РОЗПІЗНАВАННЯ ЖЕСТІВ УКРАЇНСЬКОЇ ДАКТИЛЬНОЇ АБЕТКИ

Виконав: студент 2-го року навчання
освітньої програми «Прикладна математика», 113
магістр, Бікчентаєв М. О.
Керівник: доктор технічних наук
Глибовець А. М.

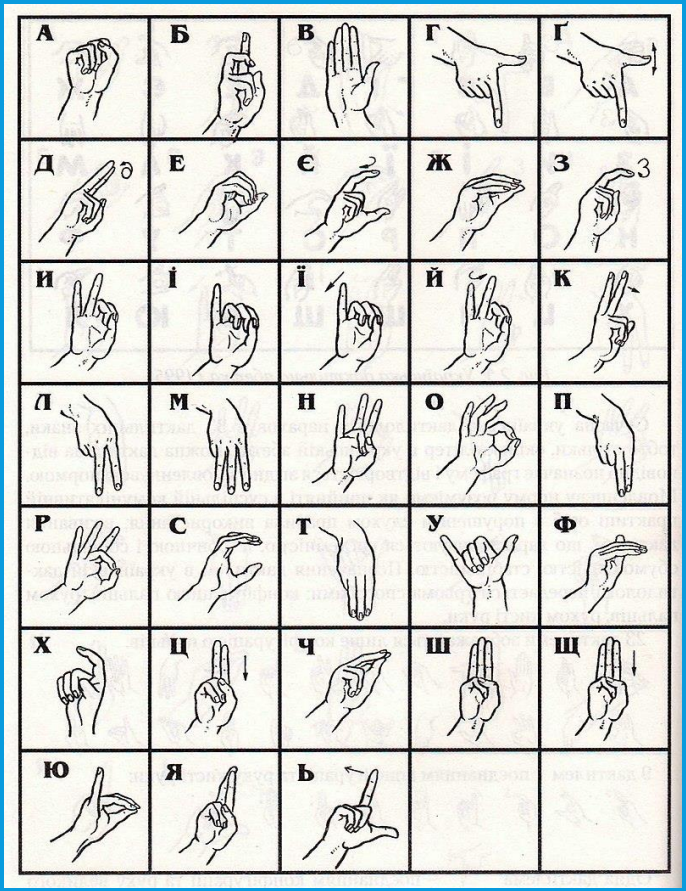
МЕТА РОБОТИ

- Проаналізувати алгоритми розпізнавання жестів та, використовуючи методи машинного навчання, побудувати штучну нейронну мережу для розпізнавання жестів української дактильної абетки із вхідного потоку зображень з камери без використання додаткового обладнання (сенсорів або рукавичок).

УКРАЇНСЬКА ДАКТИЛЬНА АБЕТКА

- *Українська дактильна абетка* - допоміжна система української жестової мови, в якій кожному жесту однієї руки відповідає літера української абетки.
- Використовується для вимови допоміжних слів, слів, які не мають жестового позначення, а також коли необхідно уточнити значення певного слова.
- Налічує 33 дактильні знаки (або *дактилеми*), що відповідає кількості літер в українській абетці.

УКРАЇНСЬКА ДАКТИЛЬНА АБЕТКА



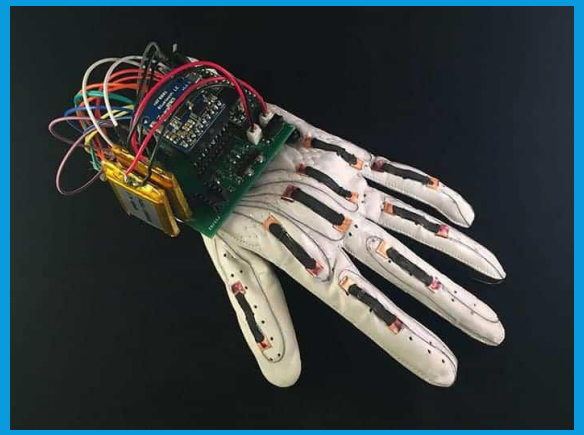
Українська дактильна абетка

23 дактилеми є статичними і потребують лише певної конфігурації пальців для зображення.

10 дактилем є динамічними та окрім, конфігурації пальців, потребують руху п’ястку руки або пальців.

GLOVE-BASED РОЗПІЗНАВАННЯ ЖЕСТІВ

- Існує два підходи до розпізнавання жестів: *glove-based* (на основі рукавичок), який передбачає носіння рукавичок з датчиками, та *CV-based* (на основі комп'ютерного зору), який використовує методи комп'ютерного зору і не вимагає носіння жодних датчиків.
- Рукавички з датчиками можуть легко надати точні координати розташування долоні та пальців і дані про їхню орієнтацію в просторі. Проте, цей підхід повністю залежить від спеціалізованих датчиків, які можуть бути досить дорогими у придбанні та обслуговуванні.



Приклад рукавички з датчиками

CV-BASED РОЗПІЗНАВАННЯ ЖЕСТІВ

- CV-based підхід не вимагає спеціального обладнання, окрім веб-камери, однак пов'язаний з певними труднощами, такими як варіації освітлення, складні фони, зображення чи відео низької якості, або оклюзія.
- Ми можемо розглядати розпізнавання жестів рук як задачу з розпізнвання образів. Для вирішення цієї задачі можна скористатися *підходами на основі машинного навчання* або ж *підходами на основі глибокого навчання*.

CV-BASED РОЗПІЗНАВАННЯ ЖЕСТІВ

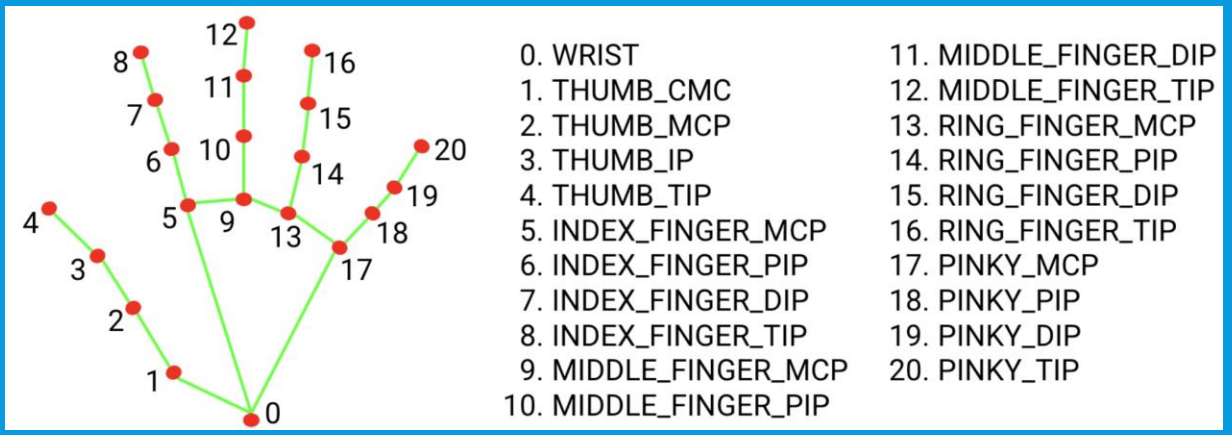
- *Підходи на основі машинного навчання* використовують особливості зображення, такі як колір або контури об'єктів, щоб ідентифікувати групи пікселів, які можуть належати жесту.
- Ця інформація потім використовуються алгоритмами машинного навчання, на кшталт SVM (support vector machine) або RDF (random decision forest), для класифікації зображеного жесту.
- Ефективність підходів, заснованих на машинному навчанні, обмежена якістю особливостей зображення, що були обрані для класифікації жесту.
- *Підходи на основі глибокого навчання*, з іншого боку, використовують нейронні мережі для автоматичного виявлення необхідних особливостей зображення, що призводить до значного підвищення точності системи розпізнавання образів.

CV-BASED РОЗПІЗНАВАННЯ ЖЕСТІВ

- Оскільки методи глибокого навчання дають гарні результати у задачі розпізнавання образів та є найбільш популярними у сфері комп'ютерного зору, в цій роботі буде використано дві глибокі нейронні мережі.
- *Перша*, попередньо навчена мережа, під назвою *HandLandmarker*, буде використана для виявлення руки на зображенні та отримання її орієнтирів.
- *Друга*, використовуючи ці орієнтири, буде виконувати класифікацію жесту.

GOOGLE MEDIA PIPE

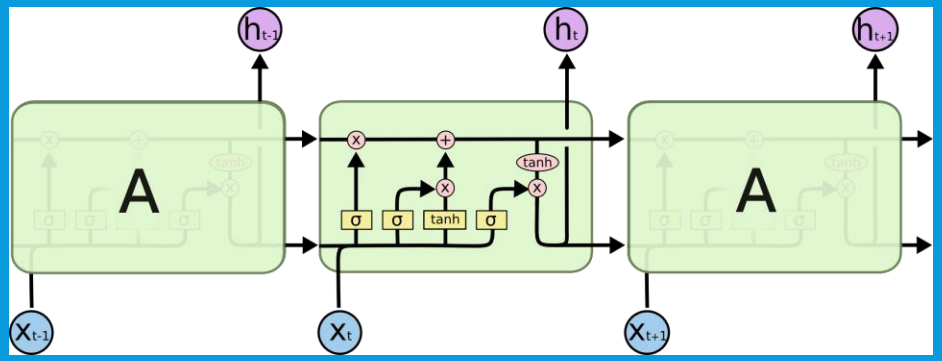
- *Google MediaPipe* – це набір інструментів та бібліотек, що дозволяють застосувати методи машинного навчання для вирішення таких задач як розпізнавання обличь, орієнтирів рук чи пози людини.
- Модель для розпізнавання орієнтирів рук, що має назву HandLandmarker, приймає на вхід зображення та повертає список з 21-ї точки. Положення кожної точки описується трьома координатами.



Назви точок та їх розташування відносно одна одної

LSTM

- LSTM (Long Short-Term Memory) мережа - це різновид рекурентних нейронних мереж (RNN), що здатна вивчати довгострокові залежності між вхідними даними та здатна обробляти всю послідовність даних, наприклад, відео, а не окремі елементи, наприклад, відеокадри.
- LSTM складається з повторюваних модулів, що пов'язані між собою. Усі модулі мають однакову структуру.
- Оскільки результатом роботи HandLandmarker є список орієнтирів, LSTM буде становити основу класифікатора.



Структура LSTM

СТВОРЕННЯ НАБОРУ ДАНИХ

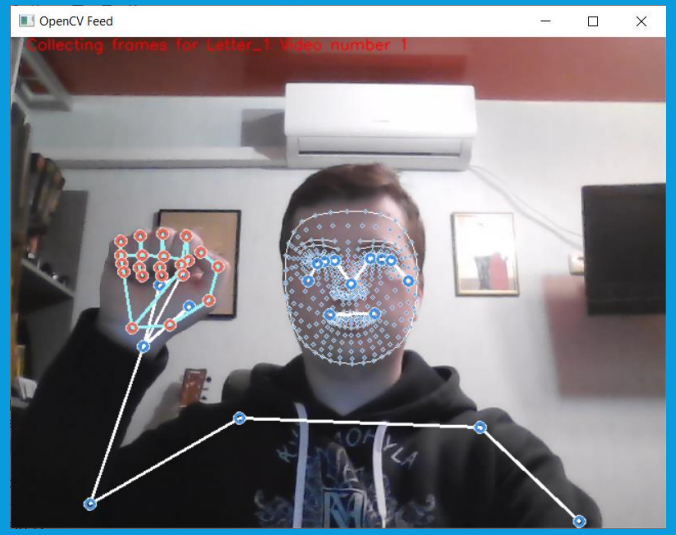
- Глибокі нейронні мережі вимагають значної кількості даних для їх тренування.
- Для української дактильної абетки відсутні набори даних які можна було б використати для тренування моделі, тому було вирішено створити власний набір.
- Отриманий набір даних складається з 33 класів. Кожному класу відповідає 50 відеозаписів, які в подальшому будуть розбиті на 65 кадрів.

СТВОРЕННЯ НАБОРУ ДАНИХ

- Для створення набору даних спочатку було вирішено використати відео з YouTube. Проте, не дивлячись на те, що якість знайдених відео була високою, їх кількість була замалою для тренування мережі.
- Загалом було використано 6 відеозаписів з YouTube, а також відеозаписи з ресурсу «Spread The Sign».
- Кожен з 6 відеозаписів з YouTube був розділений на відеофрагменти, на яких зображувався той чи інший жест дактильної абетки. Після цього, відеофрагменти були розділені на 65 кадрів; якщо відеофрагмент складався з меншої кількості кадрів – він «зациклювався» додаванням початкових кадрів у кінець фрагменту.

СТВОРЕННЯ НАБОРУ ДАНИХ

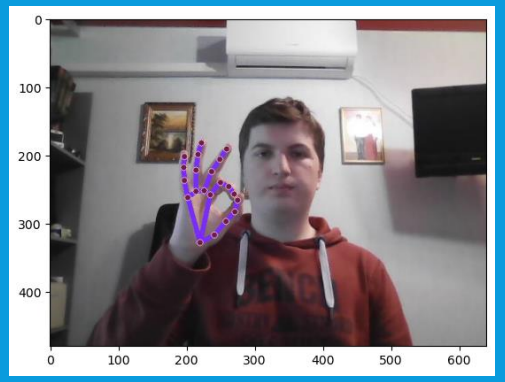
- Для запису решти відео було розроблено програму на мові програмування Python з використанням бібліотек OpenCV та Google MediaPipe.
- Програма використовує веб-камеру для запису відео з 65 кадрів. Після запису відеокадри зберігаються у форматі JPG на диску комп'ютера.

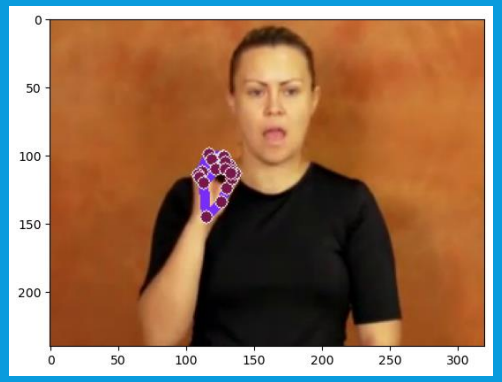


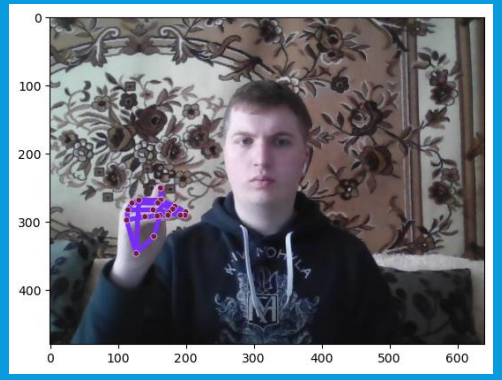
Вікно програми для запису відео.

Після кожного записаного відео, програма робить паузу у 5 секунд щоб можна було підготуватися до запису наступного.

СТВОРЕННЯ НАБОРУ ДАНИХ







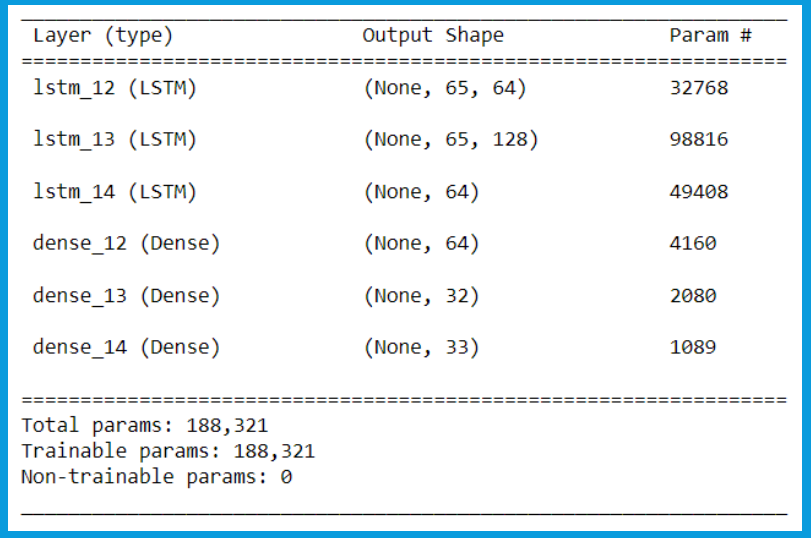
Приклад відеокадрів з набору даних.

На відеокадрах можна побачити орієнтири, що були отримані завдяки Google MediaPipe.

Кожному жесту абетки відповідає 50 відео по 65 кадрів.

ПОБУДОВА МОДЕЛІ

- Модель для класифікації жестів складається з 6 шарів.
- Перші 3 шари використовують LSTM з функцією активації $tanh$ (гіперболічний тангенс). Решта шарів є звичайними щільно зв'язаними (*densely-connected*) шарами, 2 з яких використовують ReLU у якості функції активації, у той час як останній шар використовує Softmax.



Layer (type)	Output Shape	Param #
lstm_12 (LSTM)	(None, 65, 64)	32768
lstm_13 (LSTM)	(None, 65, 128)	98816
lstm_14 (LSTM)	(None, 64)	49408
dense_12 (Dense)	(None, 64)	4160
dense_13 (Dense)	(None, 32)	2080
dense_14 (Dense)	(None, 33)	1089

Total params: 188,321
Trainable params: 188,321
Non-trainable params: 0

Структура моделі для розпізнавання жестів

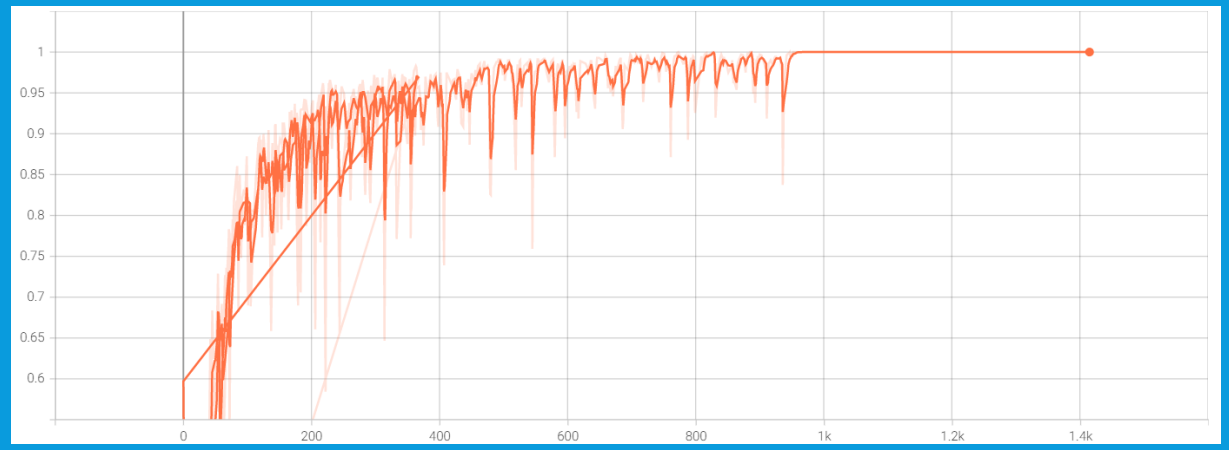
$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$Softmax(y)_i = \frac{\exp(y_i)}{\sum_{j=1}^{n} \exp(y_j)}$$

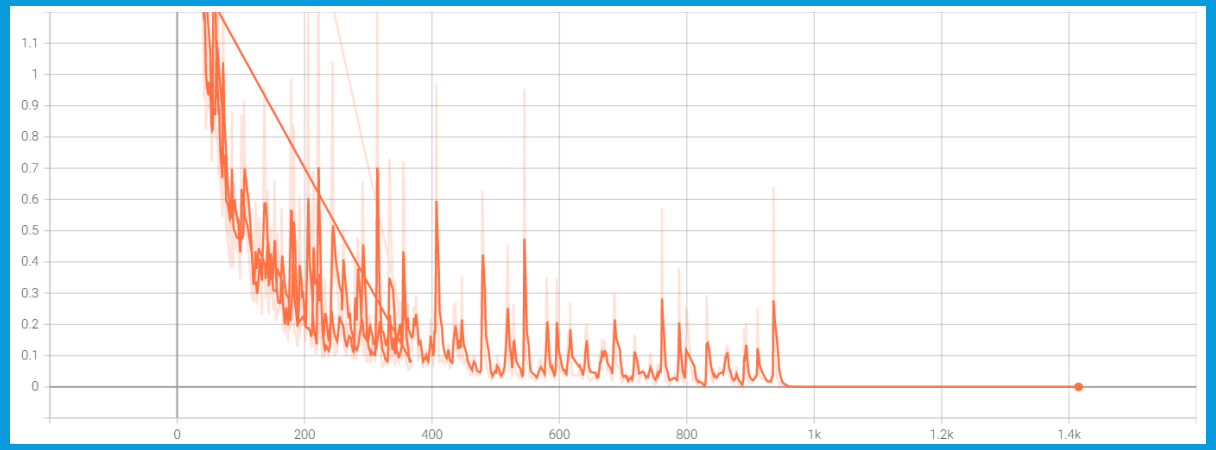
$$ReLU(x) = \max(0, x)$$

ТРЕНУВАННЯ МОДЕЛІ

- Тренування моделі відбувалося на GPU NVIDIA GeForce 1660 Ti та тривало 1416 епох, при цьому точність моделі перестала змінюватися після 960 епохи та трималася на рівні 99%.
- Набір даних для тренування становив 80% від усього набору (996 відеозаписів по 65 кадрів; 64 740 відеокадрів загалом).



Зміна точності моделі зі збільшенням кількості епох



Зміна функції втрат моделі зі збільшенням кількості епох

ОЦІНКА ЕФЕКТИВНОСТІ МОДЕЛІ

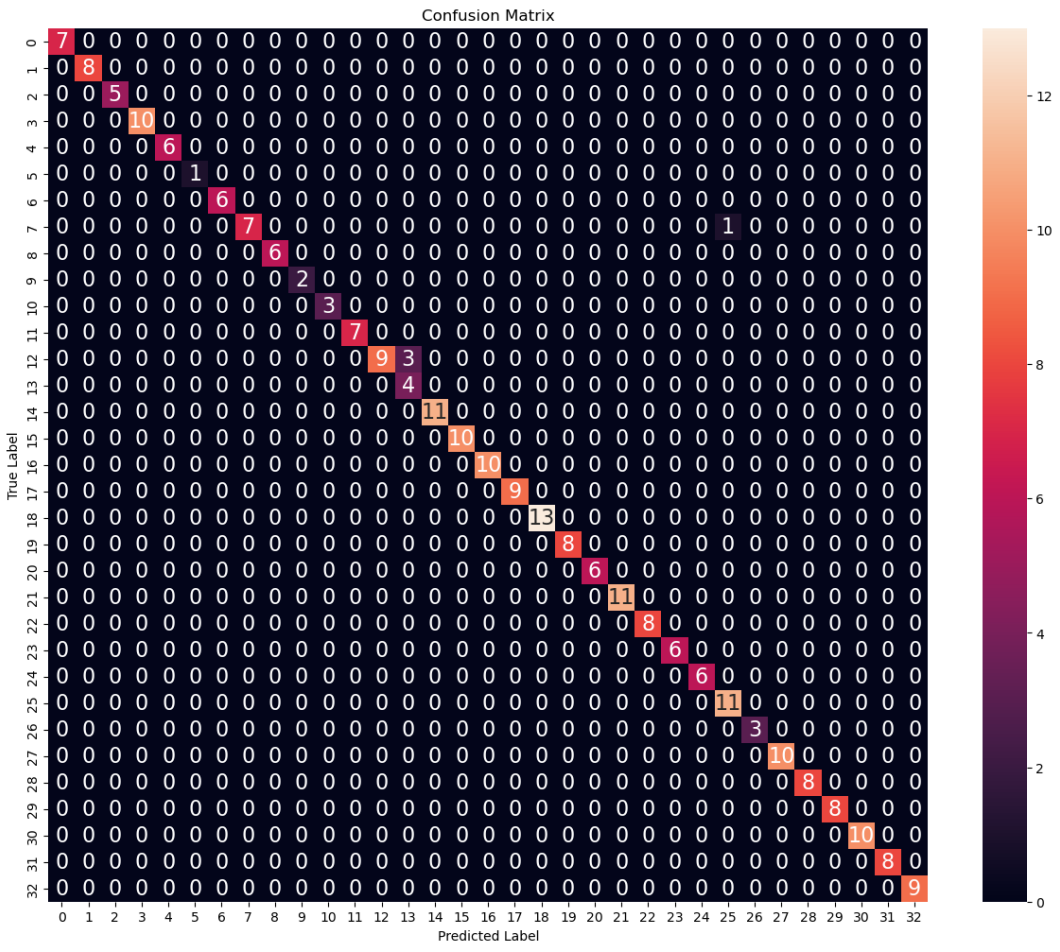
- Тестовий набір даних становить 20% від усього набору даних (250 відеозаписів по 65 кадрів; загалом 16 250 відеокадрів)
- Точність класифікатора становить 98.4% та обраховується за наступною формулою:

$$Accuracy = \frac{Num.of\ correct\ predicitons}{Total\ num.of\ predictions}$$

```
In [49]: print(f'Model accuracy: {round(accuracy_score(ytrue, yhat) * 100, 2)}%')
         Model accuracy: 98.4%
```

Точність класифікатора

ОЦІНКА ЕФЕКТИВНОСТІ МОДЕЛІ



Матриця невідповідностей класифікатора

Використовуючи матрицю невідповідностей можна порахувати кількість істинно позитивних (TP), істинно негативних (TN), хибно позитивних (FP) та хибно негативних (FN) передбачень моделі для кожного класу.

ОЦІНКА ЕФЕКТИВНОСТІ МОДЕЛІ

	precision	recall	f1-score	support
0	1.00	1.00	1.00	7
1	1.00	1.00	1.00	8
2	1.00	1.00	1.00	5
3	1.00	1.00	1.00	10
4	1.00	1.00	1.00	6
5	1.00	1.00	1.00	1
6	1.00	1.00	1.00	6
7	1.00	0.88	0.93	8
8	1.00	1.00	1.00	6
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	3
11	1.00	1.00	1.00	7
12	1.00	0.75	0.86	12
13	0.57	1.00	0.73	4
14	1.00	1.00	1.00	11
15	1.00	1.00	1.00	10
16	1.00	1.00	1.00	10
17	1.00	1.00	1.00	9
18	1.00	1.00	1.00	13
19	1.00	1.00	1.00	8
20	1.00	1.00	1.00	6
21	1.00	1.00	1.00	11
22	1.00	1.00	1.00	8
23	1.00	1.00	1.00	6
24	1.00	1.00	1.00	6
25	0.92	1.00	0.96	11
26	1.00	1.00	1.00	3
27	1.00	1.00	1.00	10
28	1.00	1.00	1.00	8
29	1.00	1.00	1.00	8
30	1.00	1.00	1.00	10
31	1.00	1.00	1.00	8
32	1.00	1.00	1.00	9
accuracy			0.98	250
macro avg	0.98	0.99	0.98	250
weighted avg	0.99	0.98	0.98	250

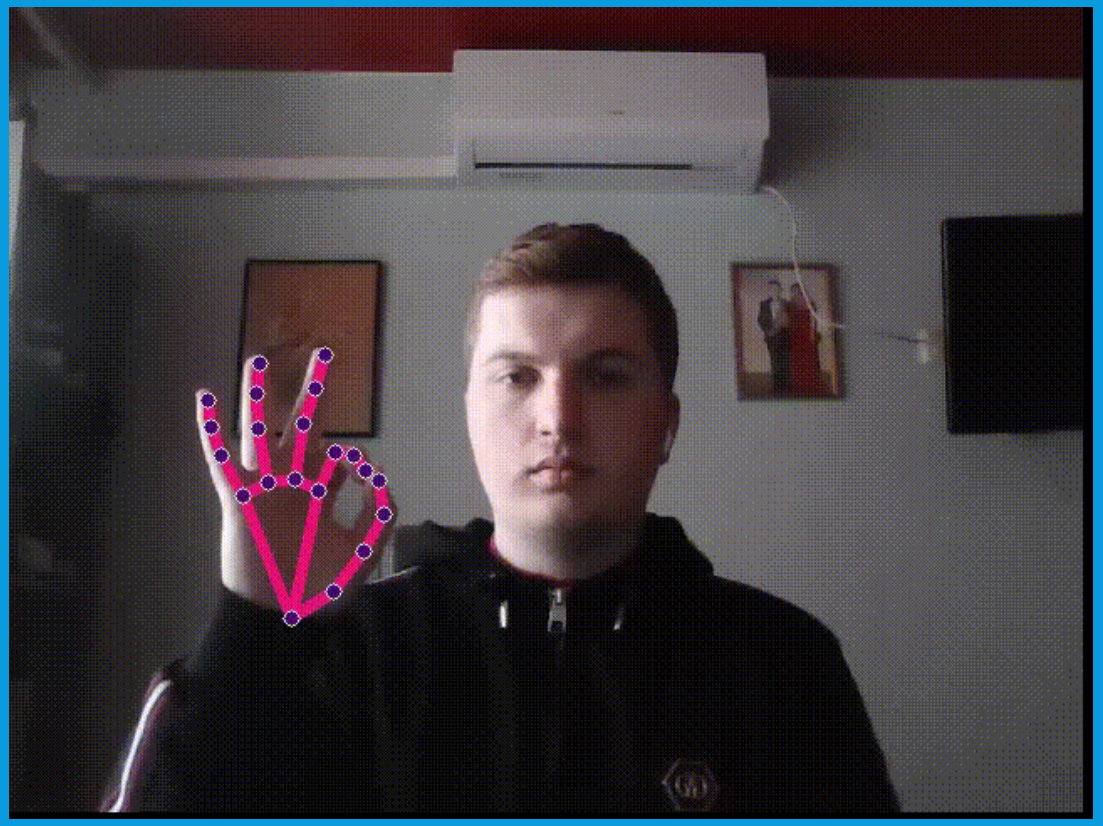
Влучність (precision), повнота (recall) та F1 score обчислені для кожного класу моделі.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precison * Recall}{Precision * Recall}$$

ПРОГРАМА ДЛЯ РОЗПІЗНАВАННЯ ЖЕСТІВ



Запис вікна програми, де можна побачити результати розпізнавання для жестів, що відповідають літерам «О», «Н», «Ш» та «Г» українського алфавіту.

Виявлення та класифікація жестів здійснюється тільки тоді, коли рука знаходиться в полі зору веб-камери.

Оскільки кириличні літери не підтримуються шрифтами, що надаються OpenCV, всі літери з українського алфавіту були транслітеровані.

РЕЗУЛЬТАТИ

- У роботі були розглянуті підходи до розпізнавання жестів та побудовано класифікатор, на базі LSTM, для класифікації усіх 33 жестів української дактильної абетки.
- Для навчання класифікатора було створено власний набір даних, де кожному жесту дактильної абетки відповідає 50 відеозаписів по 65 кадрів. Для збору даних було розроблено окрему програму.
- Використовуючи класифікатор для жестів дактильної абетки та бібліотеку Google MediaPipe, було створено програму для розпізнавання жестів української дактильної абетки у реальному часі.

ВИСНОВКИ

- Отримана модель має точність 98.4%. Зважена середня влучність та повнота рівні 99% та 98% відповідно. Зважений середній F1-score рівний 98%.
- Порівнюючи отриману модель з аналогічними роботами, можна відзначити, що її точність вища, ніж у моделей, представлених в [12] та [10], де вона рівна 92,54% та 97% відповідно.

ДЯКУЮ ЗА УВАГУ