SFT VOL. 62 ISSUE 2, 2023, PP. 172–180

Robert Wolański[a)]*, Karol Jędrasiak[b)]

[a)] *School of Aspirants of the State Fire Service in Krakow / Szkoła Aspirantów Państwowej Straży Pożarnej w Krakowie*
[b)] *WSB University / Akademia WSB w Dąbrowie Górniczej*
* *Corresponding author / Autor korespondencyjny: rwolanski@sapsp.pl*

Audio-Video Analysis Method of Public Speaking Videos to Detect Deepfake Threat

Metoda analizy audio-wideo filmów z wystąpień publicznych w celu wykrycia zagrożenia typu *deepfake*

ABSTRACT

Aim: The purpose of the article is to present the hypothesis that the use of discrepancies in audiovisual materials can significantly increase the effectiveness of detecting various types of deepfake and related threats. In order to verify this hypothesis, the authors proposed a new method that reveals inconsistencies in both multiple modalities simultaneously and within individual modalities separately, enabling them to effectively distinguish between authentic and altered public speaking videos.

Project and methods: The proposed approach is to integrate audio and visual signals in a so-called fine-grained manner, and then carry out binary classification processes based on calculated adjustments to the classification results of each modality. The method has been tested using various network architectures, in particular Capsule networks – for deep anomaly detection and Swin Transformer – for image classification. Pre-processing included frame extraction and face detection using the MTCNN algorithm, as well as conversion of audio to mel spectrograms to better reflect human auditory perception. The proposed technique was tested on multimodal deepfake datasets, namely FakeAVCeleb and TMC, along with a custom dataset containing 4,700 recordings. The method has shown high performance in identifying deepfake threats in various test scenarios.

Results: The method proposed by the authors achieved better AUC and accuracy compared to other reference methods, confirming its effectiveness in the analysis of multimodal artefacts. The test results confirm that it is effective in detecting modified videos in a variety of test scenarios which can be considered an advance over existing deepfake detection techniques. The results highlight the adaptability of the method in various architectures of feature extraction networks.

Conclusions: The presented method of audiovisual deepfake detection uses fine inconsistencies of multimodal features to distinguish whether the material is authentic or synthetic. It is distinguished by its ability to point out inconsistencies in different types of deepfakes and, within each individual modality, can effectively distinguish authentic content from manipulated counterparts. The adaptability has been confirmed by the successful application of the method in various feature extraction network architectures. Moreover, its effectiveness has been proven in rigorous tests on two different audiovisual deepfake datasets.

Keywords: analysis of audio-video stream, detection of deepfake threats, analysis of public speeches
Type of article: original research article

Received: 29.11.2023; Reviewed: 03.12.2023; Accepted: 03.12.2023;
Authors` ORCID IDs: R. Wolański – 0000-0002-5625-0936; K. Jędrasiak – 0000-0002-2254-1030;
The authors contributed the equally to this article;
Please cite as: SFT Vol. 62 Issue 2, 2023, pp. 172–180, https://doi.org/10.12845/sft.62.2.2023.10;

ABSTRAKT

Cel: Celem artykułu jest przedstawienie hipotezy, że wykorzystanie rozbieżności w materiałach audiowizualnych może znacznie zwiększyć skuteczność wykrywania różnych typów *deepfake* i związanych z nimi zagrożeń. W celu weryfikacji tej hipotezy autorzy zaproponowali nową metodę, która pozwala na ujawnienie niespójności zarówno w wielu modalnościach jednocześnie, jak i w obrębie poszczególnych modalności z osobna, umożliwiając skuteczne rozróżnienie autentycznych i zmienionych filmów z wystąpieniami publicznymi.

Projekt i metody: Zaproponowane podejście polega na integracji sygnałów dźwiękowych i wizualnych w tzw. drobnoziarnisty sposób, a następnie przeprowadzeniu procesów klasyfikacji binarnej na podstawie obliczonych korekt wyników klasyfikacji każdej modalności. Metoda została przebadana z wykorzystaniem różnych architektur sieci, w szczególności sieci typu Capsule – do głębokiego wykrywania anomalii oraz Swin Transformer – do klasyfikacji obrazów. Przetwarzanie wstępne obejmowało ekstrakcję klatek i wykrywanie twarzy przy użyciu algorytmu MTCNN, a także konwersję audio na spektrogramy mel, aby lepiej odzwierciedlić ludzką percepcję słuchową. Zaproponowana technika została przetestowana na multimodalnych zbiorach danych *deepfake*,

a mianowicie FakeAVCeleb i TMC, wraz z niestandardowym zbiorem zawierającym 4700 nagrań. Metoda wykazała wysoką skuteczność w rozpoznawaniu zagrożeń *deepfake* w różnych scenariuszach testowych.
Wyniki: Metoda zaproponowana przez autorów osiągnęła lepsze AUC i dokładność w porównaniu z innymi metodami referencyjnymi, potwierdzając swoją skuteczność w analizie artefaktów multimodalnych. Rezultaty badań potwierdzają, że skutecznie pozwala wykryć zmodyfikowane filmy w różnych scenariuszach testowych – co można uznać za postęp w porównaniu z istniejącymi technikami wykrywania *deepfake'ów*. Wyniki podkreślają zdolność adaptacji metody w różnych architekturach sieci ekstrakcji cech.
Wnioski: Przedstawiona metoda audiowizualnego wykrywania *deepfake'ów* wykorzystuje drobne niespójności cech wielomodalnych do rozróżniania, czy materiał jest autentyczny czy syntetyczny. Wyróżnia się ona zdolnością do wskazywania niespójności w różnych typach *deepfake'ów* i w ramach każdej indywidualnej modalności potrafi skutecznie odróżniać autentyczne treści od zmanipulowanych odpowiedników. Możliwość adaptacji została potwierdzona przez udane zastosowanie omawianej metody w różnych architekturach sieci ekstrakcji cech. Ponadto jej skuteczność została udowodniona w rygorystycznych testach na dwóch różnych audiowizualnych zbiorach danych typu *deepfake*.
Słowa kluczowe: analiza strumienia audio-wideo, wykrywanie zagrożeń typu *deepfake*, analiza wystąpień publicznych
Typ artykułu: oryginalny artykuł naukowy

Przyjęty: 29.11.2023; Zrecenzowany: 03.12.2023; Zaakceptowany: 03.12.2023;
Identyfikatory ORCID autorów: R. Wolański – 0000-0002-5625-0936; K. Jędrasiak – 0000-0002-2254-1030;
Autorzy wnieśli równy wkład merytoryczny w powstanie artykułu;
Proszę cytować: SFT Vol. 62 Issue 2, 2023, pp. 172–180, https://doi.org/10.12845/sft.62.2.2023.10;

Introduction

Video content has traditionally been seen as irrefutable proof of reality, being a reliable confirmation of events. However, the development of advanced video manipulation methods has disrupted this state of affairs. Due to the development of deepfake technology and its expansive spread through the Internet and social media, the credibility of video content is now in doubt [1]. High-profile cases of deepfake use, such as the false statement by the Belgian prime minister linking COVID-19 to the climate crisis [2], or a fraudulent video conference by Russian pranksters in which politicians from the UK, Ukraine and the Baltics thought they were having online video chats with Leonid Volkov, Alexei Navalny's chief of staff [3], speak volumes about the implications of using this technology. The impact is very serious, as it affects the image of well-known, often influential people, such as politicians.

In response to the emerging threats, deepfake detection methods have been developed – to counter the increasingly sophisticated techniques used by their creators. Deepfake are digital files that are created through manipulation and fabrication of audiovisual content. They are most often created using artificial intelligence algorithms [4], such as generative adversarial networks (GANs), or autoencoders. However, diffusion models and various machine learning algorithms are also used to create convincing deepfakes. The ease with which synthetic films can now be created, especially those that superimpose one person's face over another, raises serious concerns. While digital content fusion technologies [5] have legitimate applications in entertainment, multimedia or education, their potential for abuse in activities such as financial fraud is alarming. This is demonstrated by incidents such as the successful attack on a bank in the United Arab Emirates using AI-synthesized speech [6].

The proliferation of deepfakes containing disinformation poses a serious threat, and cases such as the fake video of Ukrainian President Volodymyr Zelensky [7] illustrate the chaos

Wprowadzenie

Treści wideo tradycyjnie były postrzegane jako niezbity dowód rzeczywistości, będąc wiarygodnym potwierdzeniem zdarzeń. Jednak rozwój zaawansowanych metod manipulacji wideo zaburzył ten stan rzeczy. Z powodu rozwoju technologii *deepfake* i jej ekspansywnego rozprzestrzeniania poprzez internet oraz media społecznościowe, wiarygodność treści wideo jest obecnie przedmiotem wątpliwości [1]. Głośne przypadki wykorzystania *deepfake*, takie jak fałszywe oświadczenie premiera Belgii łączące COVID-19 z kryzysem klimatycznym [2], czy oszukańcza wideokonferencja rosyjskich prankstерów, podczas której politycy z Wielkiej Brytanii, Ukrainy i krajów bałtyckich sądzili, że prowadzili internetowe wideorozmowy z Leonidem Wołkowem, szefem sztabu Aleksieja Nawalnego [3], mówią wiele o konsekwencjach użycia tej technologii. Skutki są bardzo poważne, ponieważ dotyczą wizerunku znanych, często wpływowych osób, np. polityków.

W odpowiedzi na pojawiające się zagrożenia powstały metody wykrywania *deepfake'ów* – mające przeciwdziałać coraz bardziej wyrafinowanym technikom stosowanym przez ich twórców. *Deepfake* to pliki cyfrowe, które powstają na drodze manipulacji i fabrykacji treści audiowizualnych. Najczęściej są tworzone z wykorzystaniem algorytmów sztucznej inteligencji [4], takich jak generatywne sieci przeciwstawne (GAN), czy autoenkodery. Jednakże w celu tworzenia przekonujących fałszerstw typu *deepfake* stosuje się również modele dyfuzji oraz różne algorytmy uczenia maszynowego. Łatwość, z jaką można obecnie tworzyć syntetyczne filmy, zwłaszcza te, które nakładają twarz jednej osoby na drugą, budzi poważne obawy. Chociaż technologie syntezy treści cyfrowych [5] mają uzasadnione zastosowania w rozrywce, multimediach czy edukacji, to ich potencjał do nadużyć w działaniach, takich jak oszustwa finansowe, jest alarmujący. Świadczą o tym incydenty, np. skuteczny atak na bank w Zjednoczonych Emiratach Arabskich z wykorzystaniem wypowiedzi zsyntezowanej przez sztuczną inteligencję [6].

they can cause. As a result, deepfake detection has become an essential area of research in the discipline of security engineering. Currently, this research focuses primarily on binary classification to distinguish true content from false one [1]. Traditionally, efforts in this area have focused on single modalities, usually visual or sound artefacts in films. However, as deepfakes evolve to include multimodal fraud – both audio and video – detection mechanisms must also adapt to deal with these increasingly high-tech forgeries.

Current methods of unmasking deepfakes are mainly based on image, video stream or audio analysis. While the multi-modal approach holds promise, it is rarely used in practice, as the results of past attempts to fuse multi-modal signals for deepfake threat recognition have yielded comparable or inferior results to methods that analyse a single modality. In addition, feature fusion techniques often treat different types of deepfake manipulations as homogeneous, potentially disrupting the learning process. Nowadays, in order to create deepfakes, not only easy-to-detect image manipulation techniques like head pasting or mouth shape modification are used anymore. Increasingly, we are dealing with the use of methods developed for professional film dubbing or whole picture generation [8]. The authors of the article posed a research hypothesis that the use of inconsistencies in audiovisual artifacts of various types of deepfake, together with the analysis of all available modalities, will contribute to increasing the effectiveness of recognition of threats of this type. This article presents a method for detecting visual-sound artefacts in four categories of video authenticity. The proposed solution owes its effectiveness to the fusion of audiovisual features, teaching the algorithm in each modality independently and then integrating these results. The results of the conducted tests of the developed method using the available multi-modal datasets showed the worthwhile effectiveness of the proposed method regardless of the test scenario.

Analysis of the existing solutions

In the escalating battle against deepfake threats, there are two distinct strategies for detecting them: generic methods and specific methods. Approaches that are independent of the identity of the person in the video rely on detecting manipulation through learned visual artefacts or statistical anomalies using methods such as convolutional neural networks (CNNs) [9, 10]. These techniques originally proved effective in detecting the first wave of deepfake threats, characterized by clear artefacts or traces of manipulation. Unfortunately, these methods often fail in the face of modern deepfake threats, characterized by manipulation

Rozprzestrzenianie się *deepfake'ów* zawierających dezinformację stanowi poważne zagrożenie, a przypadki, takie jak fałszywe wideo prezydenta Ukrainy Wołodymyra Zełenskiego [7], ilustrują chaos, jaki mogą one wywołać. W rezultacie wykrywanie *deepfake'ów* stało się niezbędnym obszarem badań w dyscyplinie inżynieria bezpieczeństwa. Obecnie badania te koncentrują się przede wszystkim na klasyfikacji binarnej w celu odróżnienia treści prawdziwych od fałszywych [1]. Tradycyjnie wysiłki w tym zakresie koncentrowały się na pojedynczych modalnościach, zazwyczaj artefaktach wizualnych lub dźwiękowych w filmach. Jednak w miarę jak *deepfake'i* ewoluują, obejmując multimodalne oszustwa – zarówno audio, jak i wideo, dostosowywać się do radzenia sobie z tymi coraz bardziej zaawansowanymi technologicznie fałszerstwami muszą również mechanizmy ich wykrywania.

Obecnie stosowane metody demaskowania *deepfake'ów* bazują głównie na analizie obrazu, strumienia wideo lub dźwięku. Podejście wielomodalne jest obiecujące, w praktyce natomiast jest rzadko stosowane, gdyż rezultaty dotychczasowych prób fuzji wielomodalnych sygnałów na potrzeby rozpoznawania zagrożeń typu *deepfake* uzyskiwały porównywalne lub gorsze rezultaty co metody analizujące pojedynczą modalność. Ponadto techniki fuzji cech często traktują różne typy manipulacji *deepfake* jako jednorodne, potencjalnie zakłócając proces uczenia się. Obecnie w celu stworzenia *deepfake'ów* stosuje się już nie tylko łatwe do wykrycia techniki manipulacji obrazem typu przeklejenie głowy, czy modyfikacja kształtu ust. Coraz częściej mamy do czynienia z wykorzystaniem metod opracowanych z myślą o profesjonalnym dubbingu filmów lub generacją całego obrazu [8]. Autorzy artykułu postawili hipotezę badawczą, iż wykorzystanie niespójności w artefaktach audiowizualnych różnego rodzaju *deepfake'ów* wraz z analizą wszystkich dostępnych modalności, przyczyni się do zwiększenia skuteczności rozpoznawania zagrożeń tego typu. Niniejszy artykuł przedstawia metodę wykrywania artefaktów wizualno-dźwiękowych w czterech kategoriach autentyczności wideo. Zaproponowane rozwiązanie zawdzięcza swoją skuteczność fuzji cech audiowizualnych, ucząc algorytm w każdej modalności niezależnie, a następnie integrując te wyniki. Rezultaty przeprowadzonych badań opracowanej metody z wykorzystaniem dostępnych wielomodalnych zbiorów danych wykazały wartą głębszej analizy skuteczność zaproponowanej metody niezależnie od scenariusza testowego.

Analiza istniejących rozwiązań

W nasilającej się walce z zagrożeniami typu *deepfake* można wyróżnić dwie odrębne strategie ich wykrywania: metody generyczne i metody specyficzne. Podejścia niezależne od tożsamości osoby na filmie polegają na wykrywaniu manipulacji poprzez wyuczone artefakty wizualne lub anomalie statystyczne przy użyciu metod, takich jak konwolucyjne sieci neuronowe (CNN) [9, 10]. Techniki te okazały się pierwotnie skuteczne w wykrywaniu pierwszej fali zagrożeń typu *deepfake*, charakteryzujących się wyraźnymi artefaktami lub śladami manipulacji. Niestety metody te często zawodzą w obliczu współczesnych zagrożeń typu *deepfake*,

methods that are invisible to the naked eye, or when video or audio quality is poor.

While analysis based on low-level video features has shown promise, it remains vulnerable to video stream quality degradation and sophisticated modification techniques. Analysis based on high-level semantics offers an alternative by targeting distinct anomalies in person-specific features such as eye blinking, head position, physiological signals and others. These higher-level signals can provide clues for authentication and usually allow greater generalization to new fakes. A relatively new research area is the issue of multimodal deepfake detection by combining audio and visual signals. While these approaches provide extensive feature datasets for detection [11], they often do not significantly outperform their unimodal counterparts. This paradox prompted the authors of the publication to conduct further research into more efficient use of multimodal features. The proposed method differs from traditional fusion methods by introducing a fine-grained approach that distinguishes specific inconsistencies in the two modalities of audio and video, rather than treating them as uniform.

Proposed solution

The article presents a method for analysing audio-video streams to improve the effectiveness of detecting deepfake manipulation in recordings of public speeches. The developed method is based on integrating audio and visual signals in a fine-grained manner, and then performing a binary deepfake classification process with calculated corrections based on the classification results in each modality separately. The proposed two-modal approach aims to exploit the inconsistencies of multimodal deepfakes, as well as individual artifacts introduced by manipulation or content generation in each modality independently. The beginning of the innovative method is pre-processing, followed by multi-modal feature extraction, and concluded by the adopted multi-task learning strategy.

The pre-processing procedure starts with the extraction of individual images from the input stream, adjusting the process according to the length of the video, thus ensuring standard temporal resolution at the different durations of the analysed recordings of public speeches. In the next step, the MTCNN algorithm [12] for face detection and pruning is introduced, which isolates face regions based on the detected landmarks.

The audio content of the recording is then analysed, which is typically extracted in WAV format, which is a raw representation of the audio. The sound is transformed into a mel spectrogram – a representation that better reflects human auditory perception through frequency mapping. On the mel scale, the perceived distances in height are the same. A frequency range of up to 8,000 Hz is standardized and a uniform duration of 4 seconds is set for all mel spectrograms, thus ensuring consistency across the entire data set.

The next step is to perform feature extraction. This is a particularly important part of the process, in which input data is translated into high-level features that are key to identifying

charakteryzujących się niewidoczną gołym okiem metodą manipulacji lub gdy jakość wideo lub audio jest niska.

Chociaż analiza oparta na cechach wizyjnych niskiego poziomu okazała się obiecująca, pozostaje podatna na degradację jakości strumienia wideo oraz wyrafinowane techniki modyfikacji. Analiza oparta na semantyce wysokiego poziomu oferuje alternatywę poprzez ukierunkowanie na wyraźne anomalie w cechach specyficznych dla danej osoby, takich jak mruganie oczami, pozycja głowy, sygnały fizjologiczne i inne. Te sygnały wyższego poziomu mogą dostarczyć wskazówek do uwierzytelniania i zazwyczaj pozwalają na większe uogólnienie na nowe podróbki. Relatywnie nowym obszarem badawczym jest zagadnienie wielomodalnego wykrywania *deepfake'ów* poprzez połączenie sygnałów dźwiękowych i wizualnych. Chociaż podejścia te dostarczają obszernych zbiorów danych cech do wykrywania [11], często nie przewyższają znacząco swoich jednomodalnych odpowiedników. Paradoks ten skłonił autorów publikacji do dalszych badań nad bardziej wydajnym wykorzystaniem cech multimodalnych. Zaproponowana metoda różni się od tradycyjnych metod fuzji poprzez wprowadzenie drobnoziarnistego podejścia, które rozróżnia określone niespójności w dwóch modalnościach audio i wideo, zamiast traktować je jako jednolite.

Proponowane rozwiązanie

Artykuł przedstawia metodę analizy strumienia audio-wideo mającą na celu zwiększenie skuteczności wykrywania manipulacji typu *deepfake* w nagraniach z wystąpień publicznych. Opracowana metoda opiera się na integracji sygnałów audio oraz wizualnych w sposób drobnoziarnisty, a następnie przeprowadzeniu binarnego procesu klasyfikacji *deepfake'ów* z uwzględnieniem obliczonych korekt na podstawie rezultatów klasyfikacji w każdej modalności osobno. Zaproponowane dwupłaszczyznowe podejście ma na celu wykorzystanie niespójności multimodalnych *deepfake'ów*, jak również indywidualnych artefaktów wprowadzanych na skutek manipulacji lub generacji treści w każdej modalności niezależnie. Początek innowacyjnej metody stanowi przetwarzanie wstępne, następnie przeprowadzana jest wielomodalna ekstrakcja cech, a kończy je przyjęta strategia uczenia wielozadaniowego.

Proces przetwarzania wstępnego zaczyna się od ekstrakcji poszczególnych obrazów z wejściowego strumienia, dostosowując proces do długości wideo, zapewniając w ten sposób standardową rozdzielczość czasową przy różnych czasach trwania analizowanych nagrań z wystąpień publicznych. W kolejnym kroku wprowadza się algorytm MTCNN [12] do wykrywania twarzy i przycinania, który izoluje regiony twarzy na podstawie wykrytych punktów orientacyjnych.

Następnie analizowana jest zawartość audio nagrania, która typowo jest wyodrębniana w formacie WAV, będącym surową reprezentacją audio. Dźwięk jest przekształcany w spektrogram mel – reprezentację, która lepiej odzwierciedla ludzką percepcję słuchową poprzez mapowanie częstotliwości. W skali mel postrzegane odległości w wysokości są jednakowe. Standaryzowany jest zakres częstotliwości do 8000 Hz i ustawiany jednolity czas trwania wynoszący 4 sekundy dla wszystkich spektrogramów mel, zapewniając w ten sposób spójność w całym zbiorze danych.

distinctive patterns in deepfakes. To do this, pre-processed visual and audio data are fed into deep neural networks, which autonomously learn and extract these relevant features. The proposed approach is independent of the neural network model, demonstrating flexibility for different network architectures. Selecting the optimal network architecture may be an area for further work. The experiments for this article used the Capsule type network architecture [13], known for its effectiveness in so-called deep anomaly detection, and the Swin Transformer architecture [14], known for its good results in image classification.

In the final step of the study, multitask learning was conducted. The adopted framework for the learning process was expressed by combining three loss functions and taking into account the complexity of fine-grained deep identification of fakes combined with binary classification for each modality separately. The total loss function used L_{total} was proposed as a composite of the binary cross entropy losses L_a and L_v, for audio and video modalities, respectively, and L_f, a four-class cross entropy loss that includes different types of deepfake. The audiovisual classification task aggregates the output of the video network in multiple frames to extract the overarching video features. Two variants of fusion are examined: features, combining visual and audio elements, and results, averaging the classification results from both networks. The combined computing units are then fed into a four-class classification module for video identification.

Kolejnym krokiem jest przeprowadzenie ekstrakcji cech. Jest to szczególnie ważny element procesu, w którym dane wejściowe są tłumaczone na cechy wysokiego poziomu, kluczowe z punktu widzenia identyfikacji charakterystycznych wzorców w *deepfake'ach*. W tym celu do głębokich sieci neuronowych wprowadzane są wstępnie przetworzone dane wizualne i dźwiękowe, które autonomicznie uczą się i wyodrębniają te istotne cechy. Zaproponowane podejście jest niezależne od modelu sieci neuronowej, wykazując elastyczność dla różnych architektur sieci. Dobór optymalnej architektury sieci może stanowić obszar dalszych prac. W eksperymentach na potrzeby niniejszego artykułu wykorzystano architekturę sieci typu Capsule [13], znaną ze swojej skuteczności w tzw. głębokim wykrywaniu anomalii oraz architekturę Swin Transformer [14], mającą dobre rezultat w klasyfikacji obrazów.

W ostatnim kroku badania przeprowadzono uczenie wielozadaniowe. Przyjęte ramy dla procesu uczenia się zostały wyrażone poprzez połączenie trzech funkcji strat oraz uwzględnienie złożoności drobnoziarnistej głębokiej identyfikacji podróbek w połączeniu z klasyfikacją binarną dla każdej modalności osobno. Wykorzystana całkowita funkcja strat L_{total} została zaproponowana jako złożenie binarnych strat entropii krzyżowej L_a i L_v, odpowiednio dla modalności audio i wideo oraz L_f, czteroklasowej straty entropii krzyżowej, która obejmuje różne typy *deepfake'ów*. W przypadku zadania klasyfikacji audiowizualnej agregowane są dane wyjściowe sieci wideo w wielu klatkach, tak aby wyodrębnić nadrzędne cechy wideo. Badaniu poddawane są dwa warianty fuzji: funkcji, łączącej elementy wizualne i dźwiękowe, oraz wyników, uśredniającej wyniki klasyfikacji z obu sieci. W ten sposób połączone jednostki obliczeniowe są następnie wprowadzane do czteroklasowego modułu klasyfikacyjnego w celu identyfikacji wideo.

Research results

Experimental assessment of the proposed method was carried out on multimodal deepfake datasets: the FakeAVCeleb [11], TMC [15] and an in-house collection. A total of more than 37,000 recordings have been accumulated in all three collections for research work. All three analysed collections covered a wide spectrum of actual public speaking recordings and provided a representative testing ground. The method's test scenarios included recordings representing people from different ethnic groups and genders (see Figure 1). From the FakeAVCeleb collection, 25,500 videos were used for analysis, including 570 real recordings of public speeches made available on the YouTube platform, evenly distributed among the following ethnic groups: Caucasian (Americans), Caucasian (Europeans), Black (Africans), South Asian (Indians) and East Asian (e.g. Chinese, Koreans and Japanese). The division between men and women was 50/50. The fake recordings were generated by the authors of the collection [11]. The videos varied in length and the manipulation techniques used: real audio-true video, fake audio-true video, real audio-fake video, fake audio-fake video. The second used collection [15] contained 6943 recordings divided by the applied manipulation techniques as follows: real recordings 36.92%, real audio-true

Wyniki badań

Ocena eksperymentalna zaproponowanej metody została przeprowadzona na multimodalnych zbiorach danych *deepfake*: FakeAVCeleb [11], TMC [15] oraz zbiorze własnym. Łącznie we wszystkich trzech zbiorach zgromadzono na potrzeby prac badawczych ponad 37 000 nagrań. Wszystkie trzy analizowane zbiory obejmowały szerokie spektrum rzeczywistych nagrań z wystąpień publicznych i stanowiły reprezentatywny poligon doświadczalny. Scenariusze testowe metody uwzględniały nagrania reprezentujące osoby z różnych grup etnicznych i płci (zob. ryc. 1). Ze zbioru FakeAVCeleb wykorzystano do analizy 25 500 filmów, w tym 570 prawdziwych nagrań z wystąpień publicznych udostępnionych na platformie YouTube, równomiernie rozdzielonych na następujące grupy etniczne: rasa kaukaska (Amerykanie), rasa kaukaska (Europejczycy), rasa czarna (Afrykanie), rasy Azji Południowej (Hindusi) i rasy Azji Wschodniej (np. Chińczycy, Koreańczycy i Japończycy). Podział pomiędzy mężczyzn i kobiety był w proporcji 50/50. Nagrania fałszywe zostały wygenerowane przez autorów zbioru [11]. Filmy różniły się długością i zastosowanymi technikami manipulacji: prawdziwe audio-prawdziwe wideo, fałszywe audio-prawdziwe wideo, prawdziwe audio-fałszywe wideo, fałszywe audio-fałszywe wideo. Drugi

video 10.80%, real video-fake audio 9.07%, fake video-true audio 22.97%, fake video-fake audio 20.24%. The self-collection contained 5,000 recordings, divided into 1,000 real recordings of public speeches and 4,000 manipulated recordings, equally for each type of the manipulation.

In order to evaluate the effectiveness of the proposed audio-visual deepfake detection method, taking into account different detection strategies using Capsule and Swin Transformer networks, it was compared with established deepfake detection techniques such as MesoInception-4 [16], EfficientNet [17] and FTCN [18], AVoiD-DF [19] and AV-Lip-Sync [20]. The results confirmed the effectiveness of the authors' proposed method for recognizing deepfake threats (see Table 1).

wykorzystany zbiór [15] zawierał 6943 nagrania podzielone ze względu na zastosowane techniki manipulacji w następujący sposób: nagrania prawdziwe 36,92%, nagrania typu prawdziwe audio-prawdziwe wideo 10,80%, prawdziwe wideo-fałszywe audio 9,07%, fałszywe wideo-prawdziwe audio 22,97%, fałszywe wideo-fałszywe audio 20,24%. Zbiór własny zawierał 5000 nagrań w podziale 1000 nagrań prawdziwych z wystąpień publicznych oraz 4000 nagrań zmanipulowanych, po równo dla każdego typu manipulacji.

W celu oceny skuteczności zaproponowanej metody wykrywania audiowizualnego deepfake'u, uwzględniającej różne strategie wykrywania przy użyciu sieci Capsule i Swin Transformer, dokonano jej porównania z uznanymi technikami głębokiego wykrywania podróbek, takimi jak MesoInception-4 [16], EfficientNet [17] i FTCN [18], AVoiD-DF [19] i AV-Lip-Sync [20]. Uzyskane rezultaty potwierdziły skuteczność proponowanej przez autorów metody rozpoznawania zagrożeń typu deepfake (zob. tabela 1).

Figure 1. Examples of frames from films, showing real and modified elements that are difficult to distinguish with the naked eye
Rycina 1. Przykłady kadrów z filmów, przedstawiające elementy rzeczywiste i zmodyfikowane, trudne do rozróżnienia gołym okiem
Source: Authors' test collections.
Źródło: Zbiory testowe autorów.

Table 1. Test results of analysed deepfake threat detection methods on recordings of public speeches
Tabela 1. Rezultaty testów analizowanych metod detekcji zagrożeń typu *deepfake* na nagraniach z wystąpień publicznych

Method / Metoda	AUC	ACC
MesoInception	73.25	73.42
FTCN	86.12	68.35
EfficientNet	82.37	75.80
AVoiD-DF	88.56	84.50
AV-Lip-Sync	84.32	93.00
Proposed method / Zaproponowana metoda	96.30	97.40

Source: Own elaboration.
Źródło: Opracowanie własne.

The proposed method, like the reference methods, was implemented and tested under identical conditions using the same data sets. A comparison of the results in terms of AUC (area under the ROC curve) and model accuracy shows that the proposed solution performs better than the other models, indicating its effectiveness in detecting multimodal artefacts. It is interesting to note that all methods performed better when trained on the FakeAVCeleb dataset, compared to the TMC dataset and the custom dataset. This may be due to the greater variety of recordings in FakeAVCeleb. TMC's collection contains mostly recordings by Asians, while its own collection contains recordings by Europeans, which may have affected the results.

Proponowana metoda, podobnie jak metody referencyjne, była realizowana i testowana w identycznych warunkach, używając tych samych zestawów danych. Porównanie wyników w zakresie AUC (obszar pod krzywą ROC) i dokładności modelu pokazuje, że proponowane rozwiązanie osiąga lepsze rezultaty niż pozostałe modele, co wskazuje na jego skuteczność w wykrywaniu multimodalnych artefaktów. Interesujące jest, że wszystkie metody osiągały lepsze wyniki, gdy były trenowane na zbiorze danych FakeAVCeleb, w porównaniu do zbioru TMC i zbioru własnego. Może to wynikać z większej różnorodności nagrań w FakeAVCeleb. Zbiór TMC zawiera głównie nagrania Azjatów, a zbiór własny – Europejczyków, co mogło wpływać na wyniki.

In order to evaluate the ability of the proposed method to generalize, tests of the method's performance were also conducted when the training process was carried out on one set and the tests on the other. Again, the author's method achieved the highest efficiency. The best result (see Table 1) was achieved with the diverse FakeAVCeleb dataset and testing on the TMC dataset. The final stage of testing verified the model's effectiveness against modifications, such as real videos with mismatched audio. For this purpose, cross-validation tests were performed on the TMC dataset. Most of the analysed fake videos were correctly identified by the proposed method. Moreover, it has been labelled as "real video fake audio", highlighting the method's ability to detect this type of inconsistency, commonly found in manipulated videos.

In conclusion, the proposed method demonstrated the effectiveness of detecting modified videos in various test scenarios. This represents an advance over existing deepfake detection techniques.

Conclusion

The authors of this article presented a method for audiovisual deepfake detection that takes advantage of minor inconsistencies in multimodal features to distinguish whether the material is authentic or synthetic. The proposed approach is distinguished by its ability to identify inconsistencies across different types of deepfakes and within each individual modality. It allows to effectively distinguish authentic content from manipulated counterparts. The adaptability of the presented method has been confirmed by its successful application to various feature extraction network architectures. Its effectiveness has also been confirmed through rigorous testing on two different audiovisual deepfake datasets.

As part of their future work, the authors plan to focus their efforts on developing an audio-video content analysis system based on the proposed method that can be widely used to protect against certain types of deepfake threats. In conclusion, the proposed method sets a sure reference point in detecting forgeries in public speeches, representing a first step toward a safer digital media landscape in which the authenticity of recordings can be verified with greater certainty.

Acknowledgement

The present work was co-financed as part of the implementation of the project entitled "Interdisciplinary research projects of WSB researchers".

W celu oceny zdolności zaproponowanej metody do generalizacji przeprowadzono również badania działania metody, gdy proces trenowania został przeprowadzony na jednym zbiorze, a testy na drugim. Również w tym przypadku autorska metoda osiągnęła najwyższą skuteczność. Najlepszy rezultat (zob. tabela 1) udało się uzyskano przy różnorodnym zbiorze danych FakeAVCeleb i testowaniu na zbiorze TMC. W ostatnim etapie testów zweryfikowano skuteczność modelu przeciwko modyfikacjom, takim jak prawdziwe filmy z niedopasowanym dźwiękiem. W tym celu przeprowadzono testy krzyżowe na zbiorze danych TMC. Większość przeanalizowanych fałszywych filmów została poprawnie zidentyfikowana przez zaproponowaną metodę. Ponadto uzyskała oznaczenie jako real video fake audio (prawdziwe wideo fałszywe audio), co podkreśla zdolność metody do wykrywania tego typu niespójności, powszechnie występujących w zmanipulowanych filmach.

Podsumowując, zaproponowana metoda wykazała skuteczność wykrywania zmodyfikowanych filmów w różnych scenariuszach testowych. Stanowi to postęp w stosunku do istniejących technik detekcji typu *deepfake*.

Podsumowanie

Autorzy niniejszego artykułu przedstawili metodę audiowizualnego wykrywania *deepfake'ów*, która wykorzystuje drobne niespójności cech wielomodalnych do rozróżniania, czy materiał jest autentyczny, czy syntetyczny. Zaproponowane podejście wyróżnia się zdolnością do wskazywania niespójności w różnych typach *deepfake'ów* i w ramach każdej indywidualnej modalności. Pozwala na skuteczne odróżnianie autentycznych treści od zmanipulowanych odpowiedników. Zdolność przedstawionej metody do adaptacji została potwierdzona przez jej udane zastosowanie w różnych architekturach sieci ekstrakcji cech. Jej skuteczność została także potwierdzona w drodze rygorystycznych testów na dwóch różnych audiowizualnych zbiorach danych typu *deepfake*.

W ramach dalszej pracy autorzy planują skupić wysiłki na rozwinięciu na bazie zaproponowanej metody systemu analizy treści audio-wideo, który będzie możliwy do powszechnego stosowania w celu ochrony przed określonymi typami zagrożeń *deepfake*. Podsumowując, zaproponowana metoda wyznacza pewny punkt odniesienia w wykrywaniu fałszerstw w wystąpieniach publicznych, stanowiąc pierwszy krok w kierunku bezpieczniejszego krajobrazu mediów cyfrowych, w którym autentyczność nagrań można zweryfikować z większą pewnością.

Podziękowanie

Niniejsza praca była współfinansowana w ramach realizacji projektu pt. „Interdyscyplinarne projekty badawcze pracowników naukowych WSB".

Literature / Literatura

[1] Nguyen T.T., Nguyen Q.V.H., Nguyen D.T., Nguyen D.T., Huynh-The T., Nahavandi S., Nguyen C. M., *Deep learning for deepfakes creation and detection: A survey*, „Computer Vision and Image Understanding" 2022, 223, 103525.

[2] https://brusselstimes.com/106320/xr-belgium-posts-deepfake-of-belgian-premier-linking-covid-19-with-climate-crisis [dostęp 10.09.2023]

[3] https://wiadomosci.onet.pl/swiat/politycy-padli-ofiara-technologii-deep-fake-pranksterzy-podszywali-sie-pod/16w1ep7 [dostęp: 04.12.2023].

[4] Wang X., Guo H., Hu S., Chang M.C., Lyu S., *Gan-generated faces detection: A survey and new perspectives*, „arXiv" 2022, 2202.07145.

[5] Cao Y., Li S., Liu Y., Yan Z., Dai Y., Yu P.S., Sun L. *A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt*, „arXiv" 2023, 2303.04226.

[6] https://noizz.pl/nauka-i-technologia/sztuczna-inteligencja-sklonowali-glos-dyrektora-banku-i-ukradli-miliony/mnwrnpk [dostęp: 04.12.2023].

[7] https://www.komputerswiat.pl/aktualnosci/wydarzenia/do-sieci-trafil-deepfake-z-prezydentem-zelenskim-w-falszywym-wideo-namawial-do/n40qel7, [dostęp: 04.12.2023].

[8] Xie T., Liao L., Bi C., Tang B., Yin X., Yang J., Ma, Z., *Towards realistic visual dubbing with heterogeneous sources*, Proceedings of the 29th ACM International Conference on Multimedia, 2021, 1739–1747.

[9] Amerini I., Galteri L., Caldelli R., Del Bimbo A., *Deepfake video detection through optical flow based cnn*, Proceedings of the IEEE/CVF international conference on computer vision workshops, 2019.

[10] Almutairi Z., Elgibreen H., *A review of modern audio deepfake detection methods: challenges and future directions*, „Algorithms" 2022, 15(5), 155.

[11] Khalid H., Tariq S., Kim M., Woo S.S., *FakeAVCeleb: A novel audio-video multimodal deepfake dataset*, „arXiv" 2021, 2108.05080.

[12] Zhang N., Luo J., Gao W., *Research on face detection technology based on MTCNN*, International Conference on Computer Network, Electronic and Automation (ICCNEA), 2020, 154–158.

[13] Patrick M.K., Adekoya A.F., Mighty A.A., Edward B.Y., *Capsule networks – a survey*, „Journal of King Saud University-computer and information sciences" 2022, 34(1), 1295–1310.

[14] Liang J., Cao J., Sun G., Zhang K., Van Gool L., Timofte R., *Swinir: Image restoration using swin transformer*, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 1833–1844.

[15] Chen W., Chua S.L.B., Winkler S., Ng S.K., *Trusted Media Challenge Dataset and User Study*, Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, 3873–3877.

[16] Afchar D., Nozick V., Yamagishi J., Echizen I., *Mesonet: a compact facial video forgery detection network*, In 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, 1–7.

[17] Koonce B., Koonce B., *EfficientNet. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, 2021, 109–123.

[18] Zheng Y., Bao J., Chen D., Zeng M., Wen F., *Exploring temporal coherence for more general video face forgery detection*, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 15044–15054.

[19] Yang W., Zhou X., Chen Z., Guo B., Ba Z., Xia Z., Ren K., AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake, „IEEE Transactions on Information Forensics and Security" 2023, 18, 2015–2029.

[20] Shahzad S.A., Hashmi A., Peng Y.T., Tsao Y., Wang H. M., *AV-Lip-Sync+: Leveraging AV-HuBERT to Exploit Multimodal Inconsistency for Video Deepfake Detection*, „arXiv" 2023, 2311.02733.

SEN. BRIG. ROBERT MARCIN WOLAŃSKI, PH.D. ENG. – employee of the School of Aspirants of the State Fire Service in Krakow, Department of the Training Centre for the Protection of Population and Cultural Property. He is a graduate of the AGH University of Science and Technology in Cracow, officer's studies at the Main School of Fire Service, postgraduate studies in the area of wheeled vehicle operation and road accident expertise. He defended his doctoral thesis on infrared and microwave thermal protection technologies and materials at the University of Science and Technology. He conducts scientific work in parallel with his teaching activities through projects and individual research. He focuses on safety engineering issues with a special emphasis on the safety of rescuers. He is the author of a number of publications and a reviewer of recognized publications. In innovation activities, he is co-author of the patent "Method

ST. BRYG. W ST. SP. DR INŻ. ROBERT MARCIN WOLAŃSKI – pracownik Szkoły Aspirantów Państwowej Straży Pożarnej w Krakowie, Wydziału Centrum Szkolenia Ochrony Ludności i Dóbr Kultury. Absolwent Akademii Górniczo-Hutniczej w Krakowie, studium oficerskiego Szkoły Głównej Służby Pożarniczej, studiów podyplomowych z zakresu eksploatacji pojazdów kołowych oraz ekspertyz wypadku drogowego. Obronił pracę doktorską z zakresu technologii i materiałów do produkcji ochron termicznych przed promieniowaniem podczerwonym i mikrofalowym w Akademii Górniczo-Hutniczej. Prowadzi równolegle z działalnością dydaktyczną prace naukowe w ramach projektów i badań indywidualnych. Koncentruje się na zagadnieniach inżynierii bezpieczeństwa ze szczególnym uwzględnieniem bezpieczeństwa ratowników. Jest autorem szeregu publikacji i recenzentem uznanych wydawnictw. W działalności

of manufacturing ceramic layers on fabric". He is the initiator of a number of conferences and seminars aimed at the presentation and exchange of scientific and technical ideas in the area of progressive designs, technologies and organizational solutions for reducing the risk of conducting rescue operations. Currently, as an employee of the Civil and Cultural Property Protection Training Centre at the SA PSP Krakow, he continues his activities of promoting, educating and developing initiatives in the area of cultural heritage protection.

KAROL JĘDRASIAK, PH.D. – academic teacher, didactician and manager, author of more than 81 scientific publications, including 3 scientific monographs with high citability. The author's scientific experience includes participation in 24 research and development projects, also as a manager. Active participant in 24 scientific conferences and symposia. Expert of the WSL2014-2020 ROP, member of the Steering Committee of the Game INN Sector Program and the Society for Image Processing. As a result of his previous work and cooperation with industry, he participated in the development of 27 claims of intellectual property rights (3 granted patents, 12 patent applications, 12 design registration rights). Specialist in computer vision, computer graphics, artificial intelligence tools, computer, database and sensor system development. Since 2008, he has held management positions in private companies. For many years he was CEO of VR Technology, a company developing algorithms in the area of data analysis, commercializing innovative solutions in virtual reality technology and simulation as well as coaching systems.

innowacyjnej jest współautorem patentu „Sposób wytwarzania ceramicznych warstw na tkaninie". Jest inicjatorem szeregu konferencji i seminariów ukierunkowanych na prezentację i wymianę myśli naukowo-technicznej w obszarze progresywnych konstrukcji, technologii i rozwiązań organizacyjnych w zakresie ograniczenia ryzyka prowadzenia działań ratowniczych. Obecnie jako pracownik Centrum Kształcenia Ochrony Ludności i Dóbr Kultury w SA PSP Kraków kontynuuje swoją działalność promowania, edukacji i rozwoju inicjatyw w zakresie ochrony dziedzictwa kulturowego.

DR KAROL JĘDRASIAK – nauczyciel akademicki, dydaktyk i menadżer, autor ponad 81 publikacji naukowych, w tym 3 monografii naukowych o wysokiej cytowalności. Doświadczenie naukowe autora obejmuje udział w 24 projektach badawczo-rozwojowych, w tym także w charakterze kierownika. Aktywny uczestnik 24 konferencji i sympozjów naukowych. Ekspert RPO WSL2014-2020, członek Komitetu Sterującego Programu Sektorowego Game INN oraz Towarzystwa Przetwarzania Obrazów. W rezultacie dotychczasowej pracy oraz współpracy z przemysłem uczestniczył w opracowaniu 27 zastrzeżeń prawa własności intelektualnej (3 przyznane patenty, 12 zgłoszeń. patentowych, 12 praw z rejestracji wzoru przemysłowego). Specjalista w zakresie wizji komputerowej, grafiki komputerowej, narzędzi sztucznej inteligencji, wytwarzania systemów informatycznych, bazodanowych i sensorycznych. Od 2008 roku zajmuje stanowiska kierownicze w przedsiębiorstwach prywatnych. Przez wiele lat był Prezesem Zarządu spółki VR Technology zajmującej się opracowywaniem algorytmów z zakresu analizy danych oraz komercjalizacją innowacyjnych rozwiązań z zakresu technologii wirtualnej rzeczywistości oraz systemów symulacyjnych i trenażerowych.