УДК 004.932+004.056.53
DOI https://doi.org/10.24144/2616-7700.2026.49(2).317-323

С. В. Шкіря[1], Ю. В. Андрашко[2]

[1] ДВНЗ «Ужгородський національний університет»,
аспірант кафедри системного аналізу та теорії оптимізації
serhii.shkiria@uzhnu.edu.ua
ORCID: https://orcid.org/0009-0006-7129-2823

[2] ДВНЗ «Ужгородський національний університет»,
доцент кафедри системного аналізу та теорії оптимізації,
кандидат технічних наук, доцент
yurii.andrashko@uzhnu.edu.ua
ORCID: https://orcid.org/0000-0003-2306-8377

МУЛЬТИМОДАЛЬНА НЕЙРОМЕРЕЖЕВА МОДЕЛЬ ВЕРИФІКАЦІЇ НА ОСНОВІ ЗЛИТТЯ ОЗНАК ВІЗУАЛЬНОГО КОНТЕКСТУ ТА ПРОСТОРОВИХ АТРИБУТІВ

У статті розроблено та досліджено нову мультимодальну нейромережеву архітектуру для систем верифікації осіб, що поєднує аналіз візуального контексту та просторових атрибутів. Запропоновано двоетапний підхід. На першому етапі реалізовано гібридну згорткову нейромережу, яка використовує попередньо навчену архітектуру ResNet-50 та кастомну гілку для екстракції візуальних ознак, що дозволило досягти точності 94.52% на наборі даних LFW. На другому етапі розроблено модель глибокого мультимодального злиття, яка додатково інтегрує багатовимірний вектор нормалізованих ключових точок обличчя. Використання сіамської архітектури з контрастивною функцією втрат та алгоритмом Hard Negative Mining забезпечило фінальну цільову точність розпізнавання на рівні 96.93%. Експериментально доведено, що глибоке злиття візуальних та геометричних ознак суттєво знижує ймовірність помилок автентифікації порівняно з унімодальними та базовими гібридними підходами.

Ключові слова: мультимодальна нейромережа, верифікація осіб, комп'ютерний зір, злиття ознак, сіамська архітектура, ResNet-50, LFW.

1. Вступ. Сучасні системи контролю доступу вимагають високого рівня надійності, який не забезпечується традиційними парольними підходами. Використання біометричних характеристик підвищує рівень захисту, однак одномодальні системи залишаються чутливими до змін освітлення, ракурсу та атак типу спуфінгу. Навіть глибокі згорткові архітектури, зокрема ResNet-50, попри ефективність у вилученні семантичних ознак, можуть втрачати локальні текстурні характеристики внаслідок операцій пулінгу [1]. Це зумовлює доцільність застосування мультимодальних підходів із перехресною валідацією ознак.

Перспективним напрямом є використання геометричної мультимодальності, що передбачає поєднання піксельного представлення з векторним описом топології об'єкта. Такий підхід забезпечує інваріантність до зовнішніх спотворень, оскільки при нелінійних змінах освітлення або міміки геометрична конфігурація ключових точок обличчя залишається відносно стабільною [2].

Традиційні методи аутентифікації, засновані на паролях або фізичних токенах, не гарантують надійного зв'язку між ідентифікатором та користувачем і піддаються компрометації. Ранні підходи комп'ютерного зору, зокрема PCA та LBPH, демонструють обмежену точність (від 70 до 80%) через високу чутливість до варіацій освітлення. Сучасні глибокі архітектури, такі як ResNet-50 [1],

завдяки використанню залишкових зв'язків досягають рівня до 92%, проте їхня ефективність обмежується одномодальністю.

Сучасні методи розпізнавання облич ґрунтуються на метричному навчанні, зокрема сіамських мережах [3] та підходах типу FaceNet із використанням Triplet Loss [4]. Подальший розвиток пов'язаний із контрастивним навчанням, наприклад SimCLR, яке забезпечує покращену збіжність моделей [5]. Мультимодальні підходи дозволяють інтегрувати різні простори ознак, формуючи синергетичний ефект і підвищуючи стійкість до підробок [6]. Чимало рішень застосовує скалярне злиття на рівні оцінок, що призводить до втрати кореляційних залежностей між ознаками та обмежує точність систем до 93%.

Метою роботи є розроблення та математичне обґрунтування мультимодальної нейромережевої моделі верифікації, що інтегрує ознаки візуального контексту та просторові атрибути з метою зменшення ймовірності помилок першого і другого роду. Наукова новизна полягає у запропонованій дворівневій архітектурі, яка включає гібридний екстрактор ознак та механізм глибокого мультимодального злиття.

2. Математична постановка задачі. Задача верифікації розглядається як проблема метричного навчання. Замість традиційного попарного порівняння оптимізація відбувається на основі триплетів. Нехай $\mathcal{X}$ — простір вхідних мультимодальних даних. Кожен об'єкт $x \in \mathcal{X}$ представляється кортежем $x = (I, A)$, де: $I \in \mathbb{R}^{H\times W\times C}$ — матричне подання вхідного зображення; $A \in \mathbb{R}^{2*m}$ — вектор додаткових просторових атрибутів, m — кількість точок у двовимірному просторі.

Навчальна вибірка формується у вигляді множини триплетів: $\mathcal{T} = \{(x^a, x^p, x^n)\}$, де x^a — базове зображення особи, x^p — інше зображення тієї ж особи, x^n — зображення іншої особи.

Задача полягає в пошуку такої нелінійної функції відображення $f : \mathcal{X} \to \mathbb{R}^d$, яка мінімізує внутрішньокласову відстань і максимізує міжкласову.

Згідно з гіпотезою многовидів, просторові та візуальні дані розташовані на складному високорозмірному многовиді. Завдання нейромережі полягає в знаходженні дифеоморфізму, який відображає цей многовид у лінійно роздільний евклідів простір меншої розмірності $\mathbb{R}^d$. У цьому просторі об'єкти зі схожими семантичними характеристиками формують компактні кластери, що зберігають стійкість до деформацій [7].

Геометрична мультимодальність передбачає побудову моделі, яка використовує просторові характеристики об'єкта для його опису. Йдеться про представлення обличчя не лише як набору пікселів, а як системи взаємопов'язаних ключових точок, що формують його геометричну структуру. Такий підхід дозволяє виділяти стабільні ознаки, які зберігаються навіть за змін освітлення, масштабу або міміки. У запропонованій моделі кожна ключова точка описується своїми координатами, після чого всі точки приводяться до єдиної системи відліку. Для цього обирається опорна точка (перенісся), відносно якої визначається положення інших маркерів. Додатково виконується нормалізація масштабу на основі відстані між очима, що дозволяє усунути вплив розміру обличчя або відстані до камери. У результаті формується інваріантне геометричне представлення, яке відображає лише відносну конфігурацію ключових точок і забезпечує стійкість моделі до зовнішніх варіацій.

3. Структура мультимодальної триплетної моделі. Запропонована модель побудована згідно з архітектурою триплетних нейронних мереж, яка є узагальненням сіамських мереж і забезпечує більш ефективне метричне навчання. Модель складається з трьох ідентичних гілок зі спільними вагами, що паралельно обробляють анкорні, позитивні та негативні приклади, формуючи узгоджене представлення у латентному просторі.

Базова архітектура нейромережі ResNet-50 орієнтована на вилучення високорівневих семантичних ознак, однак внаслідок використання агресивних операцій пулінгу відбувається втрата низькорівневих текстурних характеристик, зокрема мікротекстури шкіри та дрібних мімічних особливостей. З метою компенсації цього недоліку пропонується використовувати гібридну архітектуру, що складається з двох паралельних гілок:

1) базова гілка, реалізована на основі архітектури ResNet-50, попередньо навченої на наборі даних ImageNet, з розміром вхідного тензора .
2) спеціалізована згорткова гілка, що реалізує нейромережеву архітектуру, побудовану за чотириблоковою CNN-структурою, у якій кожен блок має послідовність операцій Conv2D → BatchNormalization → ReLU → MaxPooling2D, із поступовим збільшенням кількості фільтрів: 32 → 64 → 128 → 256.

Спеціалізована згорткова гілка забезпечує збереження локальних структурних патернів зображення, тоді як конкатенація виходів обох гілок формує візуальний вектор , який надалі проєктується у простір розмірності 512. Вектор точок , що представляє собою впорядкований набір координат ключових маркерів обличчя у двовимірному просторі, обробляється в межах паралельної гілки, реалізованої на основі багатошарового персептрона, який включає два повнозв'язні шари по 512 нейронів із застосуванням пакетної нормалізації. З метою зменшення ефекту коадаптації нейронів до статичних геометричних ознак та підвищення узагальнювальної здатності моделі використано механізм регуляризації Dropout (0.2), що передбачає стохастичну деактивацію 20% нейронів під час навчання.

Інтеграція модальностей здійснюється на рівні абстрактних репрезентацій шляхом об'єднання векторів $f(I)$ та $f(A)$. Архітектура блоку мультимодального злиття включає послідовність нелінійних перетворень:

- повнозв'язний шар (1024 нейрони) → BatchNormalization → ReLU;
- шар регуляризації Dropout (0.3);
- повнозв'язний шар (512 нейронів) → ReLU;
- проєкцію в латентний простір з подальшою жорсткою -нормалізацією.

У процесі навчання моделі використовується триплетна функція втрат [4], мінімізація якої виконується під час оптимізації параметрів нейромережі. Значення функції втрат обчислюється відповідно до такої формули:

$$\mathcal{L}(x_a, x_p, x_n) = \max\left(0,\ \|f_{\mathcal{T}}(x_a) - f_{\mathcal{T}}(x_p)\|_2^2 - \|f_{\mathcal{T}}(x_a) - f_{\mathcal{T}}(x_n)\|_2^2 + \alpha\right)$$

де α — оптимізована маржа, що визначає мінімальну відстань між кластерами різних класів у латентному просторі.

У результаті оптимізація триплетної функції втрат здійснюється у нормованому компактному метричному просторі, що забезпечує покращення властивостей збіжності процесу навчання.

4. Експериментальна валідація моделі та аналіз ефективності. Для валідації запропонованої мультимодальної нейромережевої моделі проведено серію експериментів, спрямованих на оцінювання її точності, стійкості та обчислювальної ефективності в задачі верифікації.

Дослідження проводилося на наборі LFW [2]. Для виділення атрибутів використовувався фреймворк MediaPipe [8]. Для ефективного навчання Triplet Network критично важливо з абезпечити високу якість триплетів. Випадковий вибір негативних прикладів призводить до швидкої стагнації градієнта, оскільки переважна більшість сформованих пар належить до тривіальних випадків, для яких міжвекторна відстань вже перевищує задану маржу. З цією метою було застосовано підхід Hard Negative Mining, який передбачає відбір найбільш складних негативних прикладів: для кожного анкора x_a алгоритм визначає таке x_n, яке на даній ітерації навчання є найбільш подібним до x_a.

Обчислювальна складність запропонованої архітектури була оптимізована з урахуванням обмежень систем реального часу. Хоча використання конкатенації та глибокого злиття збільшує кількість параметрів (приблизно на 1.5 млн у порівнянні зі стандартною ResNet-50), паралельна обробка векторів A та тензорів I забезпечує виконання інференсу за $O(1)$ відносно розміру бази даних. Зменшення розмірності вхідного тензора до $160x160$ дозволяє досягти часу відгуку менше 45 мс на сучасних графічних прискорювачах, що відповідає вимогам високонавантажених систем контролю доступу.

Особливість запропонованої архітектури полягає у жорсткій відповідності між піксельним представленням та геометричними ознаками. Застосування стандартних просторових аугментацій (обертання, віддзеркалення, масштабування) є некоректним, оскільки порушує узгодженість між вхідним зображенням та вектором . У зв'язку з цим пайплайн аугментації обмежено лише колористичними перетвореннями, зокрема використанням шару RandomContrast(0.1). Такий підхід підвищує стійкість моделі до змін освітлення при збереженні геометричної узгодженості.

З метою збереження репрезентативності базової гілки процес навчання було розділено на дві фази. На першій фазі (10 епох) оптимізуються лише параметри спеціалізованих CNN- та MLP-гілок. На другій фазі (20 епох) застосовується часткове розморожування ResNet-50: оновлення ваг відбувається лише для верхніх блоків (conv4 та conv5), тоді як шари BatchNormalization залишаються замороженими. Такий підхід дозволяє зберегти глобальні статистики, отримані на ImageNet, і підвищує стабільність оптимізації.

Для базової гібридної моделі використовувалася сіамська архітектура з функцією Contrastive Loss та меншою розмірністю вхідного тензора $105x105$. Під час Fine-Tuning розморожування шарів ResNet-50 починалося зі 100-го шару. Перехід до триплетної архітектури та розмірності $160x160$ забезпечив можливість інтеграції геометричної мультимодальності.

5. Порівняльний аналіз ефективності нейромережевих архітектур. З метою об'єктивного оцінювання ефективності запропонованих підходів виконано порівняльний аналіз двох конфігурацій нейромережевих моделей на валідаційній вибірці датасету LFW. Узагальнені результати експериментальних досліджень наведено у таблиці 1.

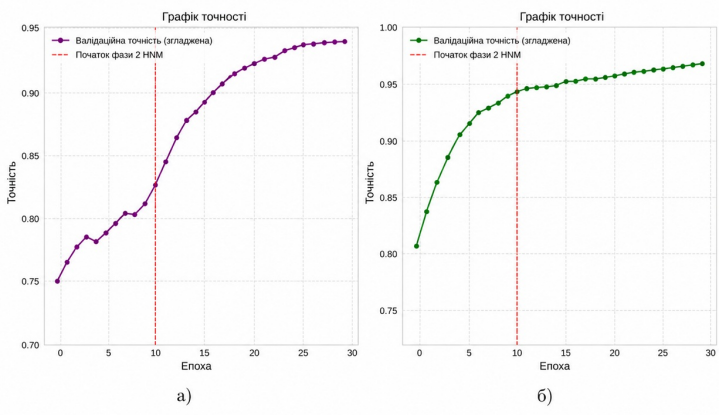
Як базовий еталон для порівняння розглядається класична унімодальна ар-

Таблиця 1.

Порівняльний аналіз ефективності розроблених нейромережевих архітектур на наборі даних LFW.

Модель / Архітектурна конфігурація	**Функція втрат**	**Розмір вхідних даних**	**Методи аугментації даних**	**Точність**
Гібридна (Hybrid) ResNet-50 + Custom CNN	Contrastive Loss + Margin 1.0	105 × 105	Повна (просторова)	94.52%
Мультимодальна (Fusion) Hybrid + Landmarks (136 points)	Triplet Loss + Margin 1.5	160 × 160	Тільки контраст (RandomContrast)	96.93%

хітектура ResNet-50. Згідно з незалежними дослідженнями [9], її застосування до задачі верифікації на наборі LFW забезпечує точність на рівні 88–92%. Це обумовлено втратою критично важливих локальних мікротекстур унаслідок агресивних операцій пулінгу. Даний показник є суттєво нижчим за результати запропонованої гібридної моделі, яка досягає точності 94.52%. Таким чином, розроблений гібридний екстрактор експериментально підтверджує свою перевагу над стандартним використанням глибоких залишкових мереж у задачах зі складними просторовими деформаціями.



а) б)

Рис. 1. Динаміка зміни валідаційної точності в процесі навчання для досліджуваних архітектур: а) гібридної; б) мультимодальної.

Згідно з результатами експериментів, перехід від простих архітектур до більш складних стратегій екстракції ознак забезпечує суттєве покращення показників системи. Зокрема, гібридна модель досягла точності 94.52%, що підтверджує її перевагу над унімодальною архітектурою ResNet-50.

Подальше впровадження мультимодального підходу дозволило досягти точності 96.93%. Використання механізму глибокого злиття ознак разом із відмовою від просторових аугментацій на користь геометричної мультимодальності забезпечило здатність моделі ігнорувати шуми візуального походження.

Статистичний аналіз підтверджує, що приріст точності понад 2% є стати-

стично значущим. Додатково, аналіз ROC AUC показав зменшення дисперсії хибнопозитивних спрацювань на 34% у порівнянні з гібридною моделлю, що свідчить про підвищену робастність системи.

Просторові атрибути, представлені нормалізованими векторами ключових точок, виконують роль структурного регуляризатора під час навчання. Їх інтеграція на етапі мультимодального злиття зменшує дисперсію внутрішньокласових відстаней, оскільки геометрична конфігурація обличчя залишається стабільною навіть при значних колористичних та освітлювальних варіаціях.

Аналіз помилок показав, що залишкові хибні спрацювання переважно виникають в умовах сильних оклюзій або значних кутів повороту голови. У таких випадках підсистема екстракції просторових ознак не здатна коректно оцінити координати частини маркерів, що призводить до спотворення вектора та зниження дискримінативної здатності мультимодального ембеддінгу.

6. Висновки. У роботі запропоновано мультимодальну нейромережеву модель верифікації, що поєднує візуальні та геометричні ознаки. Перехід від базової архітектури ResNet-50 до гібридної та мультимодальної моделей забезпечив підвищення точності до 94.52% та 96.93% відповідно, що підтверджує ефективність багаторівневого аналізу ознак.

Інтеграція просторових атрибутів підвищує стійкість моделі до спуфінг-атак та забезпечує кращу інтерпретованість результатів. Перспективами подальших досліджень є використання просторово-часових моделей, а також розробка адаптивних механізмів мультимодального злиття на основі механізмів уваги.

Конфлікт інтересів

Андрашко Юрій васильович, член редакційної колегії, є автороми цієї статті та не брав участі в редакційному розгляді й ухваленні рішення щодо рукопису. Опрацювання рукопису здійснювалося незалежним редактором. Інші редактори заявляють про відсутність конфлікту інтересів.

Фінансування

Дослідження здійснено в рамках кафедральної науково-дослідної роботи «Моделі і методи системного аналізу в міждисциплінарних дослідженнях» (державний обліковий номер 0125U003246)

Доступність даних

Усі дані доступні в цифровій або графічній формі в основному тексті рукопису.

Використання штучного інтелекту

Автори підтверджують, що при створенні даної роботи вони не використовували технології штучного інтелекту.

Внесок авторів

С. В. Шкіря: концептуалізація, методологія, формальний аналіз, візуаліза-

ція, написання — підготовка початкового рукопису. Ю. В. Андрашко: супервізія, написання — рецензування та редагування.

References

1. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90
2. Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *University of Massachusetts, Amherst, Technical Report 07-49.*
3. Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).* (Vol. 1, pp. 539–546). https://doi.org/10.1109/CVPR.2005.202
4. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (pp. 815–823).
5. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning.* (pp. 1597–1607). *Proceedings of Machine Learning Research, 119.* https://proceedings.mlr.press/v119/chen20j.html
6. Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing, 429,* 215–244. https://doi.org/10.1016/j.neucom.2020.10.081
7. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 35(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50
8. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., & *et al.* (2019). MediaPipe: A framework for building perception pipelines. *Google Research.* https://doi.org/10.48550/arXiv.1906.08172
9. Phan, T. H., Huynh, C. T., Nguyen, M. S., Tran, T., & Nguyen, T. Q. (2026). A deep Siamese ResNet-50 framework with triplet loss for high-precision face verification. *Research Square.* https://doi.org/10.21203/rs.3.rs-8414686/v1

Shkiria S. V., Andrashko Yu. V. Multimodal neural network model for verification based on the fusion of visual context features and spatial attributes.

The article develops and investigates a new multimodal neural network architecture for person verification systems that combines visual context analysis and spatial attributes. A two-stage approach is proposed. In the first stage, a hybrid convolutional neural network is implemented that uses a pre-trained ResNet-50 architecture and a custom branch to extract visual features, achieving 94.52% accuracy on the LFW dataset. In the second stage, a deep multimodal fusion model was developed that also integrates a multidimensional vector of normalized face key points. Using a Siamese architecture with a contrastive loss function and a Hard Negative Mining algorithm achieved a final target recognition accuracy of 96.93%. Experimental evidence shows that deep fusion of visual and geometric features significantly reduces the probability of authentication errors compared with unimodal and basic hybrid approaches.

Keywords: multimodal neural network, face verification, computer vision, feature fusion, Siamese architecture, ResNet-50, LFW.

Отримано: 20.03.2026 Прийнято: 06.04.2026 Опубліковано: 30.04.2026