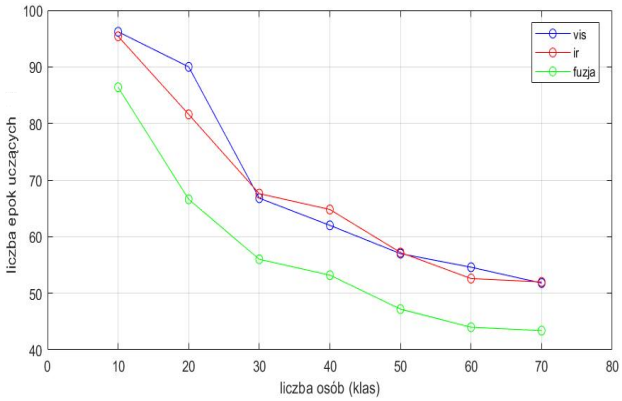
Można zauważyć, że wzrost liczby rozpoznawanych przez sieć osób wpływa na kształt krzywych uczenia i zwiększa liczbę iteracji potrzebnych do spełniania warunku zakończenia procesu uczenia.

Zgodnie z przyjętą metodyką, procedury uczenia i testowania sieci przeprowadzane były pięciokrotnie dla zadanej liczby losowo wybranych osób z całej bazy danych obrazowych, co umożliwiło wyznaczenie wartości statystycznych otrzymanych wyników. Na rys. 5 zamieszczony został wykres zależności średniej liczby iteracji potrzebnych do spełnienia warunku uczenia sieci w funkcji liczby rozpoznawanych osób. Można zauważyć, że bez względu na rodzaj przetwarzanego obrazu wzrost liczby klas zwiększa liczbę wykonanych iteracji uczących. Jest to związane ze zwiększeniem liczby obrazów w zbiorze uczącym, które muszą być poprawnie klasyfikowane przez sieć. Jednak przy wykorzystaniu obrazów będących fuzją obrazów z obu zakresów spektralnych proces uczenia sieci wymagał przeprowadzenia wyraźnie mniejszej liczby iteracji niż przy przetwarzaniu obrazów z jednego zakresu spektralnego.



Rys. 5. Wykres średniej liczby iteracji uczących w funkcji liczby rozpoznawanych osób (klas)

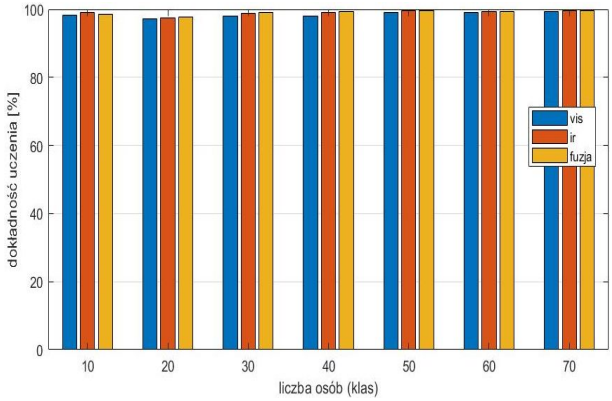
W procesie uczenia sieci oprócz iteracji wyróżnia się również epoki uczące. Każda epoka składa się z jednej lub wielu iteracji uczących w zależności od liczebności bazy danych. Jedna epoka obejmuje prezentację wszystkich obrazów uczących. Zatem im liczniejszy zbiór uczący, tym więcej iteracji przypada na jedną epokę uczącą. W przeprowadzonych eksperymentach numerycznych liczba obrazów uczących zwiększała się znacząco wskutek wzrostu liczby rozpoznawanych osób. Na rys. 6 zamieszczony został wykres średniej liczby epok uczących, potrzebnych do spełnienia warunku zakończenia procesu uczenia sieci, w funkcji liczby klas.



Rys. 6. Wykres średniej liczby epok uczących w funkcji liczby rozpoznawanych osób (klas)

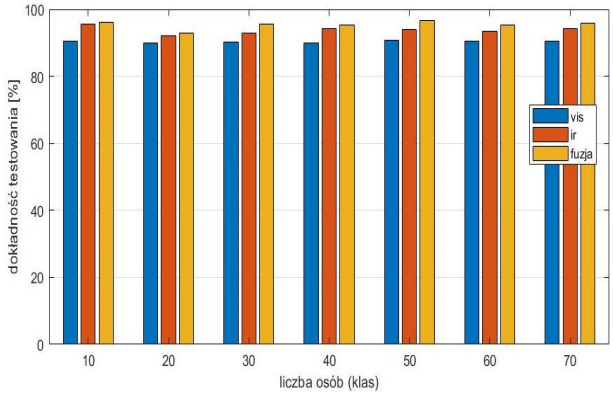
Można zauważyć, że wraz ze wzrostem liczby rozpoznawanych osób maleje liczba przeprowadzonych epok uczących. Warto przypomnieć, że proces uczenia był kończony po uzyskaniu średniej dokładności rozpoznania z 20 kolejnych iteracji uczących większej niż 95%. Otrzymane wyniki wskazują na zasadność zdefiniowania warunku zakończenia procesu uczenia sieci w ten właśnie sposób zamiast wykonywania z góry określonej, stałej liczby epok uczących. W ten sposób wyeliminowana została możliwość nie tylko „niedouczenia" ale również „przeuczenia" sieci, co również jest niekorzystne gdyż powoduje utratę generalizacji sieci, czyli poprawnej klasyfikacji obrazów ze zbioru testującego.

Wykresy zamieszczone na rys. 7 i 8 przedstawiają wartości dokładności wyznaczone dla zbiorów uczących oraz testujących uzyskane dla zmiennej liczby rozpoznawanych przez sieć osób przy przetwarzaniu różnego typu obrazów. Są to wartości średnie wyznaczone na podstawie 5 powtórzeń procesów uczenia i testowania sieci dla zadanej liczby losowo wybranych osób z całej bazy danych obrazowych.



Rys. 7. Wykres średnich dokładności uzyskanych w procesie uczenia sieci dla różnej liczby osób (klas)

Analizując wykres z rys. 7 można zauważyć, że wykorzystywany warunek zakończenia procesu uczenia pozwolił na względnie poprawne dopasowanie się sieci do danych uczących, które są klasyfikowane ze średnią dokładnością przekraczającą 97%. Jednak praktyczne zastosowanie sieci wymaga poprawnego rozpoznawania obrazów, które nie brały udziału w procesie uczenia. Do sprawdzenia generalizacji sieci służą obrazy ze zbioru testującego. Na podstawie wykresu zamieszczonego na rys. 8 można zauważyć, że uzyskane średnie dokładności rozpoznawania obrazów testujących również mają względnie duże wartości – nie mniejsze niż 90%, które nie zależą od liczby rozpoznawanych osób.



Rys. 8. Wykres średnich dokładności uzyskanych na etapie testowania sieci dla różnej liczby osób (klas)

Na podstawie zestawionych danych można stwierdzić, że rodzaj przetwarzanego obrazu ma wpływ na dokładność testowania. Zastosowanie obrazów termalnych daje lepsze wyniki niż przetwarzanie obrazów z zakresu światła widzialnego. Natomiast najwyższe dokładności testowania uzyskano dla obrazów powstałych w wyniku fuzji obu zakresów spektralnych.

Podsumowanie

Uzyskane wartości dokładności testowania wskazują, że przy wykorzystaniu bazy danych obrazowych zarejestrowanej w zmiennych warunkach akwizycji przetwarzanie obrazów termalnych daje lepsze wyniki identyfikacji tożsamości za pomocą konwolucyjnej sieci neuronowej niż w przypadku analizy obrazów z zakresu światła widzialnego. Warto przypomnieć, że w wykorzystywanej bazie danych znajdowały się obrazy zarejestrowane m.in. przy kontrolowanych zmianach kąta padania światła, natężenia oświetlenia rodzaju źródła

światła, które w znaczącym stopniu modyfikują obraz z zakresu światła widzialnego. Fakt ten może być powodem otrzymania lepszych wyników dla obrazu termalnego. Warto jednak podkreślić, że przetwarzanie obrazów będących wynikiem fuzji obrazów z obu zakresów spektralnych pozwoliło uzyskać lepsze dokładności identyfikacji niż w przypadku analizy obrazów z jednego zakresu. Można dzięki temu wnioskować, że oba rejestrowane zakresy spektralne dostarczają istotnych ale różnych informacji o tożsamości danej osoby, które się wzajemnie uzupełniają.

***Autorzy**: dr hab. inż. Jacek Jakubowski, Wojskowa Akademia Techniczna, Wydział Elektroniki, ul. Kaliskiego 2, 00908 Warszawa, E-mail: jacek.jakubowski@wat.edu.pl;*
mgr inż. Jolanta Chmielińska, Wojskowa Akademia Techniczna, Wydział Elektroniki, ul. Kaliskiego 2, 00908 Warszawa, E-mail: jolanta.pacan@wat.edu.pl.

LITERATURA

[1] Bengio Y., Courville A., and Goodfellow I., Deep Learning – systemy uczące się, *Wydawnictwo Naukowe PWN SA*, Warszawa, (2018).
[2] Taigman Y., Yang M., Ranzato M. A., Wolf L., DeepFace: Closing the Gap to Human-Level Performance in Face Verification, *2014 IEEE Conference on Computer Vision and Pattern Recognition*, materiały konferencyjne (2018), 1701-1708.
[3] Chmielińska J., Jakubowski J., Ocena powtarzalności punktów kluczowych obrazów twarzy z zakresu światła widzialnego i podczerwieni, *Przegląd Elektrotechniczny*, R. 90, nr 8, (2014), 205-208.
[4] Sun Y., Liang D., Wang X., Tang X., DeepID3: face recognition with very deep neural networks, *arXiv:1502.00873*, (2015).
[5] Schroff F., Kalenichenko D., Philbin J., FaceNet: a unified embedding for face recognition and clustering, *The IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 815-823.
[6] Coskun M., Ucar A., Yildrim O., Face recognition based on convolutional neural network, *2017 International Conference on Modern Electrical and Energy Systems*, (2017), 376-379.
[7] Landry S., Tagne E., Tonye E., CNNSFR: A Convolutional Neural Network System for Face Detection and Recognition, *International Journal of Advanced Computer Science and Applications*, 9 (12), (2018), 240-244.
[8] Wu Z., Peng M., Cheng T., Thermal face recognition using convolutional neural network, *2016 International Conference on Optoelectronics and Image Processing*, materiały konferencyjne (2016), 6-9.
[9] Kakkirala K., Chalamala S., Jami S., Thermal Infrared Face Recognition: A review, *19th International Conference on Modelling & Simulation*, materiały konferencyjne (2017), 55-60.
[10] Shao L., Zhu F., Transfer Learning for Visual Categorization: A Survey, *IEEE Transactions on Neural Networks and Learning Systems*, 26 (5), (2015), 1019-1034.