

DOI: <https://doi.org/10.36910/6775-2524-0560-2025-59-17>

УДК 004.93

Казимира Ірина Ярославівна, канд. техн. наук, доцент

<https://orcid.org/0000-0003-1597-5647>

Цапів Володимир Володимирович, магістр

<https://orcid.org/0009-0008-2420-7483>

Національний університет "Львівська політехніка", м. Львів, Україна

СИСТЕМА РОЗПІЗНАВАННЯ ДИНАМІЧНИХ ЖЕСТИВ РУК У РЕАЛЬНОМУ ЧАСІ НА ПЕРИФЕРІЙНИХ ПРИСТРОЯХ

Казимира І. Я., Цапів В. В. Система розпізнавання динамічних жестів рук у реальному часі на периферійних пристроях. У даній роботі представлено систему розпізнавання динамічних жестів рук у реальному часі, призначену для ефективної взаємодії з інтелектуальними пристроями та безконтактними інтерфейсами, зокрема на периферійних пристроях. Розроблена система інтегрує Google Mediapipe для визначення пози рук, який є достатньо легким для запуску на мобільних пристроях, з модифікованою архітектурою DD-Net, оптимізованою для онлайн класифікації жестів з використанням 2D та 3D даних. Ключові нововведення включають впровадження допоміжної класифікаційної гілки для вирішення проблеми дисбалансу класів та механізму уваги для покращення розпізнавання частково спостережуваних жестів. Розроблена система була протестована на наборах даних NVGesture та SHREC22, досягнуто точності 0.784 та 0.924, відповідно, що перевищує попередні результати. Експериментальні результати демонструють високу ефективність запропонованого підходу у задачах розпізнавання жестів в реальному часі, зокрема на мобільних платформах. Показано, що використання 3D даних та механізму уваги дозволяє значно покращити точність розпізнавання, особливо для складних та частково видимих жестів. Запропонована система може бути використана в широкому спектрі застосувань, включаючи системи віртуальної та доповненої реальності, робототехніку та безконтактні інтерфейси на периферійних пристроях.

Ключові слова: розпізнавання жестів в реальному часі, взаємодія людина-комп'ютер, динамічні жести рук, глибоке навчання, комп'ютерний зір, периферійні пристрої.

Kazymyra I., Tsapiv V. Real-time dynamic hand gesture recognition system on edge devices. This paper presents a novel real-time dynamic hand gesture recognition system designed for efficient interaction with smart devices and touchless interfaces, with a focus on edge devices. The proposed system integrates Google Mediapipe for hand pose detection, which is lightweight enough to run on mobile devices, with a modified version of the DD-Net architecture, optimized for online classification of gestures using 2D and 3D data. Key innovations include the introduction of an auxiliary classification head to address class imbalance and an attention mechanism to improve the recognition of partially observed gestures. The system is evaluated on the NVGesture and SHREC22 datasets, achieving an accuracy of 0.784 and 0.924, respectively, surpassing previous benchmarks. Experimental results demonstrate the high efficiency of the proposed approach in real-time gesture recognition tasks, particularly on mobile platforms. It is shown that the use of 3D data and an attention mechanism significantly improves recognition accuracy, especially for complex and partially visible gestures. The proposed system can be used in a wide range of applications, including virtual and augmented reality systems, robotics, and touchless interfaces on edge devices.

Keywords: real-time gesture recognition, human-computer interaction, dynamic hand gestures, deep learning, computer vision, edge devices.

Постановка проблеми. Взаємодія людини з комп'ютером (Human-Computer Interaction, HCI) активно розвивається, переходячи від традиційних методів введення даних до більш природних форм комунікації. Одним із перспективних напрямків є розпізнавання жестів рук, яке забезпечує інтуїтивне управління пристроями, зокрема у сферах доповненої та віртуальної реальності (AR/VR), робототехніки та безконтактних інтерфейсів. Важливість цієї технології суттєво зросла через необхідність мінімізувати фізичний контакт у громадських місцях, що стало особливо актуальним після пандемії COVID-19 [1].

У світовій науковій літературі дослідження у сфері розпізнавання жестів охоплюють різні методи комп'ютерного зору та глибокого навчання. Значний прогрес досягнуто у використанні згорткових нейронних мереж (CNN), рекурентних мереж (RNN), трансформерів та гібридних моделей. Однак більшість існуючих рішень орієнтовані на аналіз попередньо підготовлених послідовностей і не завжди відповідають вимогам реального часу через високі обчислювальні витрати. Це створює розрив між теоретичними досягненнями та практичними застосуваннями, особливо у випадках, коли необхідна низька затримка та ефективна робота на пристроях з обмеженими ресурсами.

Незважаючи на значні досягнення в галузі, залишається відкритим питання ефективного розпізнавання динамічних жестів у реальному часі з урахуванням обчислювальних обмежень. Відповідно, дослідження моделей і методів розпізнавання жестів у потоковому відео для

забезпечення інтерактивного управління смарт-пристроями та безконтактними інтерфейсами є актуальним завданням.

Об'єкт і предмет дослідження. Об'єктом дослідження є процеси розпізнавання жестів у реальному часі на потоковому відео. Предметом дослідження є нейронні мережі, алгоритми та інструменти, що використовуються для цієї задачі.

Мета і завдання дослідження. Основна мета роботи – розробка та оптимізація системи розпізнавання жестів у реальному часі. Для її досягнення були визначені такі завдання:

- провести аналіз існуючих методів розпізнавання жестів та їх ефективності для застосування в реальному часі;
- оцінити продуктивність сучасних моделей з урахуванням апаратних обмежень;
- розробити та оптимізувати архітектуру нейронної мережі для стабільного розпізнавання жестів;
- провести експериментальні дослідження та оцінити точність розробленої системи на різних наборах даних.

Розроблена система розпізнавання жестів сприятиме покращенню інтерактивного управління смарт-пристроями, забезпечуючи швидкість та точність роботи в реальному часі.

Аналіз існуючих моделей та підходів.

Однією з передових моделей є STA-GCN (Spatial-Temporal Attention Graph Convolutional Network), запропонована Wang та ін. [4]. Ця модель використовує двопотокову архітектуру графової згорткової мережі, де окремо обробляються дані про пози та рухи рук. Ключовою особливістю є застосування просторово-часової уваги, що дозволяє моделі виявляти найбільш важливі суглоби та часові інтервали для точного розпізнавання жестів. STA-GCN [4, 11] показала високу ефективність на наборах даних DHG14/28 та SHREC2017 [18]. Механізми уваги відіграють важливу роль у підвищенні точності, дозволяючи моделі концентруватися на найбільш інформативних частинах вхідних даних.

У роботі [5] представлено багаторівневу модель LSTM, яка комбінує дані глибини та скелетні дані для підвищення надійності розпізнавання жестів. Хоча ця модель досягла високих результатів на еталонних наборах даних, вона вимагає значних обчислювальних ресурсів через використання даних глибини та складної архітектури LSTM, що може бути обмеженням для застосувань на пристроях з обмеженою потужністю.

DD-Net (Double-feature Double-motion Network), розроблена Liu та ін. [6], є прикладом легковажкої мережі, спрямованої на швидке та ефективне розпізнавання дій на основі скелетних даних. DD-Net використовує різниці координат суглобів (JCD) та багатошкальну глобальну характеристику руху для врахування змін швидкості виконання жестів. Результати показали, що хоча JCD є інваріантними до положення камери, для представлення глобального руху потрібні двошкальні характеристики. Важливо відзначити, що, на відміну від STA-GCN [4], оригінальна версія DD-Net не використовує механізми уваги.

У контексті SHREC 2022 [7] розглядалися гетерогенні системи розпізнавання жестів, зокрема моделі, що використовують двоступеневу ST-GCN [7] та Casual TCN [7]. Ці моделі застосовують техніку ковзного вікна для ізоляції жестів у безперервному потоці даних та використовують часові згорткові мережі (TCN) для уникнення витоку інформації з майбутнього, що є важливим для онлайн-детекції. Також була представлена мережа Transformer з машиною станів у кінці (TN-FSM) [7], яка використовує 3D-позиції суглобів рук для збагачення характеристик та чіткого визначення меж жестів у часі.

Модель Stronger [8] інтегрує DD-Net з техніками ковзного вікна для безперервного виявлення жестів, використовуючи шість різних вхідних характеристик, включаючи JCD, різниці між парами суглобів та орієнтацію долоні. Хоча Stronger [8] досягає високої точності класифікації, вона страждає від великої кількості помилкових розпізнавань у моменти, коли жест насправді не виконується.

DeepGRU [9] демонструє інший підхід, використовуючи GRU (Gated Recurrent Units) та глобальний механізм уваги для навчання розпізнаванню жестів від початку до кінця на основі сирих даних скелетів та поз. Архітектура на основі GRU з увагою пропонує простішу, але ефективну альтернативу GCN та LSTM, досягаючи порівнянних результатів без необхідності складної попередньої обробки даних.

У рамках SHREC 2019 [10] були представлені нові архітектури для онлайн-сегментації жестів, такі як Divide et Agnosce та Sliding Window 3-Cent. Divide et Agnosce використовує двоступеневу мережу, де один етап сегментує жести в реальному часі, а інший класифікує кожен сегмент. Метод Sliding Window 3-Cent є адаптацією алгоритму k -найближчих сусідів (kNN) для роботи в онлайн-контексті.

Сучасні дослідження в галузі розпізнавання жестів спрямовані на розробку моделей, які є одночасно точними та ефективними, особливо для застосувань у реальному часі [8, 12, 15, 16]. Актуальність цієї галузі постійно зростає завдяки зростанню попиту на інтуїтивні та безконтактні способи взаємодії з технологіями в різних сферах, від керування смарт-пристроями до віртуальної реальності та медичних застосувань.

Невирішені проблеми та обмеження. Незважаючи на значний прогрес, існують невирішені проблеми. Деякі моделі, особливо ті, що використовують LSTM або архітектуру Transformer з даними глибини, мають високі обчислювальні вимоги [21]. Проблема помилкових розпізнавань у моменти відсутності жесту, як у моделі Stronger [8], також потребує вирішення. Оптимізація моделей для онлайн-застосувань з метою зменшення затримки залишається актуальною задачею [3]. Крім того, індивідуальні відмінності у виконанні жестів можуть впливати на точність розпізнавання [2], а надійність систем залежить від якості вхідних даних.

Матеріали і методи дослідження.

Загальний опис розробленої системи розпізнавання жестів. Розроблена система виконує розпізнавання жестів рук у реальному часі на основі послідовних 3D-точок з відеопотоку. Актуальність цієї розробки зумовлена розвитком безконтактних інтерфейсів людина-комп'ютер для різних застосувань. Архітектура системи оптимізована для низької затримки та використовує ефективні методи обробки даних і класифікації.

Система включає отримання 3D-точок рук за допомогою Mediapipe [13] або інших сенсорних пристроїв, їх буферизацію та попередню обробку. Для класифікації жестів використовується одновимірна згортова нейронна мережа DD-Net [6], яка була адаптована та оптимізована для розпізнавання жестів рук на основі скелетних даних у межах цієї роботи. Результати класифікації обробляються алгоритмом голосування для отримання надійного прогнозу.

Отримання та буферизація вхідних даних. Система розпізнавання жестів підтримує різні сенсорні пристрої для отримання вхідних даних, включаючи стандартні RGB-камери з Mediapipe [13], Intel RealSense та HoloLens 2 [14]. Гнучкість у виборі джерела дозволяє використовувати систему в різних сценаріях. Кожен сенсор має свої особливості щодо точності, вартості та умов використання. У цьому дослідженні використовувалася стандартна RGB-камера з Mediapipe як компроміс між вартістю, доступністю та точністю.

Для кожного кадру відеопотоку відбувається вилучення 3D-ключових точок рук за допомогою Mediapipe (при використанні RGB-камери). Точність вилучення ключових точок є критично важливою для якісного розпізнавання жестів. Mediapipe Hands надає інформацію про 21 ключову точку руки, і кожен жест у часі представляється як послідовність цих положень.

Для захоплення динаміки руху система використовує буфер на 16 кадрів, що працює за принципом ковзного вікна. Кадри в буфері перед обробкою організовуються у зворотному порядку, що надає системі найновіші кадри на початку послідовності для покращення розпізнавання динамічних жестів.

Попередня обробка даних. Отримані послідовності 3D-точок рук проходять етапи попередньої обробки для підвищення якості даних. Важливим етапом є обробка відсутніх даних. При короткочасних збоях у відстеженні руки деякі кадри можуть не містити дійсних 3D-точок. Для відновлення цілісності послідовності застосовується лінійна інтерполяція координат для кожного 3D-суглоба між сусідніми дійсними кадрами в межах невеликого ковзного вікна (3-5 кадрів).

Наступним етапом є нормалізація даних. Координати ключових точок кожного кадру нормалізуються для усунення варіацій масштабу, положення та міжкористувачьких відмінностей шляхом статистичного регулювання з використанням середнього значення та стандартного відхилення, розрахованих на навчальному наборі даних. Нормалізація забезпечує отримання нейронною мережею вхідних даних у подібному діапазоні до навчальних даних.

Ще одним важливим аспектом є підлаштування до частоти кадрів. Оскільки частота кадрів відеопотоку може відрізнятися від частоти навчальних даних, система виконує повторну вибірку кадрів. При вищій частоті вхідних даних вибираються фіксовані 16 кадрів через рівні проміжки часу,

а при нижчій може застосовуватися лінійна інтерполяція для отримання 16 кадрів у буфері. Це допомагає зберегти часовий контекст руху рук.

Генерування ознак для навчання нейронної мережі. Генеруються три типи ознак: різниці координат суглобів (JCD), "повільні" зміщення (M_{slow}) та "швидкі" зміщення (M_{fast}).

- Різниці координат суглобів (JCD):
Обчислюються евклідові відстані між кожною парою з J суглобів для $W = 16$ кадрів, формуючи тензор розміром $(J * (J - 1)/2) \times W$.
- "Повільні" зміщення (M_{slow}):
Обчислюються як різниця координат кожного суглоба на сусідніх кадрах, формуючи тензор розміром $J \times (W - 1)$.
- "Швидкі" зміщення (M_{fast}): обчислюються як різниця координат кожного суглоба на кадрах через один, формуючи тензор розміром $J \times (W/2 - 1)$.

Архітектура нейронної мережі розпізнавання жестів. Розроблена система використовує багатопшарову згорткову архітектуру нейронної мережі (рис. 1), яка включає три окремі гілки для обробки кожного з трьох типів згенерованих ознак (JCD , M_{slow} та M_{fast}). Такий підхід дозволяє кожній гілці спеціалізуватися на вилученні специфічних характеристик з відповідного типу ознак, а подальше об'єднання інформації з усіх гілок дає змогу отримати більш повне уявлення про виконуваний жест.

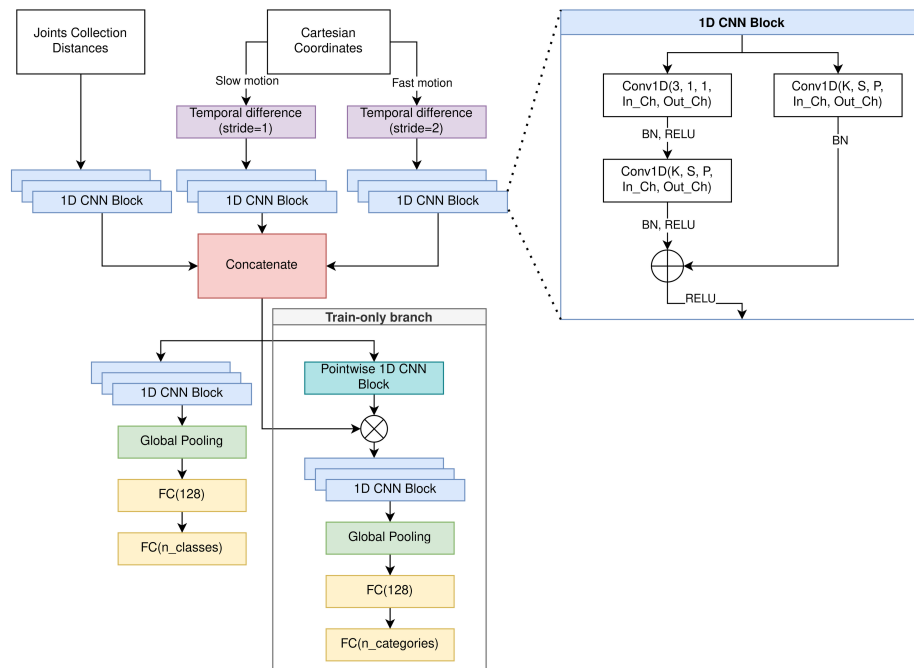


Рис. 1. Архітектура моделі розпізнавання жестів

Кожна з трьох згорткових гілок має схожу структуру та складається з серії згорткових блоків. Кожен згортковий блок включає одномірний згортковий шар (Conv1D), шар пакетної нормалізації (Batch Normalization) та шар активації Leaky ReLU. Одномірні згорткові шари ефективно обробляють послідовні дані, такі як часові ряди ознак. Пакетна нормалізація сприяє стабілізації процесу навчання, а функція активації Leaky ReLU вводить нелінійність у модель. Крім того, кожен згортковий блок містить резервне з'єднання (residual connection), що є характерною рисою архітектури ResNet [21]. Резервні з'єднання допомагають уникнути проблеми зникаючих градієнтів, що може виникати при навчанні глибоких нейронних мереж, забезпечуючи ефективніше зворотне поширення градієнта через шари.

Після обробки ознак у відповідних згорткових гілках, отримані представлення ознак подаються на дві різні класифікаційні голови: основну та допоміжну. Основна класифікаційна голова відповідає за безпосереднє прогнозування класу виконаного жесту. Її структура включає додаткові згорткові блоки та шари Max Pooling для подальшого вилучення ознак, після чого йдуть

один або кілька повнозв'язних шарів та фінальний шар Softmax, який генерує розподіл ймовірностей за всіма класами жестів. Допоміжна класифікаційна голова призначена для покращення процесу навчання, особливо в умовах онлайн-розпізнавання, де часто зустрічаються кадри, що не містять жодних жестів. Ця голова класифікує вхідні дані на три категорії: статичний жест, динамічний жест та відсутність жесту. Важливою особливістю допоміжної голови є наявність модуля уваги, що складається з точкової згортки (1x1 згортки) та функції Softmax, дозволяє мережі виділяти найбільш важливі часові сегменти вхідної послідовності ознак та ігнорувати нерелевантні частини, таким чином підвищуючи ефективність обробки та точність розпізнавання.

Алгоритм голосування для остаточного прогнозу. Алгоритм голосування за принципом автоналаштування вагомої більшості на основі ентропії (Entropy Weighted Voting) [19] підвищує точність прогнозування, застосовуючи зважене голосування до прогнозів, зроблених k моделями. Цей метод використовує ентропію для вимірювання невизначеності кожного окремого естиматора. Прогнози з високою ентропією (низька впевненість) мають менший вплив на кінцеве рішення. Ми адаптували цей алгоритм: зваження стосується не моделей (модель тут одна), але k послідовних прогнозів.

Ключові етапи пропонованого алгоритму описано нижче.

1. Матриця апостеріорної ймовірності. Для послідовності з k кадрів (або часових кроків), $P(x)$ збирає вихідні ймовірності, передбачені моделлю для кожного з m класів. У цьому випадку $P(x)$ складається з k послідовних прогнозів від однієї й тієї ж моделі. Вищі ймовірності вказують на більшу впевненість щодо конкретного класу в даний момент часу, тоді як рівномірні розподіли свідчать про вищу невизначеність.

2. Обчислення ваги на основі ентропії. Ентропія кожного прогнозу $H_i(x)$ вимірює впевненість моделі для даного кадру (або часового кроку):

$$H_i(x) = -\sum_{j=1}^m P_{ij}(x) \cdot \log(P_{ij}(x)) \quad (1)$$

де $P_{ij}(x)$ – значення ймовірності, присвоєне моделлю j -ому класу для i -го кадру (або часового кроку).

3. Ваги обчислюються за допомогою експоненти негативної ентропії, щоб їхня сума дорівнювала одиниці:

$$w_i = \frac{\exp(-H_i(x))}{\sum_{j=1}^k \exp(-H_j(x))} \quad (2)$$

де $H_i(x)$ – ентропія i -го прогнозу.

4. Зважена матриця ймовірності. Кориговані ймовірності $P'(x)$ створюються шляхом масштабування виходу кожного прогнозу відповідно до його ваги:

$$P'_{ij}(x) = w_i \cdot P_{ij}(x) \quad (3)$$

де $P_{ij}(x)$ – значення ймовірності, присвоєне моделлю j -ому класу для i -го кадру (або часового кроку), а w_i – вага i -го прогнозу.

5. Фінальне зважене голосування. Ймовірності для кожного класу агрегуються по всій послідовності прогнозів:

$$p_j^{vote} = \sum_{i=1}^k P'_{ij}(x) \quad (4)$$

де $P'_{ij}(x)$ – значення зваженої ймовірності, присвоєне моделлю j -ому класу для i -го кадру (або часового кроку).

Клас з найвищою кумулятивною ймовірністю серед p_j^{vote} є прогнозованим класом для всієї послідовності.

Результати дослідження та їх обговорення.

Для експериментів з метою оцінки розробленої моделі використовувалися набори даних NVGesture [21] та SHREC22 [7], коротка характеристика яких наводиться нижче.

SHREC'22. Набір даних SHREC'22 містить безперервні 3D положення руки, зібрані в умовах змішаної реальності за допомогою HoloLens 2. Навчальна та тестова вибірки містять по 144 послідовності з 16 класами жестів, що чергуються з незначними рухами. Жести поділені на статичні, динамічні (грубі та точні) та періодичні, з анотаціями початкових і кінцевих кадрів. Дані зібрані різними людьми для кращої узагальненості.

NVGesture. Модифікований набір даних NVGesture містить понад 1500 відеопослідовностей з 25 жестами рук від 20 учасників. Жести поділяються на статичні та динамічні. Для використання з

моделлю була розроблена автоматизована процедура анотації з Mediapipe для вилучення 2D та 3D координат суглобів. Застосовувалася фільтрація за розміром для усунення неоднозначностей з кількома руками в кадрі та інтерполяція (вікно 5 кадрів) для відновлення відсутніх даних через розмиття руху.

Метрики оцінювання. Ефективність моделі оцінювалася за допомогою стандартних метрик класифікації:

- Accuracy: $(TP+TN)/(TP+TN+FP+FN)$
- Recall: $TP/(TP+FN)$
- Precision: $TP/(TP+FP)$
- F1 score: $2*(precision*recall)/(precision+recall)$

Процедура оцінки. Оцінювання включало тестування на контрольній вибірці NVGesture з використанням зазначених метрик та оцінку в реальному часі з вебкамою (кожен жест відтворювався тричі). Обидва етапи проводилися з використанням 3D- та 2D-координат суглобів.

Набір даних NVGesture. Результати тестування на наборі даних NVGesture (див. табл. 1) демонструють порівняння 2D та 3D даних.

Таблиця 1. Результати тестування моделі на тестових даних NVGesture

	Accuracy	Recall	Precision	F1 score
2D дані	0.794	0.794	0.763	0.768
3D дані	0.784	0.758	0.784	0.751

Використання 3D-даних дозволило досягти кращих результатів порівняно з оригінальною роботою для NVGesture (ассурагу 0.74 з використанням лише кольору). Матриця заплутаності для контрольного набору даних NVGesture (рис. 2) виявила складнощі з розрізненням жестів, що виконуються схожими пальцями. Мануальна оцінка (див. табл. 2) показала кращу узагальненість моделі, навченої на 3D-даних, порівняно з 2D-даними.

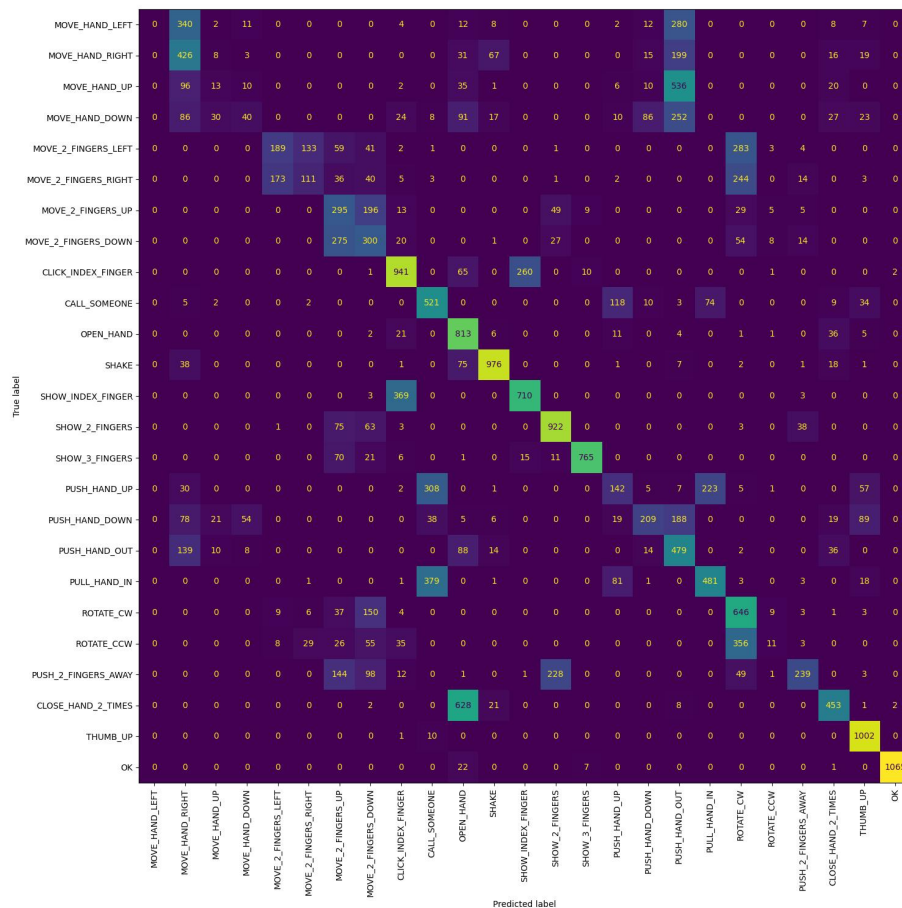


Рис. 2. Матриця заплутаності за результатами класифікації набору даних NVGesture (3D дані).

Таблиця 2. Результати тестування моделі мануально

Тип жесту	2D дані	3D дані
Статичний	0.345	0.893
Динамічний	0.567	0.712
Не жест	0.790	0.921

Приклади розпізнавання динамічних та статичних жестів під час мануальної перевірки виявили труднощі з розпізнаванням статичних жестів, що супроводжуються незначними рухами, та складних жестів з перекриттям пальців.

Набір даних SHREC22. Оцінювання на SHREC22 проводилося лише на 3D-даних. Результати представлено нижче (див. табл. 3). Матрицю заплутаності для цього набору наведено на рис. 3.

Таблиця 3. Результати тестування моделі на тестових даних SHREC22

	Accuracy	Recall	Precision	F1 score
3D дані	0.924	0.924	0.926	0.924

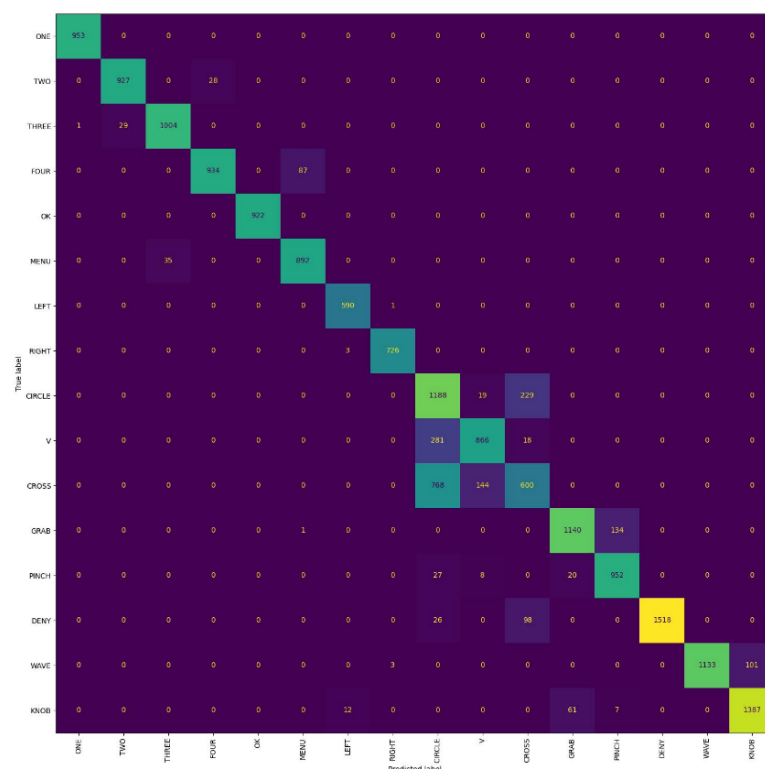


Рис. 3. Матриця заплутаності за результатами класифікації контрольного набору даних SHREC22.

Отримані результати (рис. 3 і табл. 3) підтверджують ефективність підходу на основі якісних 3D-даних для розпізнавання як статичних, так і динамічних жестів. Порівняння з 2D-даними та мануальне тестування не проводилися через специфіку набору даних SHREC22.

Обговорення отриманих результатів. Проведене дослідження було спрямоване на оцінку ефективності розробленої моделі розпізнавання жестів з використанням двох різних наборів даних: NVGesture та SHREC22. Результати, отримані на наборі даних NVGesture, показали, що використання 3D-координат суглобів рук забезпечує дещо вищу точність порівняно з використанням 2D-координат, особливо при мануальній оцінці в реальному часі. Це підкреслює важливість тривимірної інформації для розпізнавання жестів, оскільки вона дозволяє краще враховувати

глибину та складні рухи рук, що є критично важливим для динамічних жестів. Той факт, що розроблена модель на 3D-даних перевершила результати попередніх досліджень на цьому ж наборі даних, які використовували лише колірну інформацію, свідчить про ефективність обраного підходу до обробки скелетних даних.

Аналіз матриці заплутаності для NVGesture виявив певні складнощі моделі з розрізненням жестів, що виконуються схожими рухами пальців, наприклад, "MOVE_2_FINGERS_UP" та "MOVE_2_FINGERS_DOWN". Це вказує на необхідність подальшого вдосконалення моделі в частині розрізнення тонких моторних рухів. Мануальна оцінка також виявила, що модель, навчена на 2D-даних NVGesture, гірше узагальнюється на реальні сценарії порівняно з моделлю на 3D-даних, що підкреслює важливість використання тривимірних даних для практичних застосувань.

Високі результати, отримані на наборі даних SHREC22 з використанням 3D-даних, підтверджують, що якісні тривимірні дані значно покращують здатність моделі до розпізнавання як статичних, так і динамічних жестів. Відсутність можливості порівняння з 2D-даними на цьому наборі та проведення мануального тестування є обмеженням, пов'язаним зі специфікою збору даних SHREC22.

Елементи наукової новизни: адаптація архітектури DD-Net для розпізнавання жестів рук у реальному часі з використанням як 2D, так і 3D даних, отриманих за допомогою фреймворку MediaPipe; комплексне оцінювання розробленої моделі на двох різних наборах даних (NVGesture та SHREC22) з використанням як кількісних метрик, так і якісної оцінки в реальному часі; порівняльний аналіз ефективності розпізнавання жестів на основі 2D та 3D скелетних даних, що виявив кращу узагальненість 3D-даних у реальних умовах.

Практична значущість: розроблена система розпізнавання жестів у реальному часі може бути використана в різноманітних застосуваннях, включаючи безконтактне керування пристроями, інтерфейси віртуальної та доповненої реальності, а також системи допомоги для людей з обмеженими можливостями. Результати дослідження підтверджують переваги використання 3D-даних для підвищення точності та надійності розпізнавання жестів, що є важливим для практичного впровадження таких систем. Виявлені обмеження моделі, такі як труднощі з розрізненням схожих жестів, вказують на напрями для подальшого вдосконалення систем розпізнавання жестів.

Висновки та перспективи подальших досліджень. У роботі представлено розробку та оцінку моделі розпізнавання жестів рук у реальному часі. Експерименти з використанням наборів даних NVGesture та SHREC22 продемонстрували ефективність підходу, особливо при використанні 3D-даних, які забезпечили вищу точність та кращу узагальненість у реальних умовах. Модель показала високі результати на наборі даних SHREC22, що містить якісні тривимірні дані про жести. Незважаючи на досягнуті успіхи, існують певні обмеження, зокрема труднощі з розрізненням схожих жестів та статичних жестів з незначними динамічними рухами, що вказує на необхідність подальшої оптимізації алгоритмів класифікації. Отримані результати підтверджують потенціал розробленої системи для застосування в різних сферах, де потрібна інтуїтивна та безконтактна взаємодія. У майбутніх дослідженнях планується зосередитися на підвищенні робастності моделі до індивідуальних варіацій виконання жестів та розширенні її можливостей для розпізнавання більш складних та контекстно-залежних жестів.

Список бібліографічного опису

1. Iqbal, M.Z., Campbell, A.G. (2021). From luxury to necessity: Progress of touchless interaction technology. *Technology in Society*. Volume 67, November 2021, 101796. <https://doi.org/10.1016/j.techsoc.2021.101796>
2. Teslyuk, V., Chornenkyi, V., Tsapiv, V., & Kazymyra, I. (2024). Investigating hand gesture recognition models: 2D CNNs vs. visual transformers. *CEUR Workshop Proceedings*. Vol. 3688: Proc. of the 8th Int. conf. on computational linguistics and intelligent systems. Vol. III: Intelligent systems workshop. Lviv, Ukraine, Apr. 12-13, 2024, 29–38. CEUR-WS.org.
3. Tsapiv, V., & Kazymyra, I. (2024). Gesture Control in Real-time: a framework for 3D hand pose recognition for smart devices and touchless interfaces. In *1st International Scientific and Practical Conference on Computational Intelligence and Smart Systems*, 126–128.
4. Zhang, W., Lin, Z., Cheng, J., Ma, C., Deng, X., & Wang, H. (2020). STA-GCN: two-stream graph convolutional network with spatial-temporal attention for hand gesture recognition. *The Visual Computer*, 36(10–12), 2433–2444. <https://doi.org/10.1007/s00371-020-01955-w>.
5. Do, N.-T., Kim, S.-H., Yang, H.-J., & Lee, G.-S. (2020). Robust hand shape features for Dynamic Hand Gesture Recognition using multi-level feature LSTM. *Applied Sciences* (Basel, Switzerland), 10(18), 6293. <https://doi.org/10.3390/app10186293>
6. Yang, Fan, Yang Wu, Sakriani Sakti, and Satoshi Nakamura. (2019). Make Skeleton-Based Action Recognition Model Smaller, Faster and Better. In *Proceedings of the ACM Multimedia Asia*. New York, NY, USA: ACM.

7. Emporio, M., Caputo, A., Giachetti, A., Cristani, M., Borghi, G., D'Eusario, A., Le, M.-Q., Nguyen, H.-D., Tran, M.-T., Ambellan, F., Hanik, M., Nava-Yazdani, E., & von Tycowicz, C. (2022). SHREC 2022 track on online detection of heterogeneous gestures. *Computers & Graphics*, 107, 241–251. <https://doi.org/10.1016/j.cag.2022.07.015>
8. Vaezi Joze, Hamid Reza, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. (2020). MMTM: Multimodal Transfer Module for CNN Fusion. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
9. Maghoumi, Mehran, and Joseph J. LaViola Jr. (2019). DeepGRU: Deep Gesture Recognition Utility. In *Lecture Notes in Computer Science*, 16–31. Cham: Springer International Publishing.
10. Caputo, F.M., Burato, S., Pavan, G., Voillemin, T., Wannous, H., Vandeboire, J., Maghoumi, M., Taranta, E.M., A., Razmjoo, Laviola, J.J., Manganaro, F., Pini, S., Borghi, G., Vezzani, R., Cucchiara, R., Nguyen, H., Tran, M., & Giachetti, A. (2019). SHREC 2019 Track: Online Gesture Recognition.
11. Song, J.-H., Kong, K., & Kang, S.-J. (2022). Dynamic hand gesture recognition using improved spatio-temporal graph convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology: A Publication of the Circuits and Systems Society*, 32(9), 6227–6239. <https://doi.org/10.1109/tcsvt.2022.3165069>
12. Miah, A. S. M., Hasan, M. A. M., & Shin, J. (2023). Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model. *IEEE Access: Practical Innovations, Open Solutions*, 11, 4703–4716. <https://doi.org/10.1109/access.2023.3235368>
13. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C., & Grundmann, M. (2020). MediaPipe Hands: On-device Real-time Hand Tracking. ArXiv, abs/2006.10214.
14. *HoloLens 2 gestures for authoring and navigating in Dynamics 365 Guides*. (n.d.). Microsoft.com. Retrieved November 15, 2024, from <https://learn.microsoft.com/en-us/dynamics365/mixed-reality/guides/authoring-gestures-hl2>
15. *Oculus quest 2 review*. (n.d.) GSMArena. Retrieved November 25, 2024, from https://www.gsmarena.com/oculus_quest_2_review-news-46255.php
16. Bharatesh, C., Prabhu, P. B. M., & Pavan, K. B. N. (2023). A comprehensive review of leap motion controller-based hand gesture datasets. In arXiv [cs.HC]. <http://arxiv.org/abs/2311.04373>
17. Małeck, K., Nowosielski, A., & Kowalicki, M. (2020). Gesture-based user interface for vehicle on-board system: A questionnaire and research approach. *Applied Sciences (Basel, Switzerland)*, 10(18), 6620. <https://doi.org/10.3390/app10186620>
18. Quentin de Smedt, Hazem Wannous, Jean-Philippe Vandeboire, Joris Guerry, Bertrand Le Saux, et al.. SHREC'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset. 3DOR - 10th Eurographics Workshop on 3D Object Retrieval, Apr 2017, Lyon, France. pp.1-6, 10.2312/3dor.20171049_xffff_xffff_hal-01563505_xffff
19. Li, R., & Wang, X. (2017). Self-adaptive weighted majority vote algorithm based on entropy. *2017 2nd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, 73–77.
20. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., & Kautz, J. (2016). Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4207–4215.
21. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

References

1. Iqbal, M.Z., Campbell, A.G. (2021). From luxury to necessity: Progress of touchless interaction technology. *Technology in Society*. Volume 67, November 2021, 101796. <https://doi.org/10.1016/j.techsoc.2021.101796>
2. Teslyuk, V., Chornenkyi, V., Tsapiv, V., & Kazymyrya, I. (2024). Investigating hand gesture recognition models: 2D CNNs vs. visual transformers. *CEUR Workshop Proceedings*. Vol. 3688: Proc. of the 8th Int. conf. on computational linguistics and intelligent systems. Vol. III: Intelligent systems workshop. Lviv, Ukraine, Apr. 12-13, 2024, 29–38. CEUR-WS.org.
3. Tsapiv, V., & Kazymyrya, I. (2024). Gesture Control in Real-time: a framework for 3D hand pose recognition for smart devices and touchless interfaces. In *1st International Scientific and Practical Conference on Computational Intelligence and Smart Systems*, 126–128.
4. Zhang, W., Lin, Z., Cheng, J., Ma, C., Deng, X., & Wang, H. (2020). STA-GCN: two-stream graph convolutional network with spatial-temporal attention for hand gesture recognition. *The Visual Computer*, 36(10–12), 2433–2444. <https://doi.org/10.1007/s00371-020-01955-w>
5. Do, N.-T., Kim, S.-H., Yang, H.-J., & Lee, G.-S. (2020). Robust hand shape features for Dynamic Hand Gesture Recognition using multi-level feature LSTM. *Applied Sciences (Basel, Switzerland)*, 10(18), 6293. <https://doi.org/10.3390/app10186293>
6. Yang, Fan, Yang Wu, Sakriani Sakti, and Satoshi Nakamura. (2019). Make Skeleton-Based Action Recognition Model Smaller, Faster and Better. In *Proceedings of the ACM Multimedia Asia*. New York, NY, USA: ACM.
7. Emporio, M., Caputo, A., Giachetti, A., Cristani, M., Borghi, G., D'Eusario, A., Le, M.-Q., Nguyen, H.-D., Tran, M.-T., Ambellan, F., Hanik, M., Nava-Yazdani, E., & von Tycowicz, C. (2022). SHREC 2022 track on online detection of heterogeneous gestures. *Computers & Graphics*, 107, 241–251. <https://doi.org/10.1016/j.cag.2022.07.015>
8. Vaezi Joze, Hamid Reza, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. (2020). MMTM: Multimodal Transfer Module for CNN Fusion. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
9. Maghoumi, Mehran, and Joseph J. LaViola Jr. (2019). DeepGRU: Deep Gesture Recognition Utility. In *Lecture Notes in Computer Science*, 16–31. Cham: Springer International Publishing.
10. Caputo, F.M., Burato, S., Pavan, G., Voillemin, T., Wannous, H., Vandeboire, J., Maghoumi, M., Taranta, E.M., A., Razmjoo, Laviola, J.J., Manganaro, F., Pini, S., Borghi, G., Vezzani, R., Cucchiara, R., Nguyen, H., Tran, M., & Giachetti, A. (2019). SHREC 2019 Track: Online Gesture Recognition.

11. Song, J.-H., Kong, K., & Kang, S.-J. (2022). Dynamic hand gesture recognition using improved spatio-temporal graph convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology: A Publication of the Circuits and Systems Society*, 32(9), 6227–6239. <https://doi.org/10.1109/tcsvt.2022.3165069>
12. Miah, A. S. M., Hasan, M. A. M., & Shin, J. (2023). Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model. *IEEE Access: Practical Innovations, Open Solutions*, 11, 4703–4716. <https://doi.org/10.1109/access.2023.3235368>
13. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C., & Grundmann, M. (2020). MediaPipe Hands: On-device Real-time Hand Tracking. ArXiv, abs/2006.10214.
14. *HoloLens 2 gestures for authoring and navigating in Dynamics 365 Guides*. (n.d.). Microsoft.com. Retrieved November 15, 2024, from <https://learn.microsoft.com/en-us/dynamics365/mixed-reality/guides/authoring-gestures-hl2>
15. *Oculus quest 2 review*. (n.d.) GSMArena. Retrieved November 25, 2024, from https://www.gsmarena.com/oculus_quest_2_review-news-46255.php
16. Bharatesh, C., Prabhu, P. B. M., & Pavan, K. B. N. (2023). A comprehensive review of leap motion controller-based hand gesture datasets. In arXiv [cs.HC]. <http://arxiv.org/abs/2311.04373>
17. Małeck, K., Nowosielski, A., & Kowalicki, M. (2020). Gesture-based user interface for vehicle on-board system: A questionnaire and research approach. *Applied Sciences (Basel, Switzerland)*, 10(18), 6620. <https://doi.org/10.3390/app10186620>
18. Quentin de Smedt, Hazem Wannous, Jean-Philippe Vandeborre, Joris Guerri, Bertrand Le Saux, et al.. SHREC'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset. 3DOR - 10th Eurographics Workshop on 3D Object Retrieval, Apr 2017, Lyon, France. pp.1-6, 10.2312/3dor.20171049_xffff_. _xffff_ hal-01563505_xffff_
19. Li, R., & Wang, X. (2017). Self-adaptive weighted majority vote algorithm based on entropy. *2017 2nd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, 73–77.
20. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., & Kautz, J. (2016). Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4207–4215.
21. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.