УДК 004.85
DOI https://doi.org/10.32838/2663-5941/2021.5/14

Зелінський Ю.П.
Державний університет «Житомирська політехніка»

Кравченко С.М.
Державний університет «Житомирська політехніка»

РОЗПІЗНАВАННЯ ЕМОЦІЙНИХ ВИРАЗІВ ОБЛИЧЧЯ ЛЮДИНИ ЗА ДОПОМОГОЮ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ

Об'єктом дослідження є моделі й архітектури згорткових нейронних мереж для вирішення задачі розпізнавання емоційних виразів людського обличчя. Оскільки за останні роки все більше зростає інтерес до розв'язання різноманітних проблем у сфері машинного й глибокого навчання, то виключенням не стала необхідність розпізнавати експресивні ознаки на людському обличчі з якомога кращою точністю. Величезна кількість рішень і досягнень зайвий раз доводить, наскільки актуальними є вивчення подібних задач і побудова власних дієвих методів, які своєю чергою надихатимуть інших більш поглиблено занурюватись у саму сутність наукової області й розширювати горизонти для подальших досліджень. У статті наведено підхід, який дозволить для побудованої структури CNN отримати якісніші результати класифікації за відносно короткі періоди. Для розпізнавання міміки було запропоновано просте рішення з використанням комбінації згорткової нейронної мережі й попередньої обробки зображень. Альтернативним варіантом розв'язання такої проблеми виступає створення власної бази даних, яка буде містити колекції зображень інших наборів даних. Базові емоції за такої умови розділені на такі шість класів: злість, сум, радість, страх, здивування, роздратування. Значна увага дослідження приділена вибору навчальних даних для тренування нейронної мережі, а також методам попередньої обробки й збільшенню даних для підвищення показників працездатності класифікації. Також була детально описана й проілюстрована структура розробленої моделі CNN. Експериментальні результати представили хороші показники класифікації в порівнянні із сучасними експериментами з кросвалідації та перехресними базами даних. Проти інших традиційних моделей підтверджено, що запропонована структура CNN перевершує показники класифікації з меншим часом виконання.

Ключові слова: *комп'ютерне бачення, машинне й глибоке навчання, класифікація, згорткові нейронні мережі, розпізнавання емоцій, метод попередньої обробки даних, метод збільшення даних.*

Постановка проблеми. Натепер системи комп'ютерного зору й ідентифікації об'єктів на зображеннях відіграють важливу роль у сучасному світі. Повноцінне спілкування між людьми неможливе без прояву й аналізу емоцій, тому в сучасних людино-машинних системах все частіше загострюється потреба в застосуванні методів для розпізнавання емоцій.

Завдання розпізнавання емоційних виразів обличчя людини може використовуватись у різних сферах людської діяльності. Наприклад, у сфері робототехніки працюють над створенням інтелектуальних роботів, що можуть визначати емоційний стан людини й зважено на це реагувати. Усе це робиться для того, щоб забезпечити більш дружню та комфортну атмосферу під час спілкування між роботом і людиною. У маркетинговій сфері відповідні системи можуть бути використані з метою оперативного відстеження та реагування на різні проблеми в торгових центрах, супермаркетах і других місцях продажів товарів і послуг. Ефективність вирішення ряду зазначених маркетингових завдань може бути значно підвищена шляхом автоматичного розпізнавання емоцій клієнтів.

Розпізнавання емоцій застосовується також у цілому ряді інших галузей, таких як телекомунікації, системи відеоспостереження, медицина, засоби масової інформації, соціальна сфера, індустрія комп'ютерних ігор, автоматизоване навчання та багато інших.

Хоча галузь розпізнавання емоцій є досить перспективною, вона натепер не дуже розвинута [1]. Головною причиною цього є відсутність єдиних стандартів для розробки алгоритмів, а також відсутність єдиних баз даних, сформованих для навчання алгоритмів розпізнавання емоцій.

Аналіз останніх досліджень і публікацій. Починаючи з 50-х років минулого століття, перших днів зародження області штучного інтелекту, дослідники намагалися створити систему, яка зможе розуміти візуальні дані.

Через десятиліття галузь стала більш відомою як комп'ютерне бачення (computer vision). У 2012 році комп'ютерне бачення зробило стрімкий ривок вперед, коли група дослідників з університету Торонто розробила модель штучного інтелекту, яка перевершила найкращі алгоритми розпізнавання зображень.

Система штучного інтелекту AlexNet [2] (названа на честь дослідника Алекса Крижевського) виграла конкурс із комп'ютерного зору 2012 року, продемонструвавши вражаючу точність у 85%. Основу системи AlexNet склав особливий тип нейронної мережі – згорткові нейронні мережі, які здатні на високому рівні імітувати людський зір.

Постановка завдання. Важливим завданням виступає проведення збору навчальних даних для тренування згорткової нейронної мережі, використовуючи колекцію зображень людей із різних наборів даних. Для підвищення характеристик класифікації застосувати методи для попередньої обробки даних і техніку збільшення даних. Метою роботи є представлення структури розробленої згорткової нейронної мережі й одержання високих результатів експериментальних досліджень із перехресними базами даних.

Виклад основного матеріалу дослідження. Існує багато різних видів нейронних мереж, які можна використовувати в проєктах машинного навчання: рекурентні нейронні мережі, нейронні мережі з прямим зв'язком, модульні нейронні мережі й інші. Згорткова нейронна мережа – це ще один вид широко розповсюджених нейронних мереж. Вона містить згорткові шари, які мають функцію активації, повнозв'язкові й агрегувальні шари, рецептивні поля та ваги. Використовуючи ці складові частини, мережа виконує операцію на основі цих функцій.

Для того, щоб розпізнавати міміку на основі CNN, потрібно мати хорошу навчальну базу даних. У статті було зібрано відомості про десять різних баз даних із метою сформувати добре класифіковану базу даних високої якості для кожного виразу обличчя. Для зручності буде класифікуватися шість типів емоційних виразів – злість, сум, радість, здивування, страх, роздратування. Для цього слід розробити архітектуру CNN із навчальними параметрами, які будуть характерні високими показниками класифікації.

Колекція бази даних.

Для розпізнавання міміки з високою точністю потрібна база даних, що містить велику кількість зображень обличчя. База даних, використана в змаганні «Визначення мімічних виразів обличчя», що відбулася в Kaggle у 2013 році (FER 2013), складається із 40 000 зображень обличчя із сімома класами виразів обличчя [3]. Однак роздільна здатність цих зображень низька (48 х 48 пікселів), і подекуди зустрічаються неправильно позначені зображення. Якщо структура CNN розроблена для вхідного зображення з низькою роздільною здатністю, то потрібно змінити розмір вхідного зображення з високою роздільною здатністю відповідно до структури. Під час цього процесу продуктивність класифікації знижується, оскільки співвідношення зображення змінюється та виникає розмиття. Також неправильне маркування погіршує показники класифікації.

На рис. 1 показані деякі фотографії з бази даних FER2013, які були неправильно позначені для різних виразів обличчя. Для подолання цих проблем ми використаємо такі десять високоякісних баз даних, опублікованих університетами й науково-дослідними інститутами по всьому світу.

Характеристики наборів даних для розпізнавання емоцій:

1. Амстердамський набір динамічних виразів обличчя (ADFES). Цей набір оснований на 648 знятих емоційних виразів. Він демонструє дев'ять емоцій: шість основних емоцій (гнів, огиду, страх,

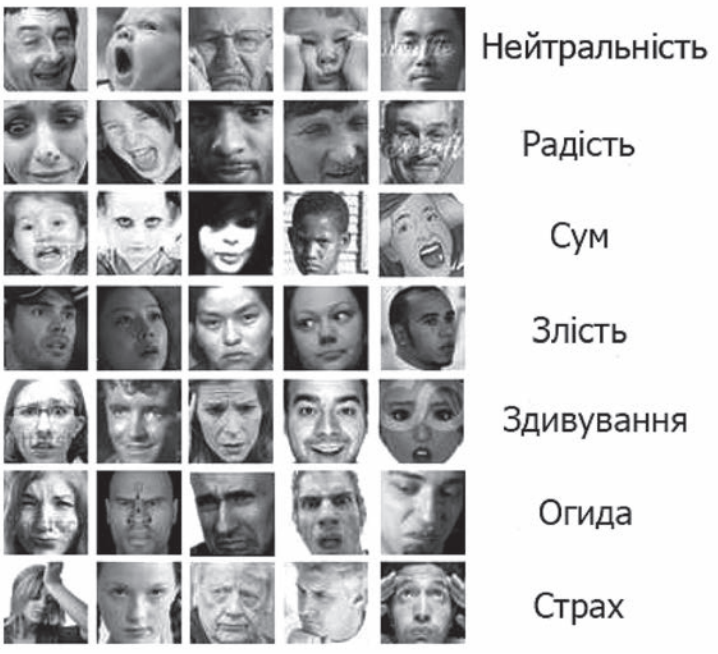

Рис. 1. Приклади зображень, які були неправильно класифіковані для облич у базі даних FER 2013

радість, сум і здивування), а також зневагу, гордість і збентеження. Вирази зображуються на 22 моделях (10 жіночих і 12 чоловічих).

2. Чиказька база даних облич (CFD). Вона містить нейтральні зображення обличчя для 597 людей у віці від 17 до 65 років. Вона складається з різних рас і містить зображення 158 осіб із такими класами: радість, злість і страх.

3. База даних Cohn-Kanade з AU-кодованою мімікою. Вона містить 593 відеопослідовностей для 123 осіб віком від 18 до 30 років. Серед них 309 послідовностей демонструють емоційні вирази: радість, сум, гнів, здивування, страх, огиду й роздратування.

4. EU-Emotion Stimulus Set. Набір складається з таких виразів, як нейтральність, радість, сум, злість, здивування та презирство для 19 акторів у віці від 10 до 70 років. Актори працювали в драматичних школах і театрах діючих агентств Великобританії.

5. База даних ESRC тривимірних облич. Вона містить зображення, зроблені під різними кутами й освітленням за допомогою чотирьох камер для 45 чоловіків і 54 жінок. База складається з класичних емоційних виразів: радість, сум, злість, здивування та роздратування.

6. База даних життєвого шляху. Вона містить дані 575 особистостей віком від 18 до 93 років. База даних була розроблена, щоб бути більш репрезентативною для вікових груп протягом усього життя з особливим акцентом на підборі людей похилого віку.

7. Каролінська база даних емоційних облич (KDEF). Вона містить 4 900 зображень, зроблених під п'ятьма кутами: -90°, -40°, 0°, 45° та 90° для 35 жінок і 35 чоловіків у віці від 20 до 30 років. База даних складається із семи виразів: нейтральність, радість, сум, злість, здивування, страх і роздратування.

8. Radboud Faces Database (RaFD). Це набір фотографій із 67 моделей (включаючи чоловіків і жінок кавказького походження, кавказьких дітей, як хлопчиків, так і дівчаток, марокканських і голландських чоловіків), що демонструють 8 емоційних виразів. RaFD – це високоякісна база даних облич, яка містить зображення восьми емоційних виразів: гніву, огиди, страху, радості, суму, здивування, зневаги й нейтральності. Кожну емоцію показували в трьох різних напрямах погляду, і всі знімки були зроблені одночасно з п'яти ракурсів камери.

9. База даних вебпошуку. База даних, отримана на основі зображень знайдених у вебпошуку.

10. Варшавський набір емоційних виразів обличчя (WSEFEP). Набір містить 210 високоякісних знімків 30 осіб. Вони демонструють шість основних емоцій і нейтральний вираз обличчя.

Методи попередньої обробки даних і збільшення даних.

Оскільки розпізнавання виразів обличчя людини вимагає лише інформації про область обличчя, необхідна попередня обробка, щоб виявити й вирізати лише область обличчя навчального зображення. У статті використано метод, заснований на характеристиках Хаара, для виявлення та вирізання області обличчя [4]. На рис. 2 показаний результат виявлення та вирізання області обличчя з вихідного зображення та перетворення його на зображення сірого кольору.

Якщо кількість навчальних зображень недостатня в порівнянні з параметрам и навчання CNN, може виникнути проблема з надмірною підгонкою, а продуктивність класифікації знизиться. Для розв'язання такої проблеми використовується техніка збільшення даних, яка додає кількість навчальних зображень.

Архітектура CNN проілюстрована на рис. 3. Мережа складається з восьми шарів. Перші п'ять шарів є згортковими (C1-5), а наступні три – повнозв'язковими (FC6-8). Вихідний сигнал останнього повнозв'язкового шару надходить на шести напрямлену функцію активації softmax, яка здійснює розподіл по шести мітках класу. Шари Maxpooling слідують за першим, другим і п'ятим згортковим шаром. Не лінійність функції ReLU (Rectified Linear Unit) застосовується до виводу кожного згорткового й повнозв'язкового шару.

Перший згортковий шар фільтрує вхідне зображення розміром 227 × 227 із 96 ядрами розмірністю 11 × 11 із кроком 4 пікселі. Другий згортковий шар бере як вихідні дані перший згортковий шар і фільтрує його зі 128 ядрами розмірністю

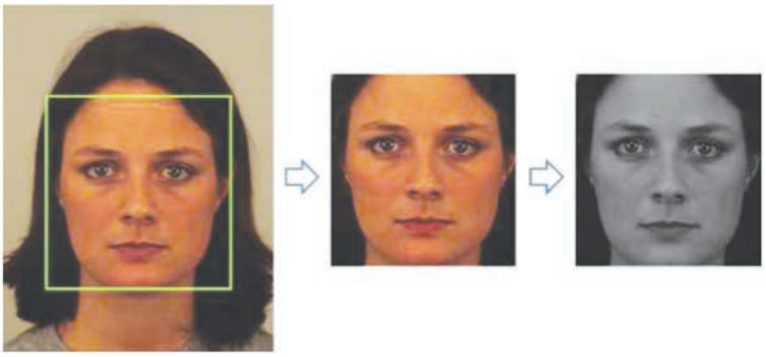

Рис. 2. Результат перетворення вирізаної області зображення в сіре зображення

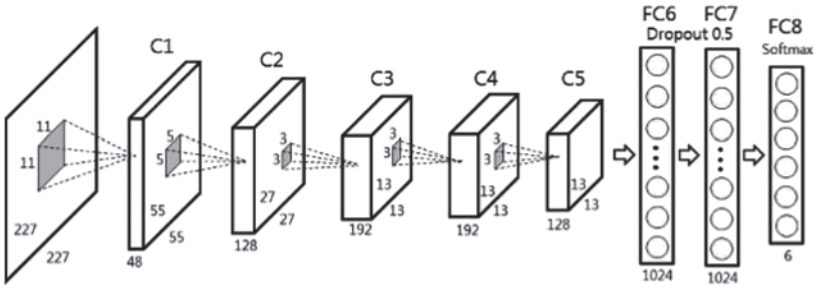

Рис. 3. Запропонована архітектура CNN

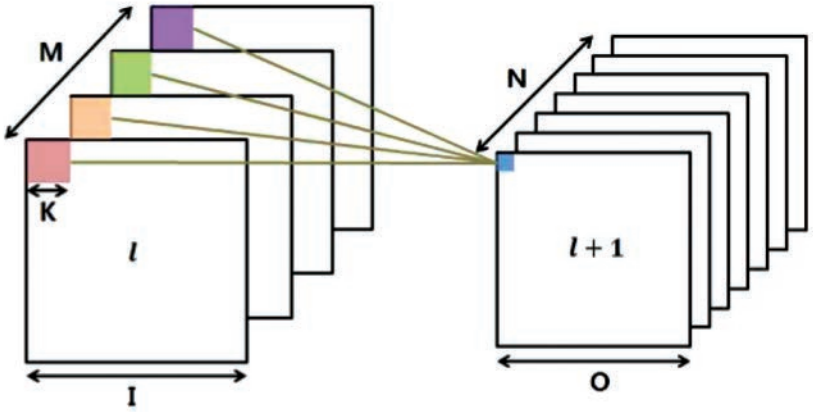

Рис. 4. Обчислювальний зв'язок між послідовними згортковими шарами

5 × 5 × 48. Третій, четвертий і п'ятий згорткові шари з'єднані між собою без втручання шарів об'єднання або нормалізації. Третій згортковий шар має 192 ядра розмірністю 3 × 3 × 128, підключених до виходів другого згорткового шару. Четвертий згортковий шар має 192 ядра розмірністю 3 × 3 × 192, а п'ятий згортковий шар має 128 ядер розміром 3 × 3 × 192. Повністю з'єднані шари мають по 1 024 нейрони. Для запобігання надмірного перенавчання даних до перших двох повнозв'язкових шарів застосовується метод dropout.

Запропонована структура аналогічна AlexNet [2]. У розпізнаванні виразів обличчя кількість каналів у згортковому шарі й кількість вузлів у повнозв'язковому шарі зменшуються для того, щоб вибрати оптимальну структуру з вищою продуктивністю класифікації, меншим часом виконання та меншими параметрами навчання.

Як бачимо на рис. 4, коли кількість каналів послідовних згорткових шарів зменшується, то кількість обчислень квадратично зменшується. Таким чином, кількість каналів кожного згорткового шару зменшується вдвічі, що значно зменшує обсяг обчислень. На відміну від AlexNet, який для класифікації містить 1 000 класів, вищезгадана структура має класифікувати лише шість виразів. Таким чином, кількість вузлів у повнозв'язковому шарі зменшується більш ніж в чотири рази.

Експерименти й результати дослідження. Реалізація кроків попередньої обробки була здійснена за допомогою бібліотеки OpenCV і мови програмування Python. Попередньо оброблені навчальні зображення сірого кольору повертаються під кутами -15°, -10°, -5°, 5°, 10° і 15° і перевертаються горизонтально. У результаті будуть отримані чотирнадцять зображень. Далі порівнюється точність розпізнавання виразів обличчя з і без збільшення даних.

Таблиця 1

Порівняння точності методів збільшення даних

Попередня обробка	Точність (%)
Без збільшення даних	89,05
Зі збільшенням даних	93,93

У табл. 1 показано вплив методів збільшення на точність розпізнавання в оптимальній структурі (48, 128, 192, 192, 128, 1024, 1024, 6). Як показано в табл. 1, за застосування техніки збільшення даних точність розпізнавання значно зростає.

Далі оцінюється точність нової архітектури глибокої нейронної мережі у двох різних експериментах, таких як перехресна валідація та перехресна база даних. В експерименті з перехресною валідацією було використано техніку перехресної перевірки з К-кратною формою, де К = 10. База даних була розділена на 10 груп без накладання предметів між групами. Ця методологія забезпечує генералізацію класифікаторів.

Таблиця 2

Середня точність (%) у разі перехресної валідації

База даних для оцінювання	Точність структури	Сучасний рівень
ADFES	100,00	96,30
CFD	97,64	-
CK+	96,83	96,76
EU-Emotion Stimulus Set	79,17	-
ESRC	84,76	-
FACE DATABASE	91,29	89,81
KDEF	91,79	89,00
RafD	99,25	93,96
Web Search	81,97	-
WSEFEP	96,11	-

У табл. 2 наведено середню точність класифікації зображень за шістьма виразами. Середню матрицю неточностей в експериментах із перехресної валідації можна побачити в табл. 3. У табл. 2 результати експерименту з перехресної перевірки кращі в порівнянні з поточним сучасним рівнем техніки. У вже опублікованих статтях немає результатів із використанням CFD, EU-Emotion Stimulus Set, бази даних ESRC і WSEFEP. Web Search також не є порівняльною базою даних, оскільки кожне зображення виразу обличчя видобувається та отримується окремо. У результаті суворого експерименту перехресної валідації можна побачити, що запропонована структура має чудові загальні показники класифікації з боку предмета в одному середовищі. Табл. 3 показує високу точність – щонайменше 86,37% – для шести виразів, які мають бути класифіковані в матриці неточностей для експерименту перехресної валідації.

Таблиця 3

Середня матриця неточностей за перехресної валідації (%)

	NE	HA	SA	AN	SU	DI
NE	97,15	0,65	1,36	0,58	0,19	0,06
HA	2,30	96,19	0,16	0,40	0,48	0,48
SA	5,18	0,43	87,19	4,60	1,15	1,44
AN	1,86	0,74	6,20	86,37	0,12	4,71
SU	1,93	0,64	0,26	0,13	96,79	0,26
DI	0,74	1,78	3,56	3,12	0,59	90,21

В експерименті між базами даних одна база даних використовується для оцінки, а решта баз даних використовується для навчання мережі. Перехресні бази даних є складним завданням, оскільки кожна база даних має різне освітлення, положення людини, кут нахилу камери й емоційний вираз. У табл. 4 наведено середню точність між базами даних під час класифікації шести виразів.

Результатом є точність класифікації з використанням методу опорних векторів з ядром функції радіальної бази (SVM + RBF) [5]. Однак найкращі результати були отримані шляхом зміни бази навчання та параметрів класифікатора. Завдяки експериментам із перехресною валідацією та перехресними базами даних можна стверджувати, що запропонована структура CNN підходить для генералізації розпізнавання виразів обличчя.

Для порівняння з іншими моделями CNN експеримент перехресної перевірки був проведений для AlexNet, VGGNet (11 шарів), OverFeat (швидка модель) і CNN. Відбувається вимір часу навчання та тестування, коли пакет проходить через кожну модель. Табл. 5 показує, що запропонована структура витрачає набагато менше часу як на навчання, так і на тестування.

Ми навчали кожну модель CNN із нуля, використовуючи той самий протокол, який використовується для навчання нашої власної мережі. У табл. 6 наведено класифікаційну точність кожної моделі. Можна підтвердити, що класифікація запропонованої моделі найкраща.

Таблиця 4

Середня точність (%) у перехресній базі даних

База даних для оцінювання	Точність запропонованої структури	Сучасний стан
ADFES	99,24	-
CFD	92,16	-
CK+	92,61	64,20
EU-Emotion Stimulus Set	73,96	-
ESRC	72,11	-
FACE DATABASE	78,32	79,27
KDEF	78,81	-
RafD	94,94	-
Web Search	82,79	-
WSEFEP	95,56	-

Висновки. Описана модель згорткової нейронної мережі CNN найкраще підходить для розпізнавання виразів обличчя для шести емоцій. Структура запропонованого алгоритму має хорошу генералізацію та ефективність класифікації. Методика збільшення даних застосовується для розв'язання проблеми надмірності, яка погіршувала показники класифікації. Створена модель CNN має оптимальну структуру для скорочення часу виконання та підвищення ефективності класифікації та визначалася шляхом коригування кількості карт характеристик у згортковому шарі й кількості вузлів у повнозв'язковому. Проведені експерименти підтвердили ефективність методів попередньої обробки й збільшення даних.

Таблиця 5

Час тренування та тестування для кожної моделі

Модель	Час тренування (пакетів за секунду)	Час тестування (пакетів за секунду)
AlexNet	0,325	0,068
OverFeat	0,593	0,125
VGGNet	2,128	0,569
Запропонована модель	0,131	0,027

Таблиця 6

Точність для кожної моделі

Модель	Точність
AlexNet	93,55
OverFeat	93,55
VGGNet	91,60
Запропонована модель	93,95

Список літератури:

1. Bartlett M.S., Littlewort G., Frank M., Lainscsek C., Fasel S., Movellan J. Recognizing facial expression: machine learning and application to spontaneous behavior. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 20–25 June 2005. San Diego, 2005. P. 568–573.

2. Krizhevsky A., Sutskever S., Hinton G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. 2012. P. 1–9.

3. Змагання в Kaggle для розпізнавання емоційних виразів обличчя людини. URL: https://www.kaggle.com/shawon10/facial-expression-detection-cnn (дата звернення: 28.08.2021).

4. Viola P., Jones M. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition*, 2001, CVPR 2001. Proceedings of the 2001, IEEE Computer Society Conference on. Vol. 1. IEEE, 2001.

5. Bekios-Calfa J., Buenaposada J.M., Baumela L. Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2011. Vol. 33. No. 4. P. 858–864.

Zelinskyi Yu.P., Kravchenko S.N. **RECOGNITION OF EMOTIONAL EXPRESSIONS OF HUMAN FACE USING CONVOLUTIONAL NEURAL NETWORK**

The object of research is the models and architectures of convolutional neural networks to solve the problem of recognizing the emotional expressions of the human face. As interest in solving various problems in the field of machine and deep learning has been growing in recent years, it is no exception that it is not necessary to recognize expressive features on the human face with the best possible accuracy. The huge number of solutions and achievements proves once again how relevant it is to study such problems and build their own effective methods, which in turn will inspire others to delve deeper into the essence of the scientific field and expand horizons for further research. The article presents an approach that will allow for the constructed structure of CNN to obtain better classification results in relatively short periods of time. A simple solution using a combination of convolutional neural network and image pre-processing was proposed to recognize facial expressions. An alternative solution to this problem is to create your own database, which will include collections of images from other datasets. Basic emotions are divided into the following six classes: anger, sadness, joy, fear, surprise, irritation. Much attention is paid to the choice of training data for neural network training, as well as methods of pre-processing and data augmentation to improve the performance of the classification. The structure of the developed CNN model was also described in detail and illustrated. The experimental results showed good classification performance compared to modern cross-validation experiments and cross-databases. Compared to other traditional models, it is confirmed that the proposed structure of CNN exceeds the classification indicators with less execution time.

Key words: *computer vision, machine and deep learning, classification, convolutional neural networks, emotion recognition, data preprocessing method, data augmentation method.*