

РОЗПІЗНАВАННЯ КОНФІГУРАЦІЙ РУК УКРАЇНСЬКОЇ ЖЕСТОВОЇ МОВИ НА ІЗОЛЬОВАНИХ ВІДЕО

К. Д. Грішин^{1,а}, Н. І. Недашківська^{1,б}

¹ Навчально-науковий інститут прикладного системного аналізу, КПІ ім. Ігоря Сікорського

Анотація

Автоматичне розпізнавання жестової мови на основі штучного інтелекту є актуальною проблемою. Для підвищення точності таких систем доцільне використання додаткових ознак, зокрема конфігурацій рук. У роботі представлено нейронну мережу прямого поширення на основі повнозв'язних шарів для ідентифікації конфігурацій української жестової мови (УЖМ) у відеопотоці. Запропоновано додавати до вхідних даних математичні показники, що обчислюються на основі координат ключових точок, отриманих системою захоплення пози MediaPipe. Для навчання моделей зібрано відео ізолюваних жестів УЖМ з Інтернету вебскрепінгом, частину відео знято самостійно на камеру. Експерименти підтвердили високу точність ідентифікації конфігурацій. Нові ознаки суттєво підвищували узагальнювальну здатність нейронних мереж порівняно з моделями, що використовували лише координати ключових точок. Розроблена система може слугувати компонентом для побудови комплексних архітектур розпізнавання жестового мовлення.

Ключові слова: українська жестова мова, конфігурація, машинне навчання, нейронні мережі, комп'ютерний зір, розпізнавання жестів рук, оцінка якості моделі, інженерія ознак, системний аналіз.

Вступ

Понад 70 мільйонів осіб у світі використовують більше ніж 300 різних жестових мов. Станом на 2025 рік в Україні близько 2 млн громадян мають порушення слуху, для понад 230 тисяч осіб українська жестова мова (УЖМ) є рідною. Водночас спостерігається критичний дефіцит фахівців, здатних забезпечити комунікацію між нечуючими та чуючими верствами населення. Зокрема, кількість перекладачів УЖМ, атестованих Українським товариством глухих, становить 165 осіб [1]. Оскільки неможливо забезпечити індивідуальний супровід сурдоперекладача для кожної особи з порушеннями слуху, розроблення автоматизованих систем на основі штучного інтелекту (ШІ) для розпізнавання УЖМ є актуальною проблемою. Попри активну реалізацію подібних проєктів для інших мов, сегмент систем ШІ для підтримки УЖМ залишається недостатньо розробленим [2, 3].

У даній роботі представлено фреймворк для системи розпізнавання жестової мови. На початковому етапі розроблюється модель розпізнавання ізолюваних жестів. Для того, щоб така модель ідентифікувала жести, перед тим необхідно виділити інформативні ознаки. Перспективним дескриптором є конфігурація рук – базова лінгвістична компонента жесту, що визначається специфічним положенням пальців та руки/рук. Будь-який жест однозначно ідентифікується через поєднання конфігурації, локації та трає-

кторії руху [4]. Попереднє визначення конфігурацій, задіяних під час виконання жесту, дозволяє суттєво звужити коло пошуку. У роботі побудовано модель машинного навчання для розпізнавання конфігурацій УЖМ у відеопотоці із застосуванням системи запропонованих математичних показників.

1. Сучасний стан проблеми

1.1. Особливості розпізнавання жестової мови

Проблему розпізнавання жестової мови поділяють на декілька задач: ізолюване розпізнавання жестової мови (ISLR, Isolated Sign Language Recognition), неперервне розпізнавання жестової мови (CSLR, Continuous Sign Language Recognition), переклад жестової мови (SLT, Sign Language Translation).

ISLR – класифікація жестів, коли на кожному відео представлено лише один жест. Її розв'язання недостатньо для побудови перекладача, оскільки в реальних умовах надходить послідовність жестів. Однак це може бути першим етапом побудови системи для перекладу жестової мови в звичайну мову [5].

CSLR – розпізнавання послідовності жестів $\{y_1, \dots, y_m\}$ на відео. Часові межі, протягом яких показується кожен жест y_i , невідомі. Створення датасетів для CSLR складніше, ніж для ISLR, оскільки різко зростає комбінаторика простору ознак. Розпізнана послідовність ще не є перекладом, так як в ній відсутні відмінки, а граматики не відповідає звичайній мові [5].

SLT – переведення відео з послідовністю жестів в

^аconstantine1223h@gmail.com

^бn.nedashkivska@gmail.com

текст звичайною мовою. Її розв'язання здатне зменшити комунікаційний бар'єр між особами з порушеннями слуху та чуючим населенням. Така система повинна не лише розпізнавати послідовності жестів, але й граматично правильно відтворювати повідомлення звичайною мовою. Ця проблема охоплює комп'ютерний зір та машинний переклад. Особливості розглянутих задач підсумовано в таблиці 1.

Вхідними даними для моделей можуть бути відео, оптичний потік, координати ключових точок тіла, отриманих системою захоплення пози. У відкритому доступі відсутні датасети УЖМ для розв'язання будь-якої з зазначених задач [5].

1.2. Конфігурації рук як додаткові ознаки для ізолюваного розпізнавання жестової мови

Конфігурація – найменша структурна одиниця жесту разом з локалізацією та рухом руки/рук, подібно до фонем в словах звичайної мови; характеризується в одnorучних жестах положенням пальців та кисті руки, у дворучних – положенням пальців і кисті кожної руки, взаємним розміщенням рук. В УЖМ виділяється 40 конфігурацій. У рамках однієї конфігурації можливо декілька варіантів показу [4]. Під час аналізу відеоматеріалів УЖМ виділено наступні особливості:

- при виконанні дворучних жестів ліва та права рука можуть мати різні конфігурації;
- при показі окремих жестів інколи використовуються декілька конфігурацій одночасно.

Конфігурації як додаткові ознаки вперше використано у 2017 р. для датасету RWTH-PHOENIX-Weather-2014 німецької жестової мови. Нейронна мережа, натренована розпізнавати форму руки, генерувала послідовності векторних представлень, які були вхідними даними для моделі. Модель ідентифікації конфігурацій була навчена на понад 1 млн розмічених зображень 60 різних форм рук, зібраних із датської, новозеландської та німецької мов [6]. Подальші модифікації підходу на основі розпізнавання форми рук дозволили підвищити точність [7]. Також ведуться розробки мереж для визначення конфігурацій різних мов, найбільш активно досліджується для американської [8].

1.3. Постановка задачі

Дано: множина конфігурацій УЖМ $\{1, \dots, 40\}$; навчальна вибірка

$$(X_i, y_i), i = \overline{1, N_s} \quad (1)$$

де $X_i \in R^{M \times T}$ – відео із демонстрацією одиничного жесту у вигляді векторів ознак розмірності M ;

T – максимальна кількість кадрів відео (відео меншої довжини доповнюються нулями);

$y_i \in \{1, \dots, 40\}$ – конфігурація правої руки при показі жесту X_i ;

N_s – обсяг навчальної вибірки.

Потрібно: побудувати модель машинного навчання $f: R^{M \times T} \rightarrow \{1, \dots, 40\}$.

Критерії оптимальності: максимізація Ассигасу моделі машинного навчання на перевіірочній вибірці.

Обмеження: в межах жесту одна конфігурація на руку; домінуюча права рука для одnorучних жестів.

Тип задачі: навчання з вчителем; небінарна класифікація.

Аналогічно можна поставити задачу розпізнавання конфігурацій лівої руки. У даній роботі розглядалися одnorучні та дворучні жести. У подальших дослідженнях можна включити жести, в яких більше однієї конфігурації на руку, перейшовши до багатоміткової класифікації. У якості вхідних даних обрано вектори координат ключових точок, отриманих системою захоплення пози Mediapipe.

2. Запропонована система розпізнавання жестів

Традиційні підходи часто спираються на розробку End-to-End рішення для SLT, де моделі глибокого навчання встановлюють відповідність між вхідним відео та відновленим текстом. Однак для УЖМ він малоефективний через відсутність датасетів достатнього обсягу. У роботі пропонується фреймворк для неперервного розпізнавання жестової мови, що базується на декомпозиції задачі. Запропонована архітектура складається з кількох модулів:

1. Система захоплення пози *MediaPipe* – відкрита інфраструктура, розроблена Google, для виявлення ключових точок тіла на основі машинного зору. *MediaPipe* дозволяє фіксувати та аналізувати рухи людини у режимі реального часу кадр за кадром.

Таблиця 1. Задачі розпізнавання жестової мови

Задача розпізнавання жестової мови	Вхід моделі	Вихід моделі	Функціональність
Ізольоване (ISLR)	Відео, що містить одиничний жест	Розпізнаний одиничний жест	Може бути першим етапом побудови системи перекладу жестової мови.
Неперервне (CSLR)	Відео з повідомленням жестовою мовою	Послідовність розпізнаних жестів	Не надає повноцінний переклад, тому що не враховуються відмінки та граматика.
Переклад (SLT)	Відео з повідомленням жестовою мовою	Переклад звичайною мовою	Здатен виконувати функції сурдоперекладача.

Обробка відеозаписів полягає у трансформації відеопотоку в структурований набір даних, що містить координати траєкторій ключових точок. Це забезпечує ефективну фільтрацію неінформативних даних (фон, одяг, тощо), дозволяючи зосередити обчислювальні ресурси виключно на динамічних характеристиках жестового мовлення.

2. *Модуль сегментації* – компонент, що ділить вхідне відео на фрагменти, які відповідають окремим жестам (словам). Його призначення – визначити межі між жестами у відеопотоці. Фрагменти є вхідними даними для моделі ізольованого розпізнавання. Таким чином, задача CSLR зводиться до ISLR. Сегментацію можна реалізувати на основі порогового значення швидкості рухів рук або моделей ШП.

3. *Підсистема розпізнавання конфігурацій* – ідентифікує конфігурації для моделі ізольованого розпізнавання жестів. Приймає на вхід ключові точки та запропоновані математичні показники, які розраховуються відповідним модулем у підсистемі. Застосування конфігурацій підвищує точність моделі [4].

4. *Моделі ізольованого розпізнавання* – ідентифікує жести на фрагментах відео, отриманих модулем сегментації. Використовує розпізнані конфігурації як додаткові ознаки для підвищення точності. Здійсненість моделі ISLR доведена прикладами для різних жестових мов. Зокрема, відомо про успішний приклад побудови нейронної мережі, що розпізнавала 38 жестів УЖМ [9]. Запропонована система показана на рис. 1.

Модульний підхід дозволяє будувати надійні системи розпізнавання жестів навіть в умовах обмежених даних. На виході вона видаватиме послідовність жестів, розв'язуючи задачу CSLR. У подальшому для формування зв'язного тексту на основі результатів роботи системи можна використати великі мовні моделі. Однак це виходить за рамки даного дослідження.

Слід зазначити, що форми руки часто перетинаються для різних жестових мов. Тому модель можна натренувати на зразках датасетів інших мов, щоб далі застосувати для ідентифікації конфігурацій УЖМ [6]. Першим етапом розробки запропонованої системи є побудова системи розпізнавання конфігурацій рук, що розглянуто далі в роботі.

3. Вхідні дані для моделей розпізнавання конфігурацій

У якості вхідних даних обрано ключові точки тіла, отримані системою захоплення пози MediaPipe. Вона використовує моделі комп'ютерного зору без сенсорного обладнання. Всього MediaPipe виділяє 543 точки в тривимірному просторі, що відповідають тулубу, кінцівкам та обличчю (рис. 2). Цей формат забезпечує високу точність та низьку розмірність. Недоліком є поодинокі пропуски в даних, оскільки система не завжди детектує позу при високій швидкості рухів рук.

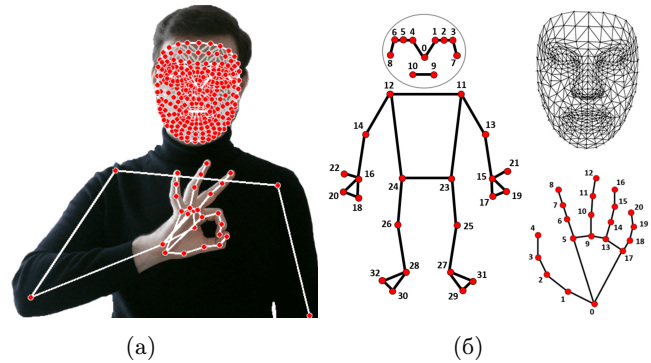


Рис. 2. Ключові точки системи захоплення пози Mediapipe

Навіть після представлення відео у вигляді ключових точок тривимірного простору, розпізнавання форм руки залишається складним через значну кількість класів та варіантів показу однієї конфігурації. Доцільно додати до вхідних даних ознаки, які спростять класифікацію. У роботі запропоновано таку систему математичних показників та досліджено їх ефективність (таблиця 2).

$$Bent(\vec{a}, \vec{b}, \vec{c}) = |\angle(\vec{a}, \vec{b})| + |\angle(\vec{b}, \vec{c})| \quad (2)$$

де $\vec{a}, \vec{b}, \vec{c}$ — вектори, що з'єднують початок та кінець проксимальної, середньої та дистальної фаланги відповідно.

$$Joined(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\| \quad (3)$$

де $\vec{x}, \vec{y}$ — вектори координат кінчиків великого та довільного іншого пальців відповідно;

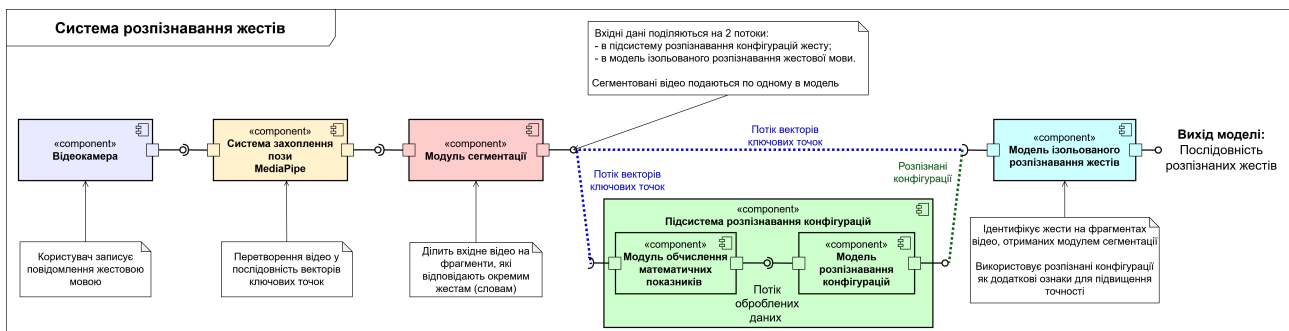


Рис. 1. UML-діаграма компонентів архітектури системи розпізнавання жестів

Таблиця 2. Система математичних показників для вхідних даних

Показник	Опис	Формула / Алгоритм розрахунку	Ілюстрація
<i>Bent</i>	Вимірює «зігнутість» пальця.	Сума модулів кутів між двома парами фаланг пальця. Формула (2).	Рис. 3(а)
<i>Joined</i>	Оцінка відстані між пальцями.	Відстань між кінчиками великого пальця та довільного іншого. Формула (3).	Рис. 3(б)
<i>Angle</i>	Оцінює кут між обраними двома пальцями.	Побудова площин лінійної регресії через 4 точки першого пальця та через 4 точки другого пальця, оцінка кута між їх нормаллями.	Рис. 3(в)
$\overrightarrow{Dir}$	Зберігає інформацію про напрямок пальця.	Різниця координат точки кінчика пальця та точки початку пальця.	Рис. 3(г)
<i>StraightHand</i>	Оцінює близькість форми руки до випрямленої долоні.	Розраховується площиною лінійної регресії через 21 точку руки та обчисленням метрики R^2 . Для прямої долоні R^2 близький до 1.	Рис. 4

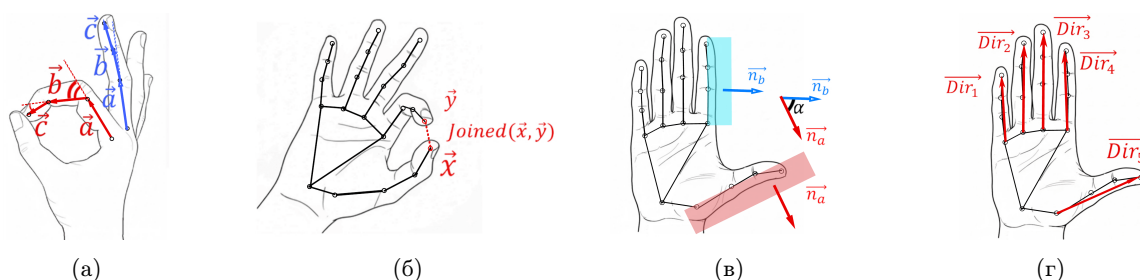


Рис. 3. Приклади розрахунку показників: (а) *Bent* для зігнутого вказівного та випрямленого середнього; (б) *Joined* – відстань між великим та вказівним; (в) *Angle* для великого та вказівного; (г) $\overrightarrow{Dir}$ для кожного пальця.

$\| \cdot \|$ — евклідова відстань між точками.

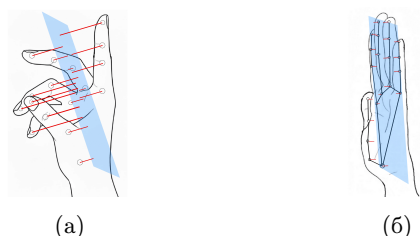


Рис. 4. Розрахунок *StraightHand*: (а) для напівзігнутої руки; (б) для випрямленої руки.

4. Формування та опис датасету

Навчальна вибірка складається з двох типів відеоматеріалу:

- зібраного з мережі Інтернет вебскрепінгом;
- знятого самостійно на камеру.

Вебскрепінг проводився на ресурсах, що не декларували заборону на використання свого матеріалу в дослідницьких цілях: жестові словники, YouTube-канали. Більшість відео містили демонстрацію жестів із підписом на екрані. Виокремлення ізольованих жестів здійснювалося розробленим програмним забезпеченням. Система фіксувала ключові точки через MediaPipe, визначала момент опускання рук (характерна ознака завершення жесту для більшості

зразків), вилучаючи фрагмент. Назва файлу генерувалася розпізнаванням тексту за допомогою бібліотеки Tesseract. Отриманий матеріал перевірено та, за наявності помилок сегментації, скориговано. Відеоматеріал, знятий самостійно, отримано власним програмним забезпеченням; варіювалась швидкість жестикулювання, відстань до камери. У частині записів була присутня артикуляція. Через MediaPipe на відео визначено ключові точки. Кожну точку центровано та нормовано за алгоритмом:

- розраховано умовний «центр тіла» – середнє координат точок 23 і 24 пози (рис. 2(6));
- розраховано розмах плеч – евклідову відстань між точками 11 і 12 пози (рис. 2(6));
- від усіх ключових точок віднято «центр тіла» та поділено на розмах плеч.

У вхідні дані включались наступні значення:

- ключові точки правої та лівої рук (21 точка на руку, 63 координати в тривимірному просторі);
- середнє точок пози, середнє точок обличчя (як «точки відліку»);
- точки губ (для аналізу артикуляції).

Було встановлено обмеження на довжину відео в 30 кадрів. Якщо відео містило менше кадрів, залишок заповнювався нульовими значеннями; якщо більше – відео обрізалася. Жести згруповано за конфігураціями правої руки на основі класифікації Кульбіди С.В. [4]. Після групування отримано вибірку

вигляду (1) обсягом 2917 зразків. Діаграма розподілу кількості зображень для 40 класів у датасеті показано на рис. 5. Нерівномірний розподіл зумовлений різною частотою використання конфігурацій в жестах.

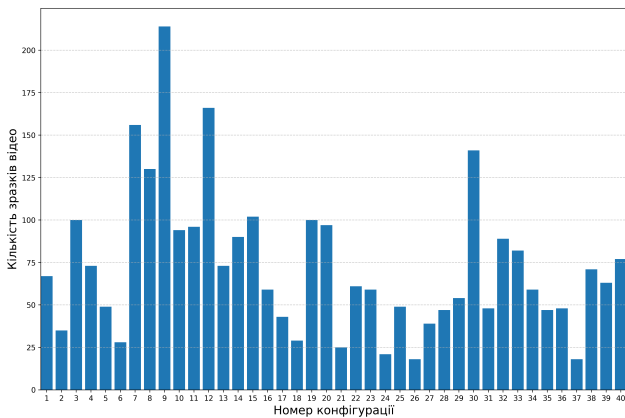


Рис. 5. Діаграма розподілу кількості зображень для 40 класів у датасеті

5. Експерименти та результати

Для оцінки ефективності запропонованих показників проведені експерименти із розпізнавання конфігурацій руки. Порівнювались моделі з двома варіантами вхідних даних:

- необроблені – нормовані координати ключових точок правої та лівої рук, губ, середнє точок обличчя, середнє точок пози;
- оброблені – необроблені дані з додаванням вектора запропонованих показників

$$\vec{X} = [Bent_i, \dots, Joined_j, \dots, Angle_k, \dots, StraightHand, Dir_{mx}, Dir_{my}, Dir_{mz}, \dots] \in \mathbb{R}^{29} \quad (4)$$

де $Bent_i, i = \overline{1, 5}$ – показники зігнутості для 5 пальців;

$Joined_j, j = \overline{1, 4}$ – показники з'єднаності із великим пальцем інших пальців;

$Angle_k, k = \overline{1, 4}$ – кути між великим і вказівним, вказівним і середнім, середнім і безіменним, безіменним і мізинцем;

$StraightHand$ – оцінка наскільки рука пряма;

$Dir_{mx}, Dir_{my}, Dir_{mz}, m = \overline{1, 5}$ – координати напрямку пальця m .

У вибірці на основі ключових точок рук програмним забезпеченням розраховано відповідні вектори (4). Для розпізнавання конфігурацій обрано мережу прямого поширення, що складається з блоків. Кожен блок має наступну структуру:

- повнозв'язний шар;
- шар батч-нормалізації;
- функція активації ReLU.

Із кожним блоком кількість нейронів у повнозв'язному шарі зменшувалась удвічі. Вихідний шар містив 40 нейронів із функцією активації Softmax. Модель побудовано на основі бібліотеки

TensorFlow Keras. Гіперпараметри моделей відображені в таблиці 3. Варіювалась кількість блоків та число нейронів першого блоку (параметри, що повністю визначають архітектуру). Ассигасу випробуваних архітектур занесено в таблицю 4. Зеленим виділено кращі результати для оброблених та необроблених даних.

Таблиця 3. Гіперпараметри моделей

Гіперпараметр	Значення
Відношення навчальної до перевірконої	8 : 2 (стратифікований)
Learning rate	0.0001 із фактором редукції 0.8 через кожні 4 епохи при відсутності покращення
Розмір батчу	64
Оптимізатор	Adam
Функція втрат	Крос-ентропія
Dropout	0.2
Кількість епох	200
Patience перед ранньою зупинкою	50
Прискорювач	GPU

Експерименти свідчать, що математичні показники значно підвищували точність моделей. Кращою на перевірконій вибірці (82, 88%) серед моделей, що приймали на вхід оброблені дані, була чотирьохблокова архітектура, де у першому повнозв'язному шарі 1024 нейронів. Для необроблених даних, кращою виявилась двоблокова модель з початковим шаром на 2048 нейронів (72, 95%).

Зафіксовано розрив між результатами на навчальній та перевірконій вибірках, що вказує на перенавчання. Попри це, мережі успішно ідентифікували конфігурації. Зменшення похибки можливо досягнути збільшенням даних. Залежність точності кращої архітектури від обсягу вибірки показано на рис. 6. Очікується, що модель генеруватиме інформативні ознаки для більш складних моделей.

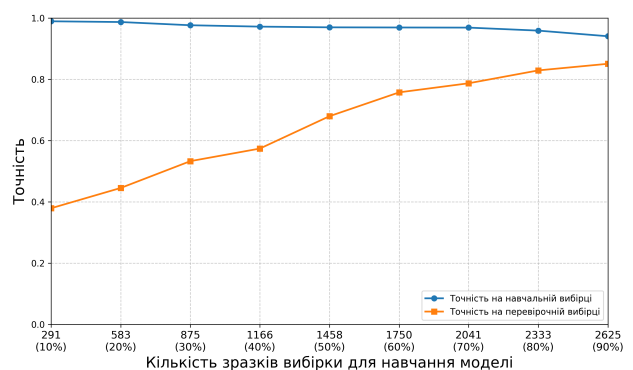


Рис. 6. Залежність точності моделі від обсягу навчальної вибірки (у відсотках вказано доля від загального обсягу вибірки)

Таблиця 4. Результати класифікації конфігурацій моделями

Кількість блоків у мережі	Число нейронів у початковому шарі	Необроблені вхідні дані		Оброблені вхідні дані	
		Точність на навчальній вибірці	Точність на перевіірочній вибірці	Точність на навчальній вибірці	Точність на перевіірочній вибірці
1	512	98,37%	70,55%	99,83%	81,68%
1	1024	99,27%	69,35%	99,96%	80,82%
1	2048	99,61%	69,18%	100,00%	81,34%
2	512	92,71%	70,03%	98,89%	82,19%
2	1024	97,04%	71,92%	99,70%	81,68%
2	2048	98,67%	72,95%	99,87%	82,36%
3	512	76,68%	67,81%	94,04%	81,85%
3	1024	87,66%	71,23%	98,89%	82,36%
3	2048	93,57%	71,40%	99,57%	81,85%
4	512	45,95%	52,05%	79,00%	76,20%
4	1024	73,00%	65,75%	95,84%	82,88%
4	2048	89,58%	72,43%	97,56%	82,02%

Висновки

У роботі натреновано нейронну мережу прямого поширення на основі повнозв'язних шарів, що розпізнає з високою точністю 40 конфігурації УЖМ на відео із ізольованими жестами. У якості вхідних даних обрано ключові точки тіла, отримані системою захоплення пози MediaPipe до яких додається запропонована система математичних показників. Показники обчислюються на основі 63 координат ключових точок руки/рук. Ці ознаки зберігають інформацію про розташування пальців, забезпечуючи інтерпретованість та низьку розмірність даних. Для створення моделі сформовано власну вибірку ізольованих жестів УЖМ; варіювались демонстранти, відстань до камери, швидкість рухів рук. Випробувались різні архітектури із додаванням до вхідних даних запропонованих математичних показників та без них. Показники суттєво підвищували точність нейронних мереж порівняно з такими, що приймали на вхід лише координати ключових точок. Створена модель ефективно розпізнає конфігурації рук УЖМ на відеопотоці, задовольняючи вимогам високої точності та застосування в реальному часі. У подальших роботах планується її інтеграція до системи ізольованого розпізнавання жестової мови.

Перелік використаних джерел

1. Українське товариство глухих. Реєстр перекладачів жестової мови УТОГ. — 2026. — URL: <https://utog.org/perekladachi/reestr-perekladachiv/>; Дата звернення: 15.04.2026.
2. SignGPT Project Team. SignGPT: Building Generative Predictive Transformers for Sign Language – An EPSRC Programme Grant (1/5/2025–30/4/2031). — 2025. — URL: <https://sites.google.com/view/signgpt/team> (visited on 04/15/2026).
3. Google DeepMind. SignGemma Announcement. — 2024. — URL: <https://x.com/GoogleDeep>

[Mind/status/1927375853551235160](https://doi.org/10.1016/j.iswa.2021.200056) (visited on 04/15/2026).

4. Кульбіда С. В. Конфігурація як компонент жести // Особлива дитина: навчання і виховання. — 2018. — № 4. — С. 15—30.
5. Adeyanju I., Bello O., Adegboye M. Machine learning methods for sign language recognition: A critical review and analysis // Intelligent Systems with Applications. — 2021. — Vol. 12. — P. 200056. — DOI: [10.1016/j.iswa.2021.200056](https://doi.org/10.1016/j.iswa.2021.200056).
6. SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition / N. C. Camgoz, S. Hadfield, O. Koller, R. Bowden // Proceedings of the IEEE International Conference on Computer Vision (ICCV). — Venice, Italy, 2017. — P. 3075–3084. — DOI: [10.1109/ICCV.2017.332](https://doi.org/10.1109/ICCV.2017.332).
7. Zhang X., Kevin D. Handshape-Aware Sign Language Recognition: Extended Datasets and Exploration of Handshape-Inclusive Methods // Findings of the Association for Computational Linguistics: EMNLP 2023. — Singapore, 2023. — P. 2993–3002. — DOI: [10.18653/v1/2023.findings-emnlp.198](https://doi.org/10.18653/v1/2023.findings-emnlp.198).
8. Carbo A., Nalisnick E. Improving Handshape Representations for Sign Language Processing: A Graph Neural Network Approach // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Suzhou, China, 2025. — P. 29122–29135. — DOI: [10.18653/v1/2025.emnlp-main.1483](https://doi.org/10.18653/v1/2025.emnlp-main.1483).
9. Бурхан І. О., Барабанов О. В. Розпізнавання української жестової мови на основі ключових точок з використанням рекурентних нейронних мереж // Наука і техніка сьогодні. — 2023. — 4(18). — С. 280—292. — DOI: [10.52058/2786-6025-2023-4\(18\)-280-292](https://doi.org/10.52058/2786-6025-2023-4(18)-280-292).