ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ТЕЛЕКОМУНІКАЦІЙ

Пояснювальна записка

до магістерської кваліфікаційної роботи

на тему: **«РОЗРОБКА ІНФОРМАЦІЙНОЇ СИСТЕМИ ДЛЯ РОЗПІЗНАВАННЯ ОБ'ЄКТІВ З ВІДЕОПОТОКУ ЗА ДОПОМОГОЮ ТЕХНОЛОГІЇ КОМП'ЮТЕРНОГО ЗОРУ»**

Виконав: студент 6 курсу, групи ПДМ-61
спеціальності 121 Інженерія програмного забезпечення
(шифр і назва спеціальності)

Треньов М.Г.
(прізвище та ініціали)

Керівник Гребенюк В.В.
(прізвище та ініціали)

Рецензент
(прізвище та ініціали)

Нормоконтроль Трінтіна Н.А.
(прізвище та ініціали)

Київ – 2022

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра Інженерії програмного забезпечення
Ступінь вищої освіти - «Магістр»
Спеціальність - 121 «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
Завідувач кафедри Інженерії програмного забезпечення автоматизованих систем

О.В.Негоденко
«____»_____2022 року

З А В Д А Н Н Я
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Треньов Микита Георгійович
(прізвище, ім'я, по батькові)

1. Тема роботи: Розробка інформаційної системи для розпізнавання об'єктів з відеопотоку за допомогою технології комп'ютерного зору

Керівник роботи Гребенюк Віктор Вікторович,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом закладу вищої освіти від «11» жовтня 2021 року № 170.

2. Строк подання студентом роботи_____
3. Вихідні дані до роботи:
   1. Принципи і методи технології комп'ютерного зору.
   2. Методи розпізнавання об'єктів у відеопотоці
   3. Методи інтеграції результатів розпізнавання об'єктів.
   4. Науково-технічна література.
4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно вирішити):
   1. Дослідження актуальності розпізнавання об'єктів у відеопотоці.

2. Аналіз оптимальних методів розпізнавання об’єктів у відеопотоці.
3. Побудова математичної моделі системи розпізнавання об'єкта у відеопотоку, що дозволяє досліджувати якісні характеристики результату та час, необхідний його отримання;
4. Дослідити вплив параметрів вхідних даних на вибір оптимальної стратегії комбінування результатів розпізнавання одиночних зображень;

5. Перелік графічного матеріалу:
   1. Розпізнавання на мобільних пристроях
   2. Розпізнавання даних у відеопотоці
   3. Існуючі підходи
   4. Модель системи розпізнавання у відеопотоці
   5. Стратегії комбінування покадрових результатів класифікації
   6. Інтеграція одиночних символів
   7. Схема обробки кадру у системі розпізнавання документів у відеопотоці
   8. Приклад розпізнавання документу

6. Дата видачі завдання_____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів магістерської роботи	Строк виконання етапів роботи	Примітка
1	Аналіз актуальності теми дипломної роботи	20.09.2021	
2	Підбір науково-технічної літератури	27.09.2021	
3	Особливості принципів і методів технології комп’ютерного зору. Аналіз вже існуючих розробок. Постановка задачі.	23.10.2021	
4	Дослідження методів розпізнавання об’єктів у відеопотоці	19.11.2021	
5	Розробка додатку для розпізнавання об’єктів	03.12.2021	
6	Вступ, висновки, реферат	07.12.2021	
7	Розробка обов‘язкових демонстраційних матеріалів	12.12.2021	
8	Попередній захист роботи	24.12.2021	

Студент _____ Треньов М.Г.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Гребенюк В.В.
підпис) (прізвище та ініціали)

РЕФЕРАТ

Текстова частина, магістерської роботи включає: 68 с., 29 рис., 2 табл., 24 джерела.

КОМП'ЮТЕРНИЙ ЗІР, ДЕТЕКТОР КУТІВ ХАРРІСА, ЗГОРТКОВІ НЕЙРОННІ МЕРЕЖІ, ЗАДАЧА ЗУПИНКИ, OPENCV, РОЗПІЗНАВАННЯ ОБ'ЄКТІВ.

Об'єкт дослідження – мобільна система для розпізнавання документів, що засвідчують особу.

Мета роботи - оптимізація процесу розпізнання об'єктів у відеопотоці за рахунок комбінування множини результатів спостережень.

Методи дослідження – технологія комп'ютерного зору, нейронні мережі та алгоритми комбінування результатів класифікації.

У роботі проведено аналіз основних понять, структури систем комп'ютерного зору. Досліджено методи розпізнавання об'єктів у відеопотоці мобільного пристрою.

Розглянуто математичну модель системи оптичного розпізнавання об'єкта у відеопотоку з блоком комбінування результатів розпізнавання об'єкта на одиночних зображеннях та з блоком зупинки процесу розпізнавання.

Розробка алгоритму комбінування (інтеграції) результатів розпізнавання рядкового об'єкта відеопотоку в рамках моделі результату, що враховує альтернативні варіанти класифікації окремих символів. Описано постановку задачі, формальний опис алгоритму, а також представлено результати порівняльного експериментального дослідження запропонованого алгоритму та алгоритму ROVER.

ЗМІСТ

ВСТУП

В останні роки набувають великої популярності цифрова обробка та цифровий аналіз зображень. Вони застосовуються в різних прикладних областях. Від автоматизованої медичної діагностики до безпілотного керування автомобілем, від глибоководних апаратів, до керування штучними космічними супутниками. Можливість застосування «комп'ютерного зору» багатогранна, а розв'язання завдання у сфері може бути перенесено на зовсім не пов'язану з нею прикладну область. З обробки образів створено чималу кількість робіт, проте «універсального» рішення для будь-якої задачі на даний момент не створено. Кожна прикладна область накладає свої специфічні умови. Виходячи з даних умов необхідно побудувати алгоритм, який вирішує задачу з максимальною швидкістю та точністю, але як правило, залишаються умови, які подолати набагато складніше за інші. Наприклад, більшість методів нестійкі до зміни інтенсивності або напряму освітлення, тоді як інші погано справляються із зміною масштабу тощо.

У рамках даної роботи як прикладна область розглядається автоматизований аналіз відео потоку з відеокамери.

Рівень економіки країни безпосередньо пов'язаний із її технічним розвитком. Інноваційні технології допомагають у вирішенні різних практичних завдань. У той час, як цілі цих завдань можуть бути абсолютно різними: збільшення прибутку від економічних процесів, прискорення доставки товарів, збільшення потужності виробництв, тощо. Однією з таких цілей є збільшення безпеки процесу.

Поширеними методами та підходами до вирішення завдань детектування, розпізнавання та класифікації є:

- *Порівняння із зразком* – класифікація за найближчим середнім, на відстані до найближчого сусіда. Також у групу порівняння із зразком можна віднести структурні методи розпізнавання.

- *Зіставлення з шаблоном* – метод розпізнавання, в якому використовується невелике зображення або шаблон для пошуку співпадаючих областей у збільшеному зображенні.

- *Нейронні мережі* – клас методів глибокого навчання, який використовується для автоматичного вивчення властивостей об'єкта та його подальшої ідентифікації. Відмінною особливістю цих методів від інших є здатність вчитися.

Мета даного проекту – створення комп'ютерного сервісу пошуку та класифікації об'єктів з відеопотоку.

Для досягнення поставленої мети в роботи були поставлені та вирішені такі завдання:

1. Розглянути існуючі алгоритми для детектування та класифікації об'єктів
2. Обрати мови та середовища програмування для реалізації обраних алгоритмів.
3. Програмно реалізувати розпізнавання об'єктів.
4. Навчити класифікатор з використанням розроблених моделей та провести його тестування.

1 ОГЛЯД МЕТОДІВ РОСПІЗНАВАННЯ ОБ’ЄКТІВ, КОМП’ЮТЕРНОГО ЗОРУ ТА МАШИННОГО НАВЧАННЯ

1.1 Введення у комп'ютерний зір

Комп'ютерний зір (computer vision, машинний зір, технічний зір) - це технологія, яка в залежності від поставленого завдання може знаходити, відстежувати, класифікувати та ідентифікувати об'єкти, витягуючи та аналізуючи отриману інформацію з зображень або відео. Цей напрямок виник у рамках штучного інтелекту. Основним завданням є розпізнавання образів, тому для повної і правильної інтерпретації того, що зображено, потрібно мати необхідну інформацію як масив пікселів, витягнутих із зображення. Кожне зображення складається з набору пікселів. Піксель є найбільшою деталізацією зображення для комп’ютера.

Під комп'ютерним зором також розуміється автоматичне отримання інформації з зображень. У ролі інформації може виступати 3D-моделі, положення камери, виявлення та розпізнавання об'єктів, групування зображень та пошук зображень за змістом. Застосування комп'ютерного зору набуває досить великої значимості у різних сферах нашого життя. За допомогою цієї технології реалізовано розумну систему відеоспостереження, яка обробляє вхідні відео та, відповідно до навченого алгоритму, приймає необхідні рішення.

До вирішення завдань детектування та класифікації застосовуються різні підходи: статистичні, спеціально розроблені теорії ключових точок, застосування алгоритмами класифікації зображень за змістом, а також машинне навчання.

Найбільш поширені завдання комп'ютерного зору[1]:

- *Розпізнавання* – одне з базових та першорядних завдань у обробці зображень, комп'ютерному та машинному зору. Воно допомагає класифікувати і ідентифікувати об'єкти, що характеризуються певним набором властивостей та ознак.

- *Відновлення зображень* – це видалення шуму з використанням різних методів, наприклад, розмиття за допомогою фільтрів на основі машинного навчання (шум датчика, розмитість рухомого об'єкта і т. д.).

- *Аналіз руху* - завдання використовує комп'ютерний зір для оцінки швидкості руху об'єктів у відео. Також застосовується для оцінки рухів, у яких послідовність відеоданих обробляється для знаходження швидкості кожної точки зображення чи 3D сцени.

- *Відновлення чи реконструкція сцени* – допомагає відтворити тривимірну модель зображення або сцени, що вводиться за допомогою зображень чи відео. Найчастіше моделлю служить набір точок тривимірного простору.

- *Обробка та аналіз зображення* – завдання зосереджено на перетворення одного 2D-зображення в інше. Реалізується за допомогою піксельних операцій, таких як підвищення контрастності або поворот зображення.

- *Високорівнева обробка* – невеликий набір даних. Вона використовує різні методи для вилучення інформації з сигналів в цілому, наприклад, набір точок або ділянка зображення, в якому імовірно знаходиться певний об'єкт, що цікавить частина даних.

1.1.1 Принципи роботи комп'ютерного зору

Комп'ютери інтерпретують зображення як послідовність пікселів, кожен з яких має власний набір значень кольору. Пікселі є необробленими будівельними блоками зображення. Кожне зображення складається з набору пікселів. Піксель вважається «кольором» або «яскравістю» світла, що з'являється на зображенні. Якщо ми розглядаємо зображення як сітку, то кожен квадрат містить один піксель.

Рисунок 1.1 – Приклад зображення шириною 1000 пікселів та висотою 750 пікселів

Зображення на малюнку 1 має роздільну здатність 1000 × 750 пікселів, де 1000 – це ширина, а 750 – висота. Ми можемо уявити зображення у вигляді матриці. У цьому випадку наша матриця має 1000 стовпців (ширина) та 750 рядків (висота) та містить 1000 × 750 = 750 000 пікселів, які представлені двома способами: а) відтінки сірого (один канал); б) колір.

У зображеннях із градаціями сірого кольору кожен піксель є скалярним значенням від 0 до 255, де нуль відповідає "чорному" кольору, а 255 - "білому". Значення між 0 та 255 мають різні відтінки сірого, де значення ближче до 0 темніше, а значення ближче до 255 світліше. Градієнтне зображення у градаціях сірого (рисунок 2) демонструє темніші пікселі з лівого боку, а світліші з правого. З малюнка 2 можна зрозуміти те, як значення в градаціях сірого перетворюються на двовимірний масив цілих чисел:

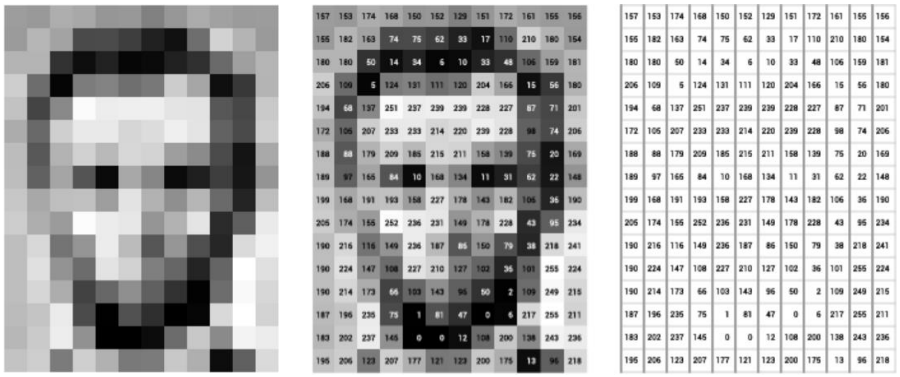


Рисунок 1.2 – Зображення та їх матричні уявлення

Ряди чисел праворуч (крайнє праве зображення на рисунку 2) – комп'ютерне представлення введеного зображення. У прикладі зображення має 12 стовпців та 16рядків, що означає 192 вхідних значень цього зображення.

Ще необхідно описати як комп'ютер перетворює для себе кольорові зображення у вигляді матриць. Пікселі у кольоровому просторі RGB більше не є скалярними значеннями, як було у зображеннях у градаціях сірого, на одному каналі. Натомість пікселі представлені списком з трьох значень: одне значення для компонента червоного (Red), друге для зеленого (Green) та третє для синього (Blue). Щоб визначити колір у колірної моделі RGB, все, що нам потрібно зробити, це підрахувати кількість червоного, зеленого та синього кольору, що містяться в одному пікселі. Кожен канал Red, Green та Blue може мати певні значення в діапазоні [0, 255], всього 256 відтінків, де 0 означає відсутність кольору, а 255 - це повна наявність кольору. Враховуючи що значення пікселя має бути лише в діапазоні [0, 255], ми зазвичай використовуємо 8-бітові цілі числа без знака для представлення яскравості. Розглянемо приклад створення різних кольорів зображення на малюнку 2а.

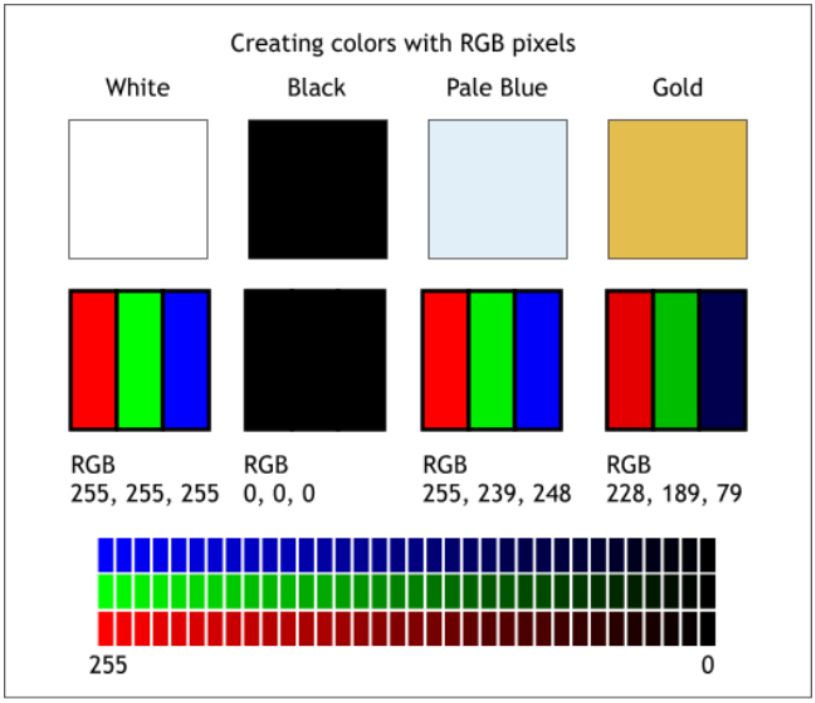


Рисунок 1.2а – Створення квітів у пікселі RGB

З цього малюнка 2а можна зрозуміти, що кожен піксель складається з трьох різних кольорів, при зміні їх відтінків можна отримати абсолютно різні кольори.

Для вирішення завдань детектування, розпізнавання та класифікації об'єктів на графічному зображенні необхідно попередньо обробляти зображення (наприклад, виконати операцію середнє віднімання чи масштабування). Оскільки типи даних, які використовуються бібліотеками (наприклад, OpenCV), завантажують зображення з диска, то їх необхідно перетворювати, перш ніж безпосередньо застосовувати алгоритми навчання до зображень. Враховуючи наші три значення Red, Green та Blue, ми можемо поєднати їх у кортеж RGB (червоний, зелений, синій). Цей кортеж представляє цей колір у колірному просторі RGB.

Ми можемо перетворити зображення RGB як три незалежних матриць, шириною W і висотою H, по одній для кожного з компонентів RGB, як показано на малюнку 3. Ми можемо об'єднати ці три матриці для отримання багатовимірного масиву з формою W × H × D, де D - глибина чи кількість каналів. Для кольорового простору RGB глибина D = 3.

Рисунок 1.3 – Вихідне зображення та його RGB канали

1.2 Огляд методів комп'ютерного зору завдання розпізнавання об'єктів

Для комп'ютерів завдання інтерпретації вмісту зображення менш тривіальна, ніж просто завдання відображення зображення. Все, що бачить наш комп'ютер - це велика матриця чисел. Щоб зрозуміти зміст зображення, ми повинні застосувати класифікацію зображень, що є завданням використання алгоритмів комп'ютерного зору та машинного навчання. У цій роботі були застосовані різні методи, крім глибокого навчання, що покращують результати комп'ютерного зору. Тим не менш, вони добре працюють для більш простих завдань, але оскільки дані стають величезними, але оскільки завдання стають більш складними, вони не замінюють згорткові нейронні мережі.

1.2.1 Перетворення рівня яскравості

Зчитавши за допомогою NumPy зображення в масив, ми можемо застосувати щодо нього різні математичні операції. Простий приклад

перетворення рівня яскравості напівтонового зображення. Візьмемо довільну функцію f , що відображає інтервал 0 ... 255 (або, якщо завгодно,0…1) у собе, тобто область значень збігається з областю визначення.

Іншим прикладом перетворення яскравості є вирівнювання гістограми. Ця операція змінює гістограму яскравості, так щоб результуюча гістограма містила всі можливі значення яскравості та при цьому приблизно в однаковій кількості. Вона часто застосовується для нормування яскравості перед подальшою обробкою, а також для підвищення контрастності. В даному випадку для перетворення використовується функція розподілу (cumulative distribution function, CDF) значень пікселів у зображенні. Приклад цієї операції наведено на малюнку 19.

1.2.2 Детектор кутів Харріса

Алгоритм виявлення кутів Харріса (детектором кутів Харріса-Стівенса) – один із найпростіших детекторів кутів об'єктів. Ідея полягає в тому, щоб знайти особливі точки, в околиці яких є межі у кількох напрямах, і є кутові точки.

Визначимо позитивно-напіввизначену симетричну матрицю

$$M_I = M_I(x), \tag{1.1}$$

де x – точка всередині зображення:

$$M_I = \nabla I \nabla I^T = \begin{bmatrix} I_x \\ I_y \end{bmatrix} \begin{bmatrix} I_x & I_y \end{bmatrix} = \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}. \tag{1.2}$$

Тут ∇ I – вектор градієнта зображення, що містить похідні I x та I_y (Визначення похідних були дані вище). За побудовою, M_I має ранг 1, а її власні значення дорівнюють

$$\lambda 1 = |\nabla I|\,2\,,\ \lambda 2 = 0. \tag{1.3}$$

Таким чином, у нас є по одній матриці для кожної точки зображення. Залежно від значень ∇I в області є три випадки:

- Якщо λ_1 і λ_2 – великі позитивні числа, то у точці x є кут.

- Якщо λ_1 велике, а $\lambda_2 \approx 0$, то існує межа і за усереднення M_I по області власні значення змінюються не сильно.

- Якщо $\lambda_1 \approx \lambda_2 \approx 0$, то в точці x немає жодних особливостей.

Розглянемо спосіб з прикладу зображення будівлі державного університету телекомунікацій:

Рисунок 1.4 – а) вихідне зображення; б) зображення з виявленими кутами за алгоритмом Харріса

1.2.3 Фільтрування контурів

Контури дуже корисні, коли ми хочемо перейти від роботи з зображенням для роботи з об'єктами на цьому зображенні. Коли об'єкт досить складний, але добре

виділяється, то найчастіше єдиним методом роботи з ним є виділення його контурів.

Оператор Кенні є популярним алгоритмом виявлення кордонів і найчастіше використовується для виділення контуру об'єктів[2]. Алгоритм Кенні для виявлення меж об'єктів складається з п'яти етапів:

1. Придушення шуму;

2. Розрахунок градієнта;

3. Пригнічення країв зображення;

4. Подвійний поріг;

5. Відстеження країв по гістерезі.

Ще одна важлива річ, яку варто згадати, що алгоритм заснований на зображеннях у градаціях сірого. Отже, попередньою умовою є перетворення зображення у відтінки сірого кольору перед виконанням вищезгаданих кроків. Коротко дамо визначення етапів алгоритму Кенні.

Коротко дамо визначення етапів алгоритму Кенні.

Придушення шуму – є один із способів позбутися шумів на зображенні. Для того щоб згладити шум застосовують розмиття по Гауссу. Для цього застосовується метод згортки зображень з гаусовим ядром (наприклад, 3х3, 5х5, 7х7 тощо пікселів). Розмір ядра залежить від очікуваного ефекту розмиття. Здебільшого найменше ядро – це менш помітна пляма. При обробці зображення, ядро(згортка матриці) є невеликою матрицею. Воно використовується для розмиття, підвищення різкості, тиснення, виявлення країв та іншого.

Розрахунок градієнта – визначає інтенсивність та напрямок краю шляхом обчислення градієнта зображення з використанням операторів виявлення краю. Градієнт – це векторна величина, що показує напрямок якнайшвидшого зростання двовимірної функції яскравості зображення.

Краї відповідають зміні інтенсивності пікселів. Щоб виявити її, найпростіше застосувати фільтри, які виділяють ці зміни інтенсивності в обох напрямках: горизонтальному (x) та вертикальному (y).

Придушення країв зображення – в ідеалі кінцеве зображення має мати тонкі краї. Таким чином, ми повинні виконати не-максимальне придушення, щоб виявити краї. Для цього можна застосувати наступний алгоритм. Обійти через усі крапки на матриці інтенсивності градієнта і знайти пікселі з максимальним значенням у напрямку ребер (рисунок 5).

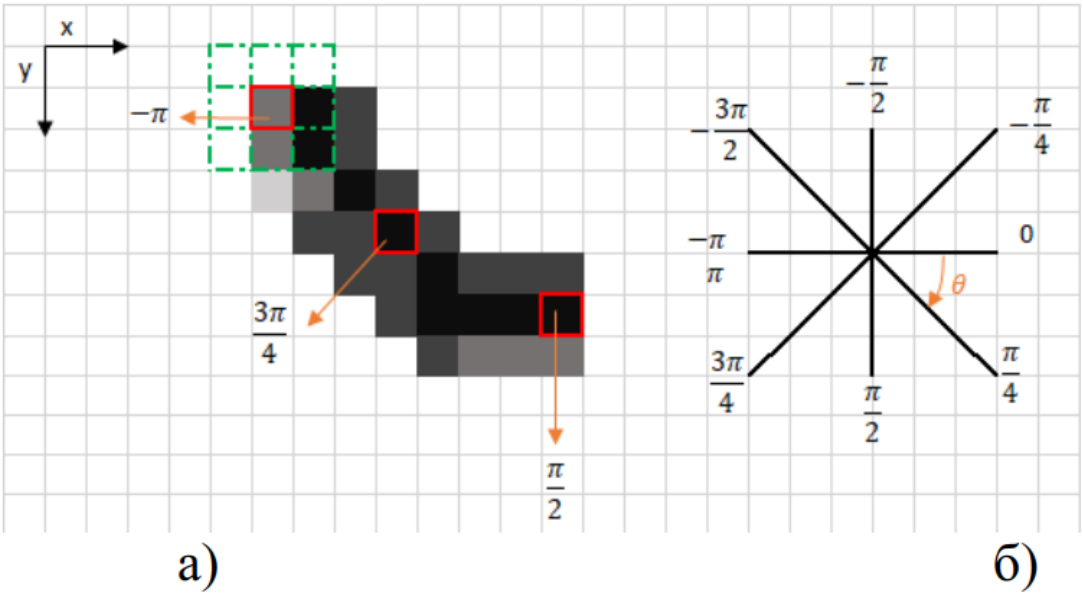


Рисунок 1.5 – а) частина краю об'єкта у вигляді пікселів

Червоний прямокутник у верхньому лівому кутку (рисунок 5а), представляє піксель інтенсивності оброблюваної градієнтної матриці інтенсивності. Відповідний напрямок країв позначено оранжевою стрілкою з кутом -π радіан.

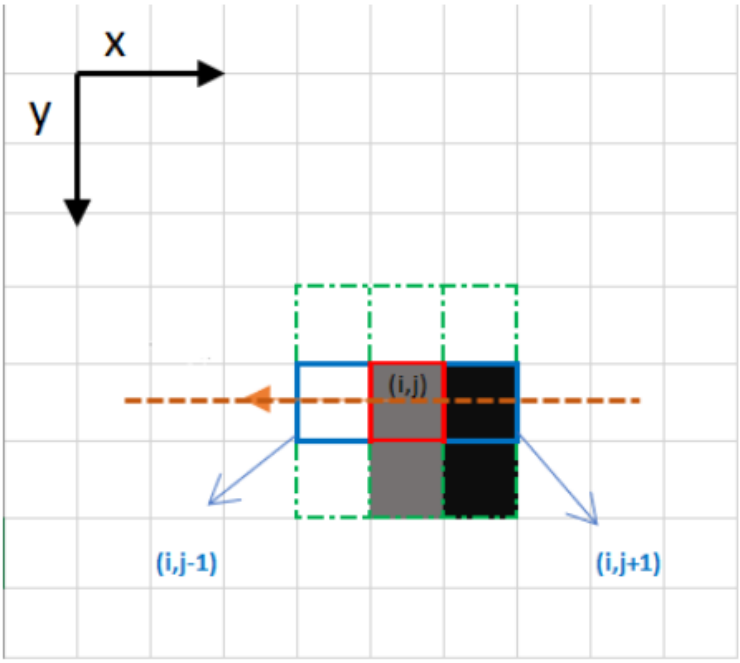


Рисунок 1.6 – зображення з неповним придушенням країв

Напрямок краю – помаранчева пунктирна лінія (горизонтальна лінія зліва направо, (рисунок 6). Мета алгоритму – перевірити, чи є пікселі в тому самому напрямку більш менш інтенсивними, ніж оброблювані[3].

У наведеному вище прикладі піксель (i, j) обробляється, і пікселі в тому ж напрямку виділяються синім кольором (i, j - 1) і (i, j + 1). Якщо один із цих двох пікселів є більш інтенсивним, ніж оброблюваний, то зберігається лише інтенсивніший. Піксель (i, j - 1) здається більш інтенсивний, тому що він білий (значення 255). Отже, значення інтенсивності поточного пікселя (i , j) встановлюється 0. Якщо в напрямі краю відсутні пікселі, що мають більш інтенсивні значення, то значення поточного пікселя зберігається.

Подвійний поріг – мета застосування цієї операції спрямована на виявлення трьох видів пікселів та за допомогою них виявляти контури:

- Сильні пікселі – пікселі з високою інтенсивністю (яскравістю).

- Слабкі пікселі – це пікселі, які мають достатнє значення інтенсивності. Ці пікселі не можна вважати сильними, але їх значення інтенсивності не є маленькими, щоб їх вважати, як пікселі, що не мають відношення до країв.

- Інші пікселі вважаються такими, що не мають відношення до країв.

Відстеження країв – результати порогового значення гістерезис, складається з перетворення слабких пікселів у сильні, якщо хоча б один із пікселів навколо краю, що обробляється, є сильним. Подивимося отримані нами результати після застосування алгоритму виявлення країв різних об'єктів (малюнки 7-8).

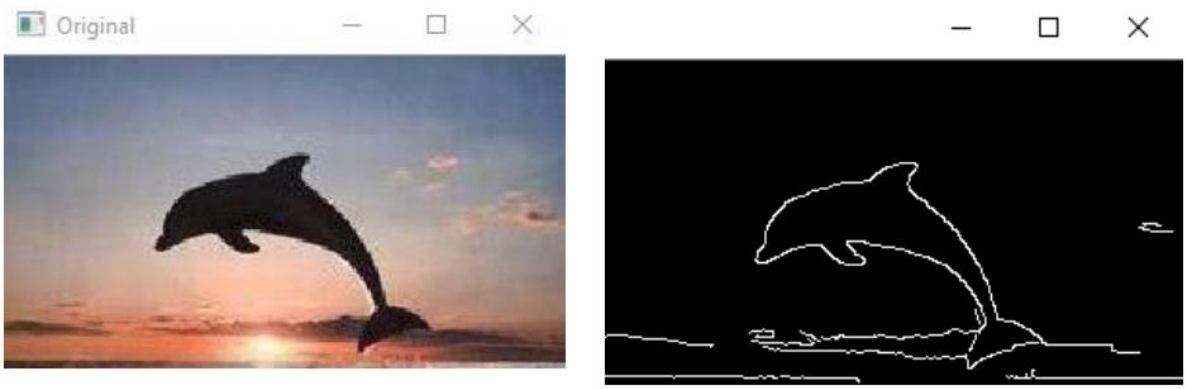


Рисунок 1.7 – а) вихідне кольорове зображення ,б) результат виявлення країв об'єкта

В результаті видно, що цей метод може знайти всі краї об'єкта(Дельфін), видаляючи при цьому багато країв об'єктів на задньому фоні (хмари). Далі розглянемо складніший об'єкт і застосуємо той самий алгоритм(Рисунок 8).

Рисунок 1.8 – а) вихідне кольорове зображення, б) результат виявлення країв об'єкта

З прикладів видно, що алгоритм виявлення країв Кенні дає задовільний результат, як простих, так складніших об'єктів.

1.3 Згорткові нейронні мережі

Штучні нейронні мережі (нейромережі або просто мережі) – це клас моделей машинного навчання, в основі яких лежить система центральної нервової системи ссавців.

Нейронна мережа складається з кількох взаємопов'язаних різних шарів, таких як вхідний шар, щонайменше один прихований шар і вихідний шар (рисунок 9). Їх найкраще використовувати при виявленні об'єктів для розпізнавання образів, країв (вертикальні/горизонтальні), форми, кольори та

текстури. Приховані шари є згортковими шарами. У даному типі нейронної мережі, згорткові шари діють як фільтр, який спочатку отримує вхідні дані, перетворює їх, використовуючи певний алгоритм або функцію, та відправляє його на наступний шар. Основні параметри нейронів є вхідний (синій колір) та вихідний шар (зелений колір).

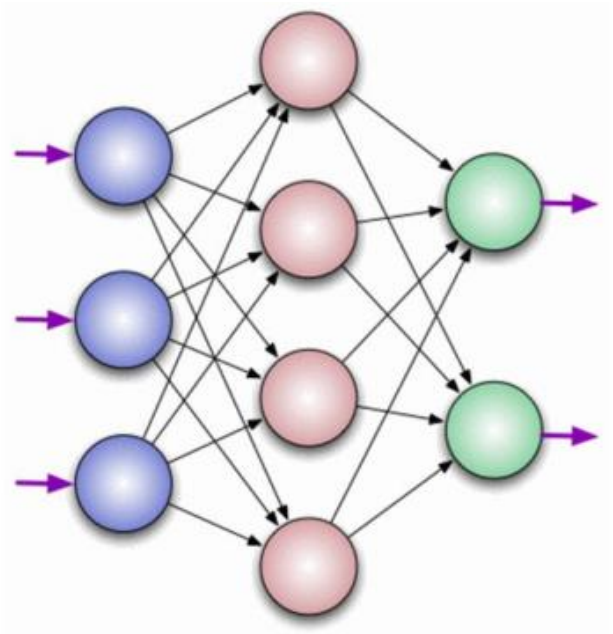


Рисунок 1.9 – Модель нейронної мережі

З більшою кількістю згорткових шарів, кожен раз, коли новий вхід відправляється на наступний згортковий шар, він змінюється по-різному. Наприклад, у згортковому шарі фільтр може ідентифікувати форму/колір у певній області, останній згортковий шар, може класифікувати об'єкт.

У загальному випадку згорткова нейронна мережа складається з великої кількості шарів. На останніх етапах зазвичай використовується один чи кілька повнозв'язкових шарів.

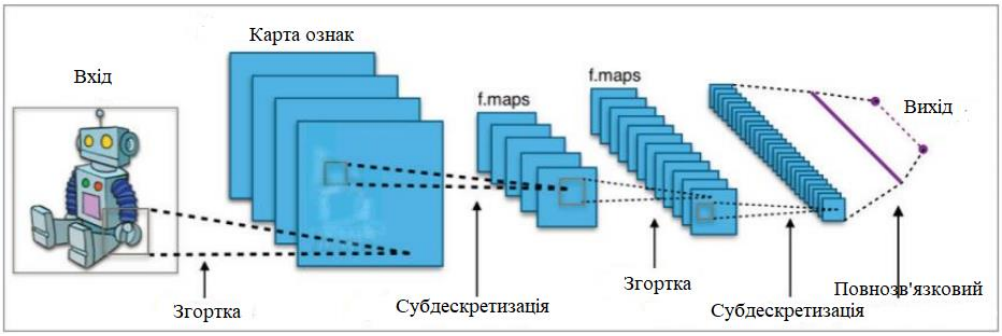


Рисунок 1.10 – Топологія згорткової нейронної мережі

Згорткові нейронні мережі забезпечують часткову стійкість до змін масштабу, зсувів, поворотів, зміни ракурсу та інших спотворень. Згорткові нейронні мережі поєднують три архітектурні ідеї, для забезпечення інваріантності до зміни масштабу, повороту, зсуву та просторовим спотворенням:

- локальні рецепторні поля (забезпечують локальну двовимірну зв'язність нейронів);

- загальні синоптичні коефіцієнти (забезпечують детектування деяких рис у будь-якому місці зображення і зменшують загальне число вагових коефіцієнтів);

- ієрархічна організація із просторовими підвибірками.

1.3.1 Розпізнавання об'єктів за допомогою нейронної мережі

Для того, щоб навчити нейронну мережу виявляти об'єкти на будь-якому зображенні, з приблизно однаковою формою і кольорами, слідує застосовувати різні фільтри (рисунок 11).

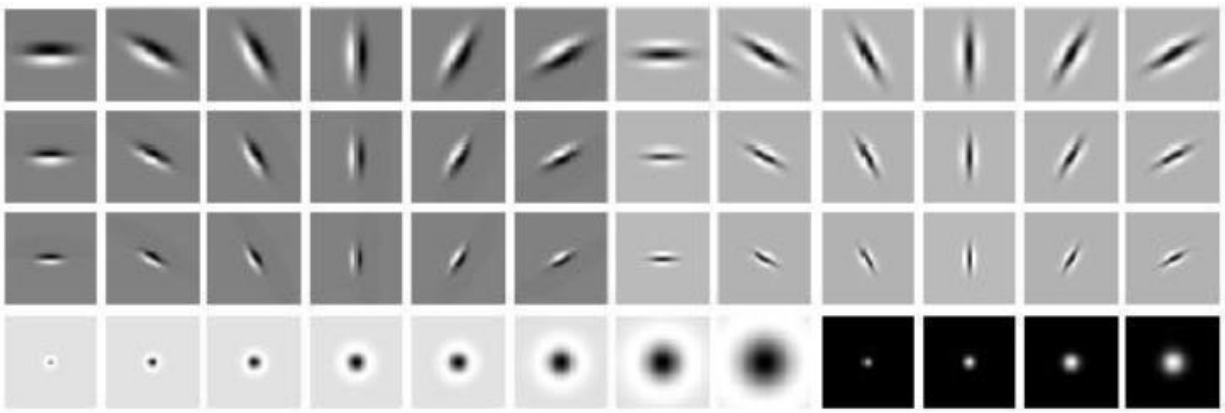


Рисунок 1.11 – Фільтри для різних фрагментів зображення

За допомогою різних фільтрів можна виділяти різні фрагменти зображення, які потім виходить виявити та дослідити у вигляді окремих властивостей та передавати іншими шарами нейрона (рисунок 11).

Щоб мережам не доводилося окремо розпізнавати об'єкти у різних частинах зображення, ми «поділяємо» ваги, що відповідають за розпізнавання між різними фрагментами вихідного зображення. Розглянемо зображення, на якому потрібно не просто виділити об'єкт, а встановити кількісну точність рішення для чотирьох фрагментів (рисунок 12).

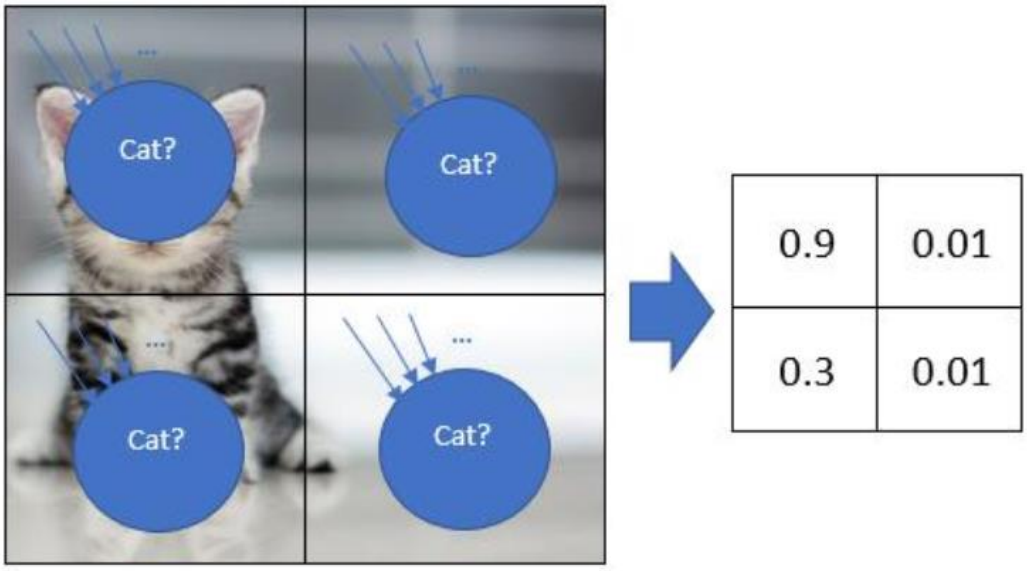


Рисунок 1.12 – Розпізнавання об'єкта та характеристики точності розв'язання

У таблиці вказані кількісні характеристики рішення, які показують ступінь точності розпізнавання об'єкта у цьому фрагменті зображення. Знаходимо максимальне значення **max**{0.9, 0.3, 0.01, 0.01} = 0.9, яке і буде необхідною характеристикою (рисунок 13).

Рисунок 1.13 – Отримання результату для виявленого об'єкта

Основна ідея алгоритму полягає в наступному:

- Використовуємо поділ ваг (weight sharing) для створення "фільтруючого вікна", що пробігає по зображенню.
- Застосований до зображення фільтр допомагає виділити фрагменти, важливі для розпізнавання.
- У той час як у традиційному машинному зорі фільтри конструювали вручну, нейромережі дозволяють нам сконструювати оптимальні фільтри з допомогою навчання.
- Фільтрування зображення можна поєднати з обчисленнями нейронної мережі.

1.3.2 Згортковий шар нейронної мережі

Згортковий шар є набір карт (карти ознак, features maps). Кожна карта має скануюче ядро (скануюче ядро являє собою фільтр, який ковзає по всьому зображенню і знаходить задані ознаки у будь-якому його місці). Кількість карт визначається вимогами до завдання, якщо взяти велику кількість карт, то підвищиться якість розпізнавання, але збільшиться обчислювальна складність.

Розмір у всіх карт згорткового шару однаковий і обчислюється за формулою: $(w, h) = (mW - kW + 1, mH - kH + 1)$ (1.3)

де (w , h) – розмір обгорткової карти, що обчислюється, mW – ширина попередньої карти, mH – висота попередньої карти, kW – ширина ядра, kH –висота ядра.

Ядро ковзає по попередній карті і здійснює операцію згортка, яка часто використовується для обробки зображень:

$$(f*g)[m, n] = \sum f[m - k, n - l]\, k,l \cdot g[k, l] \quad (1.4)$$

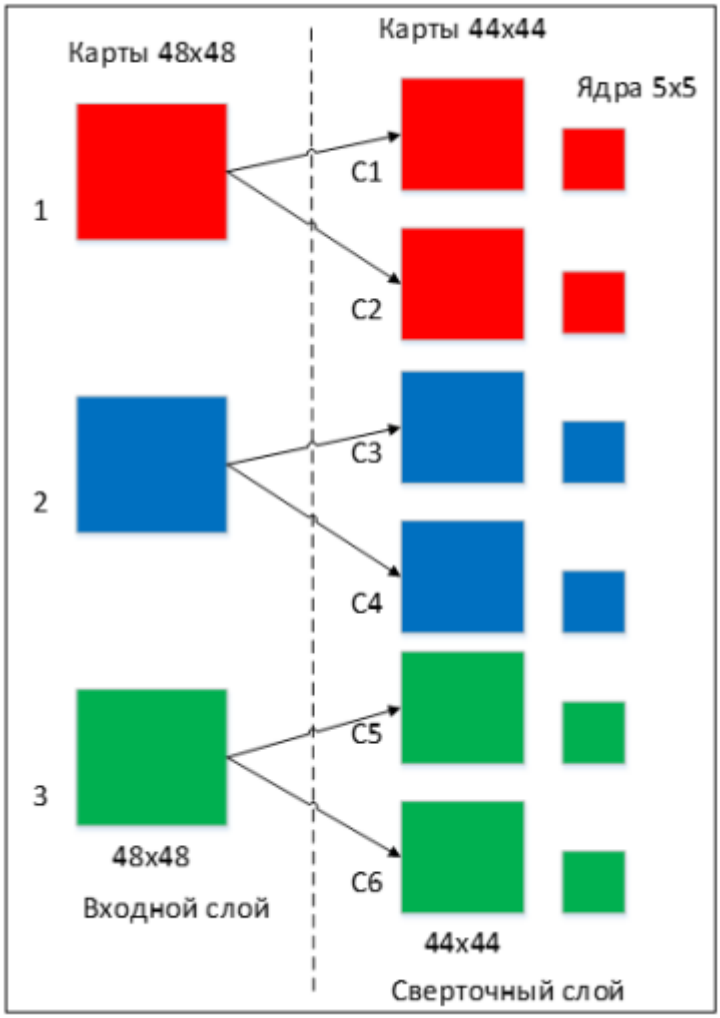


Рисунок 1.14 – Організація зв'язків між картами згорткового та вхідного(попереднього) шару

де f - вихідна матриця зображення,g - ядро згортки.

Відбувається наступне: вікном розміру ядра g проходимо із заданим кроком (зазвичай 1) все зображення f , на кожному кроці поелементно множимо вміст вікна на ядро g результат підсумовується і записується в матрицю результату, як у малюнку 15.

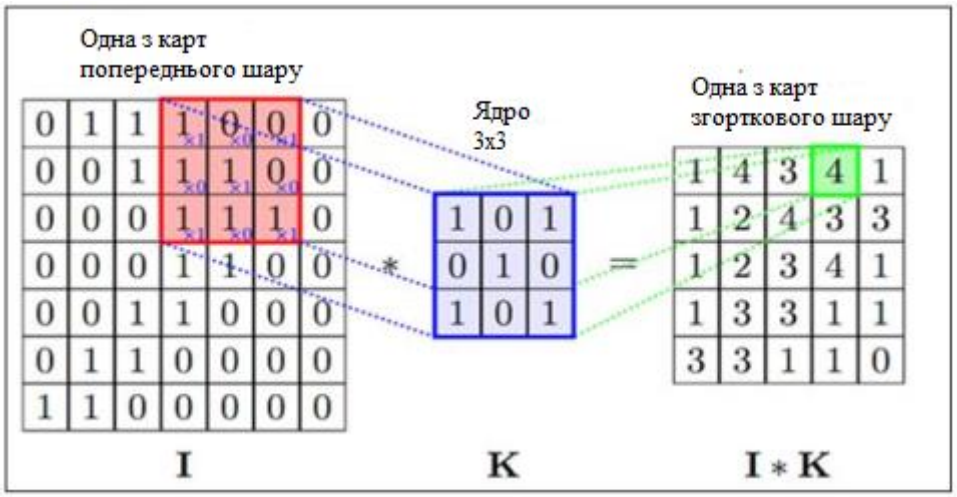


Рисунок 1.15 – Операція згортки та отримання значень карти ознак

1.3.3 Підвибірковий шар

Підвибірковий шар також, як і згортковий має карти, але їх кількість збігається з попереднім (згортковим) шаром. Ціль використання даного шару – зменшити розмірності карток попереднього шару. Якщо на попередній операції згортки вже було виявлено деякі ознаки, то для подальшої обробки настільки докладне зображення вже не потрібне, і воно ущільняється до менш детального. До того ж фільтрація вже непотрібних деталей допомагає не перенавчити мережу.

У процесі сканування ядром підвибіркового шару (фільтром) карти попереднього шару, скануюче ядро не перетинається на відміну від згорткового шару. Зазвичай кожна карта має ядро розміром 2x2, що дозволяє зменшити попередні карти згорткового шару в 2 рази. Вся карта ознак поділяється на комірки 2x2 елементи, з яких вибираються максимальні за значенням. Зазвичай у підвиборному шарі застосовується функція активації (ReLU, Rectifiedlinearunit). Операція підвиборки виконується (Max-Pool – вибір максимального) відповідно до рисунка 16.

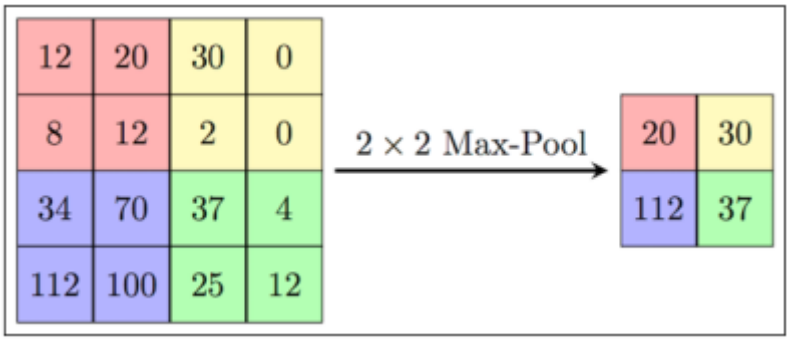


Рисунок 1.16. Формування нової карти підвиборного шару на основіпопередньої картки згорткового шару. Операція підвиборки (MaxPooling)

Шар може бути описаний формулою:

$$x^l = f(a^l \cdot subsample(x^{l-1}) + b^l), \quad (1.5)$$

Де x^l - Вихід шару, l, f - функція активації, a^l, b^l - Коефіцієнти зсуву шару l, *subsample* - операція вибірки локальних максимальних значень.

1.3.4 Повнозв'язковий шар

Останній з типів шарів – це шар звичайного багатошарового персептрона. Ціль застосування даного шару – це звернення до виходу попереднього шару та визначення властивостей, які більше пов'язані з певним класом.

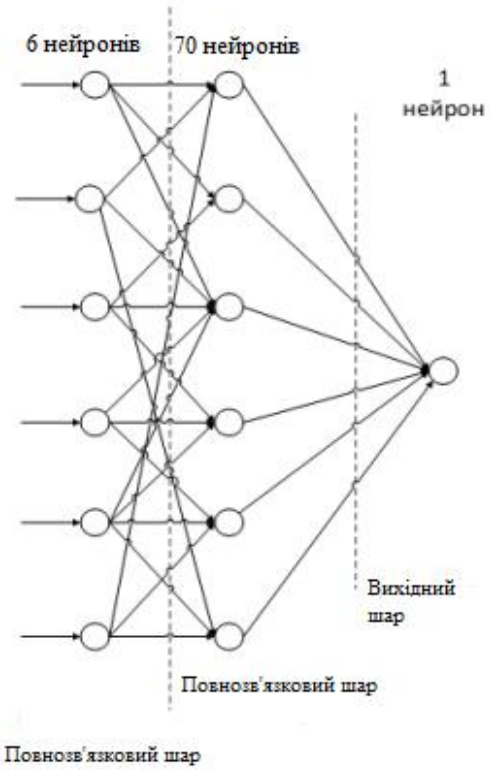


Рисунок 1.17 – Приклад повнозв’язкових шарів

Нейрони кожної карти попереднього підвиборчого шару пов’язані з одним нейроном прихованого шару. Таким чином число нейронів прихованого шару дорівнює числу карт підвиборного шару, але зв’язку можуть бути необов’язково такими, наприклад, лише частина нейронів якоїсь із карт підвибіркового шару пов’язана з першим нейроном прихованого шару, а частина, що залишилася з другим, або всі нейрони першої карти пов’язані з нейронами 1 та 2 прихованого шару.

2. МОДЕЛЬ СИСТЕМИ РОЗПІЗНАВАННЯ ОБ'ЄКТІВ У ВІДЕОПОТОЦІ МОБІЛЬНОГО ПРИСТРОЮ

2.1 Вступ

Впровадження технологічних, соціальних і комерційних процесів, заснованих на використанні мобільних пристроїв та технологій, в умовах сучасного світу вже є буденністю. Системи технічного зору з використанням мобільних технологій, наприклад, системи автоматичного введення і аналізу документів на мобільних пристроях продовжують витісняти традиційні стаціонарні системи, і розвиток технологій і технічного зору із застосуванням мобільних пристроїв в умовах в апаратних обмежень, пов'язаних з ними, стає все більш актуальним завданням.

Класичні системи розпізнавання і автоматичного введення передбачають використання відсканованого зображення або фотографії об'єкта у якості його оцифрованого представлення. При використанні мобільних пристроїв для оцифрування зображень розпізнаних об'єктів можна використовувати відеопотік цифрової камери окрім практичних фотографій або кадрів. Процес фотографування об'єкта за допомогою сучасних мобільних пристроїв передбачає етап «наведення» об'єктива камери на об'єкт з відображенням кадрів відеопотоку на екрані пристрою в режимі реального часу для управління оператором. Якщо зображення обробляється з одного зображення, інформація, що міститься в захоплених кадрах попереднього перегляду, використовується тільки оператором. При розгляді цілого відеопотоку як цифрового зображення об'єкта стає можливим використовувати набагато більше візуальної інформації[4]. Схема розглянутих систем автоматичного введення документів у відеопотік представлена на рисунку 2.1.

Використання відеопотоку дозволяє вирішувати проблеми, недоступні для вирішення при аналізі однієї фотографії. Зовнішні умови зйомки можуть

призвести до того, що розпізнаний об’єкт сильно спотворений в одному зображенні. Прикладом є відблиски від довгого джерела світла, що з’являються на глянцевій поверхні плоского об’єкта (див. Рис. 2.2).

Оскільки геометричне положення об’єкта, що знімається, має тенденцію змінюватися між кадрами у відеопотоці, відблиски також «зміщені», що дозволяє отримати інформацію про прихований об’єкт на іншому кадрі відеопотоку. Існує також важливий клас об’єктів, які не можуть бути виявлені і розпізнані в окремих зображеннях, наприклад, захисні голографічні елементи , які на окремих зображеннях можуть не відрізняються від відблисків або малюнків .

Рисунок 2.2 — Процес зйомки документа, що посвідчує особу за допомогою мобільного пристрою (в якості документа використовується макет німецької ID-картки).

В таких умовах виникає проблема вибору оптимальної стратегії об’єднання результатів покадрового розпізнавання. Ця проблема практично не описана у літературі, а найближчий спектр методів стосується проблеми об’єднання результатів розпізнавання одного і того ж об’єкта, але різних

класифікаторів. Крім основних стратегій об'єднання оцінок в роботах, що впливають на неоднорідні методи об'єднання результатів класифікаторів, розглядаються стратегії зважування рівнів значень класифікаторів, методи навчання правил поєднання, враховуючі статистичні особливості комбінованих класифікаторів і методи, які не прив'язані до статистичних особливостей класифікаторів, але використовують апарати мультимножин для побудови моделі групової класифікації об'єктів[5].

Основна відмінність відеопотоку як цифрового зображення розпізнаного об'єкта полягає в тому, що для одного і того ж об'єкта існує послідовність спостережень, які відрізняються один від одного. Припущення, що система завжди діє детерміновано, тобто у будь-який момент часу і при будь-яких зовнішніх умовах результати розпізнавання одного і того ж набору вхідних даних завжди збігаються. Таким чином, будь-яка помилка є наслідком нездатності системи одного разу розпізнати об'єкт певного класу. Помилки розпізнавання можна розділити на три групи:

1. *Помилки, пов'язані з недосконалістю алгоритму розпізнавання, тобто* помилки, які є «внутрішніми» з точки зору системи розпізнання об'єктів і які можуть проявлятися навіть при ідеальному функціонуванні інших підсистем.

2. *Помилки через дефекти попередньої обробки.* Як правило, система розпізнавання одиночних зображень є однією з підсистем певного комплексу, а зображення, що подаються в систему розпізнавання, формуються в результаті дії інших підсистем (див. Рис. 2.3). В результаті можуть виникнути помилки, пов'язані з недосконалістю попередніх підсистем. Наприклад, нехай в результаті розщеплення зображення текстового рядка на зображення окремих символів була зроблена помилка, в результаті якої положення правої межі зображення латинської літери «P» було знайдено невірно, в результаті чого на зображенні було втрачено перемичку між двома горизонтальними штрихами. Отримане

зображення, з точки зору системи розпізнання одного символа, може бути не відрізнятися від латинської літери «F».

Помилки, викликані навколишнім шумом. Ці помилки виникають, якщо в умовах зовнішнього середовища, в якому розпізнається об'єкт, зображення цього об'єкта стає не відрізнити від зображення об'єкта іншого класу. Наприклад, припустимо, що робиться фотографія документа, що засвідчує особистість, яке містить поле ім'я з істинним значенням «HANNA». Поле вписане на білому тлі, а документ покритий захисною поверхнею блиску. На момент зйомки документ показав відблиски від зовнішнього джерела світла, повністю прикривши букву «H» і залишивши зображення інших букв незмінними. Таким чином, зображення цього поля буде не відрізнити від зображення поля «АННА» на аналогічному документі.

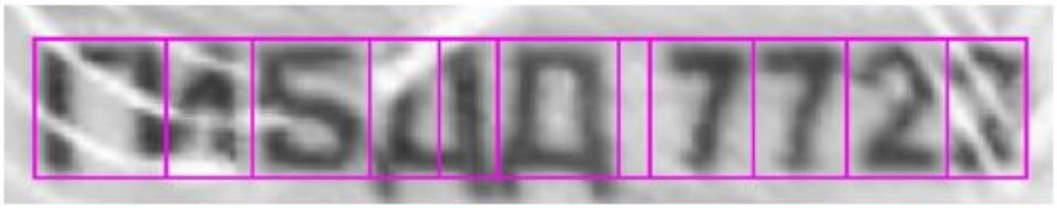


Рисунок 2.3 — Приклад помилкової сегментації текстового рядка на окремі символи в умовах розмитого зображення і дефектів, пов'язаних із захисним голографічним шаром документа.

По відношенню до єдиної системи розпізнавання зображень помилки, пов'язані з шумом навколишнього середовища або дефектами попередньої обробки, є результатом спотворення вхідного зображення. При можливості використання декількох спостережень об'єкта можна очікувати, що вплив навколишнього середовища і дефекти попередньої обробки на ці спостереження будуть різними. Однак навіть при фіксації системи розпізнавання одного об'єкта, незалежно від попередньої обробки, через недосконалі моделі класифікації залишаються помилки[6]. Сучасні

дослідження показують, що найбільш ефективним методом розпізнавання зображень є згорткові нейронні мережі, що в ряді окремих завдань показують результати, здатні конкурувати з людиною, але все ж можуть показати нестабільний результат з мінімальними змінами вхідного зображення, навіть якщо ці зміни стосувалися лише одного пікселя. Таким чином, навіть використовуючи найточніший метод розпізнавання, але маючи єдине вхідне зображення об’єкта, неможливо відокремити корисний сигнал від шуму, вплив якого може бути кардинально змінити результат.

Таким чином, розглядаючи цифрове зображення об’єкта не єдине зображення, а як відеопотік, стає можливим зменшити виникнення помилок через мінливість шуму по відношенню до окремих кадрів відеопотоку, якими класичні системи розпізнавання об’єктів не мають.

Одним з методів аналізу декількох зображень однієї і тієї ж сцени з метою зменшення впливу шуму оптичної системи і дефектів, пов’язаних з неконтрольованими умовами зйомки, є техніка «супер-роздільної здатності» - це процес отримання зображення з високою роздільною здатністю з декількох зображень одного і того ж об'єкта з більш низькою роздільною здатністю. Цьому завданню приділялася велика увага в літературі і запропоновано велику кількість підходів з урахуванням специфіки кінцевого завдання обробки зображень і розпізнавання об'єкта або сцени. Однак, як зазначалося раніше, подальша обробка отриманого єдиного зображення об'єкта залишається схильною до помилок в алгоритмі розпізнавання, зокрема, через нестабільність згорткових нейронних мереж.

2.2 Модель системи розпізнавання об’єктів у відеопотоці

Розглянемо модель єдиної системи розпізнавання об’єктів. Нехай буде дана множина, яка містить K класів $C = \{c_1, c_2, \ldots, c_K\}$. Наприклад, при розгляді завдання розпізнавання окремих символів поля «Прізвище»

громадянина України багато класів є українським алфавітом з доданими до нього символами пробілу і дефісів. Після локалізації його меж і проектної корекції заданим класом може бути сукупність типів сторінок документів, доступних для подальшої обробки. Окремо слід зазначити, що іноді в задачах розпізнавання об'єктів і явищ допускається мати «порожній клас», який повинен бути відповіддю системи розпізнавання на вхідне зображення об'єкта, про який система не знає, або на зображення, яке не містить об'єкта[7].

Нехай зображення об'єкта $I()$ з усіх можливих зображень I і в рамках моделі взаємодії системи розпізнавання з користувачем/оператором (або з іншими компонентами системи) існує клас $c^*() \in C$, до якого належить об'єкт. Задача розпізнання зображення одного об'єкта полягає у визначенні цього класу. Результат системи розпізнавання в загальній формі представлений як певне відображення з набору класів C до набору оцінок під приводом: $\hat{f} : C \rightarrow$ R. Враховуючи, що багато класів C містять рівно K елементів: (2.1)

$$f\,(I(x)) = \{(c1, q1),(c2, q2), \ldots ,(cK, qK)\}\ ,$$

де $qi \in$ R, $i \in \{1, \ldots , K\}$ – реальні оцінки приналежності до об'єкта до класу $c \in C$, за умови, що спостерігається зображення об'єкта $I()$. В якості кінцевого рішення класифікації приймається клас $c^*\ (I(x)) = \arg\max f(I(x))$. Тривіальна схема системи розпізнавання, наприклад, описаної моделі, представлена на рис. 2.4.

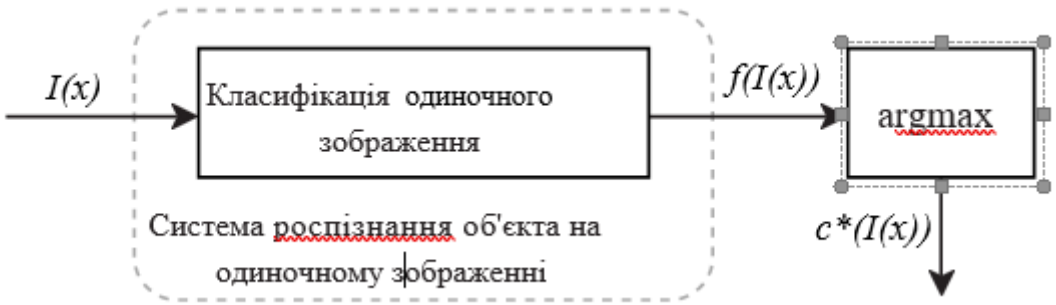


Рисунок 2.4 — Тривіальна схема системи розпізнавання єдиного об’єкта.

Якщо виключити з розгляду процес валідації результатів знань і процес навчання параметрів системи розпізнавання (якщо використовуються методи машинного навчання, наприклад, штучні нейронні мережі, для вирішення задачі класифікації), і розглянути безпосередньо процес розпізнавання, то система розпізнавання статична і не передбачає зворотного зв’язку.

Розглянемо тепер завдання розпізнавання об’єкта у відеопотоку. Відеопотік генерується за допомогою якогось пристрою захоплення, який представляє послідовність кадрів, кожен з яких є незалежним зображенням об’єкта. В умовах фіксованої кількості кадрів можна розглядати проблему розміщення об’єкта у відеопотоку як статичну систему, подібну до тієї, що представлена на рис. 2.4,але з більш складною моделлю входу. Тоді послідовність з N кадрів може розглядатися як сукупність багатьох зображень об’єктів x: $I(x) = \{I1(x), I2(x), \ldots, IN(x)\} \subset I$. При цьому модель виходу системи залишається незмінною.

Реалізація такої системи може відрізнятися в підходах до інтеграції даних. Можливо, тривіальний розгляд процесу класифікації як «чорного ящика», який обробляє відразу багато зображень (діаграма на рис. 2.5(а)). Інші варіанти здійснення частково або повністю використовують методи розпізнавання окремих зображень об’єкта та здійснюють інтеграцію на рівні вхідних зображень (рис. 2.5(б)) або на рівні результатів розпізнавання виникнення кожного окремого зображення (рис. 2.5в).

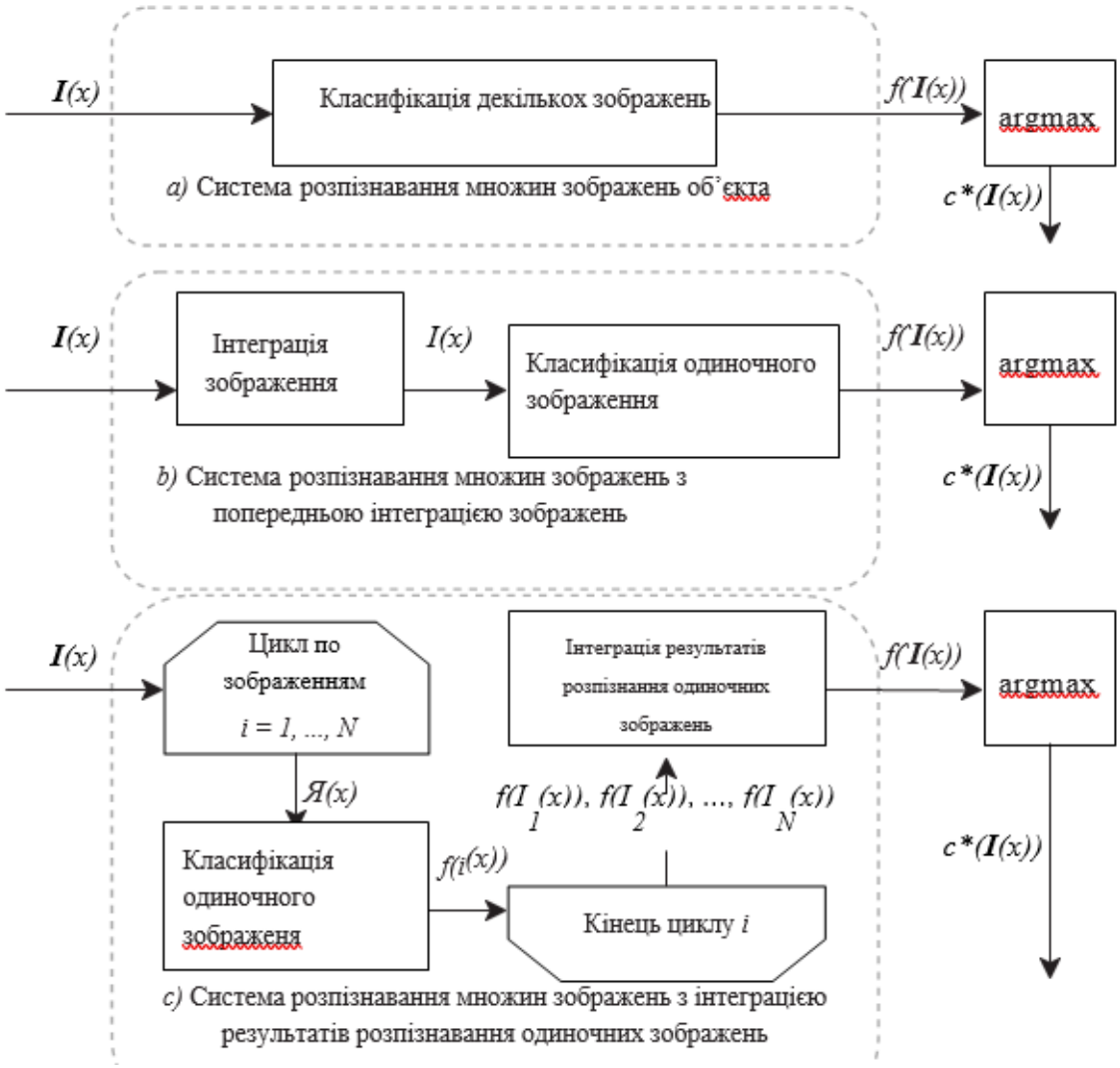


Рисунок 2.5 — Варіанти статичних систем розпізнавання множин зображень об’єкта.

Однак представлені статичні моделі системи розпізнавання об’єктів у відеопотоку не повною мірою відображають сценарій розпізнавання за допомогою мобільного пристрою – так як ці моделі припускають в якості входу лише багато кадрів, без упорядкування, і не передбачають зміни стану системи в процесі зйомки. Також, в умовах апаратних обмежень мобільних пристроїв, зберігання та обробки кількох зображень можуть бути недоречною або неможливою. Для більш точного дотримання процесу розпізнавання об’єкта у відеопоці мобільного пристрою пропонується розглянути динамічну модель з дискретним часом.

Для цілей формалізації уявимо відеопотік як послідовність створених у

часі об’єкта. Таким чином, дискретний час встановлюється t = 0, 1, 2 ,. . . та відеопотік, що містить зображення спостережуваного об’єкта $I_t(x) \in I$. Ця дискретна модель відеопотоку відповідає принципам представлення закодованого відеопотоку в програмних системах[8].

Щоб визначити систему розпізнавання об’єктів у відеопотоку, який генерується незалежно, необхідно визначити сервісну модель, яка є проміжним шаром між відеопотоком і потоком, обробленим системою розпізнавання. Найбільш тривіальна – це схема обслуговування, при якій зображення, згенеровані під час обробки попередньою системою розпізнавання зображень, скидаються. Якщо є можливість зберігати колекцію зображень, альтернативною моделлю є буферизована схема обслуговування, яка дозволяє накопичувати і видавати вхідні зображення за запитом системи в довільний момент часу, без обмежень, пов’язаних з дискретизацією генерації зображень джерелом. З точки зору самої системи розпізнавання послідовностей зображень набір методів і алгоритмів розпізнавання та інтеграції результатів не залежить від сервісної схеми, тому в рамках цієї роботи в майбутньому ми будемо вважати, що існує «поточне» зображення $I_t(x)$, а зображення можуть бути скинуті під час перезавантаження системи.

Система розпізнавання підтримує деякий внутрішній стан $s_t \in S$, змінний у часі. Час Δt, необхідний для отримання оновленого результату після введення наступного зображення $I_t(x)$, загалом, є функцією від зображення та внутрішнього стану системи: $\Delta t = \Delta\ (I_t(x),\ s_t)$, які може бути необчислювана в момент часу t. Результат роспізнання, який враховує інформацію, яка міститься на зображенні, яка була захоплена під час часу t, може бути доступна лише під час часу $T(t) = t + \Delta t$.

У початковий момент часу $t = 0$ ініціалізується внутрішній стан системи s_0. Нехай під час t захоплюється зображення $I_t(x)$, яке подається до модуля розпізнавання f. Результат роспізнання $f(I_t(x))$ стає доступним в момент часу t’ >= t та записується в системний модуль пам’яті (тобто стає частиною стану $s_{t'}$).

Після цього існує поєднання результатів розпізнання зображень об’єкта, накопичених на у поточний момент, і під час часу $T(t) > t'$ виводять результат розпізнавання $R_{T(t)}$. Після відображення результату, відбувається наступне захоплення зображення $I_{T(t)}(x)$ та процес продовжується. Таким чином, результат $R_{T(t)}$ враховує інформацію яка міститься у зображеннях з індексами 0, $T1(0)$, $T2(0)$, . . . , t(під надстрочним знаком функції $T(t)$ мається на увазі не ступінь, а множинну композицію функції). Якість результату характеризується близькістю результату $R_{T(t)}$ до справжнього значення $\nu(x)$ об’єкта x, за деякими метриками. Схема описаної системи розпізнавання представлена на рис. 2.6.

Методи розподілу ознак та класифікації об’єктів, що застосовуються в статичних системах (див. Рис 2.5), також застосовуються в динамічній моделі, однак, динамічна модель системи розпізнавання об’єкта у відеопотоці – це ряд специфічних властивостей. Перш за все, необхідно відзначити посилений ефект ефективності алгоритмів розпізнавання одиночного зображення на вивід системи. Дійсно, зменшення часу Δt, необхідного на розпізнання одного зображення $I()$, дозволяє обробляти більше інформації про об’єкт для такого ж абсолютного часу (тобто в той же час від точки зору користувача/оператора) . Крім того, в рамках такої системи існують завдання[9], які є атиповими для традиційних систем розпізнавання об’єктів на зображенні. Перше таке завдання полягає в тому, щоб отримати результат $R_{T()}$ – проблема поєднання (інтеграції) результатів розпізнання одного й того ж об’єкта з різних зображень у єдиний результат. Друге завдання – зупинити процес розпізнавання – оскільки захоплення зображень не може бути обмежена природньо, на момент часу $T(t)$ існує завдання прийняття рішення про те, що процес захоплення слід припинити та накопичений результат прийняти як остаточний.

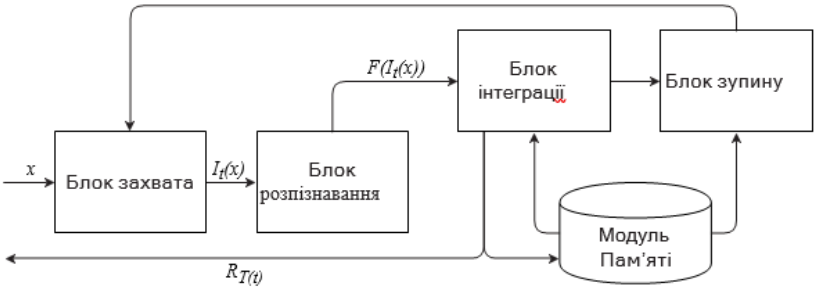


Рисунок 2.6 — Схема системи розпізнавання об’єктів у відеопотоку з зупинкою.

В якості функціональності ефективності системи під час зупинки $t = t_{\text{stop}}$, пропонується розглянути лінійну комбінацію:

$$a \cdot \rho(R_{t_{\text{stop}}}, \nu(x)) + b \cdot W(t_{stop}), \tag{2.2}$$

де a, b – константи, $\rho(R_t, \nu(x))$ – відстань від інтегрованого результату R_t до справжнього значення $\nu(x)$, що характеризує якість результату, а $W(t)$ – це штрафна функція від часу. Спеціальний випадок штрафної функції $W(t)$ – це кількість оброблених зображень:

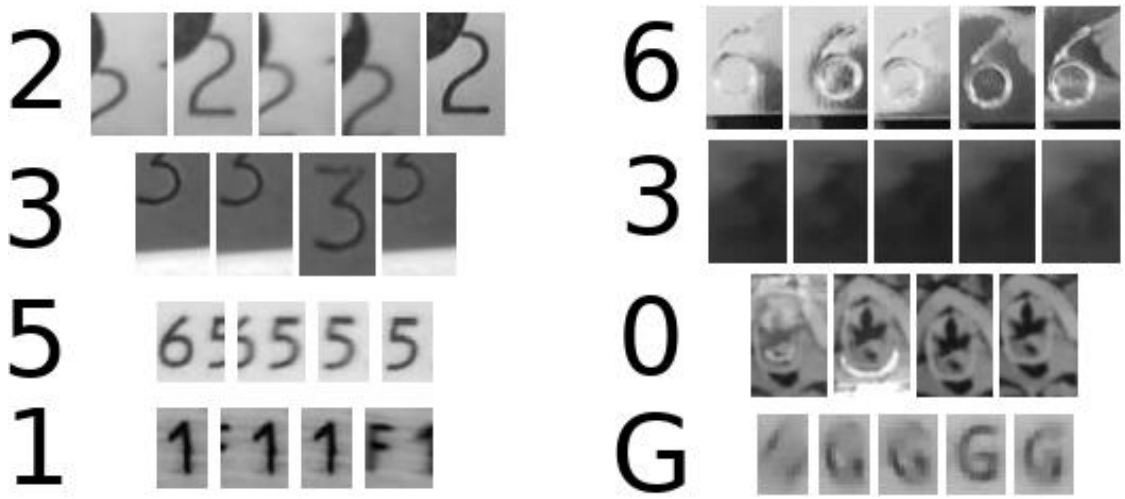
$$W(t) = \max\{i \mid T^i(0) \leqslant t\}. \tag{2.3}$$

2.3 Завдання інтеграції результатів розпізнавання об'єктів

Основним завданням традиційних систем розпізнавання об'єктів є максимізація точності розпізнавання (тобто максимізація частки «правильних» класифікацій об'єктів). *Завдання інтеграції результатів розпізнавання об'єктів* полягає в максимальній точності результату розпізнавання багатьох різних зображень одного і того ж об'єкта при заданих результатах розпізнавання єдиного зображення.

На рисунку 2.7б показані приклади послідовностей, що зображують один і той же об'єкт, які схильні до характерних спотворень, які можна віднести до шуму середовища: спотворення, пов'язані з оптичним контуром

малоформатних цифрових камер, аберації, відблиски і відбиття всередині оптичної системи, цифровий шум, нерівномірне або недостатнє освітлення сцени, расфокусування зображення і «розмиття» через рух оптичного датчика по відношенню до об’єкта, відблиски від зовнішнього джерела освітлення, геометричні спотворення, такі як перспективні спотворення зображення об'єкта або нелінійні спотворення, викликані вигинами середовища, перешкоди, створені голографічним захисним шаром і т.д. На рисунку 2.7а також представлені приклади послідовностей зображень об'єктів, що схильні до дефекту попередньої обробки, в даному випадку помилки в пошуку і локалізації об'єкта на вхідному кадрі, помилки в аналізі структури і локалізації текстових рядків, помилки сегментації текстових рядків на окремі символи[10].



а) б)

Рисунок 2.7 — Приклади послідовностей зображень об'єктів з дефектами попередньої обробки, що генерують зображення (а) і без дефектів попередньої обробки, але при впливі шуму середовища (б)

Для формалізації формулювання задачі інтеграції з точки зору моделі системи розпізнавання об'єктів у відеопотоці, припустимо, що множина

об'єктів $X = \{x_1, x_2, \ldots, x_M\}$ потужності M і набір відеопослідовностей

$$B = \{\mathbf{I}_1(x_{b_1}), \mathbf{I}_2(x_{b_2}), \ldots, \mathbf{I}_H(x_{b_H})\} \quad (2.4)$$

потужності H, де b_h – індекс об'єкта з множини X для кожного h ∈ {1, 2, . . . , H}, та кожна відеопослідовність $I_h(x_{bh}) = \{I_{h1}(x_{bh}), I_{h2}(x_{bh}), \ldots, I_{hNh}(x_{bh})\}$ – послідовність зображень об'єкта $x_{bh} \in X$, які можуть бути схильні до шумів середовища та дефектам попередньої обробки (див. розділ 2.1). Також задано безліч класів $C = \{c_1, c_2, \ldots, c_K\}$ та інформація про ідеальну приналежність кожного об'єкта до відповідного класу $\nu: X \to C$.

Завдання розпізнавання об'єкта у відеопотоці може бути сформульовано як пошук за класифікуючою функцією F: I* → C, що максимізує точність розпізнавання:

$$V_F(B) = \frac{1}{H}\sum_{h=1}^{H}\left[F(\mathbf{I}_h(x_{b_h})) = \nu(x_{b_h})\right] \to \max_F. \quad (2.5)$$

Більш конкретне завдання інтеграції результатів розпізнавання одиночного об'єкта передбачає функцію інтеграції результатів розпізнавання R : $(R^C)^* \to R^C$, перетворення послідовності результатів розпізнавання одиночних зображень в єдиний результат розпізнавання відео (тут R^C - це набір всіляких відображень від набору класів C до набору оцінок R, тобто набору R оцінок різних результатів класифікації). Оскільки кінцевою відповіддю на розпізнавання відео є клас, що відповідає максимальній оцінці розпізнавання $F(\mathrm{I}) = \arg\max R(f(\mathrm{I}))$(див. розділ 2.2), постановка задачі інтеграції будується на основі і приймає форму:

$$V_R(B) = \frac{1}{H}\sum_{h=1}^{H}\left[\arg\max R\left(\hat{f}(\mathbf{I}_h(x_{b_h}))\right) = \nu(x_{b_h})\right] \to \max_R. \quad (2.6)$$

В ідеалі, функція класифікації F або функція інтеграції результатів R повинні бути в змозі фільтрувати викиди, які з'являються в потоці вхідних даних через навколишній шум або дефекти попередньої обробки, і мати можливість здійснювати фільтрацію шуму класифікатора, вирівнюючи випадкові внутрішні помилки.

В ідеальному випадку класифікуюча функція F або функція інтегрування результатів R повинна мати можливість фільтрувати викиди, що з'являються у вхідному потоці даних через шум середовища або дефектів попередньої обробки, і мати можливість проводити фільтрацію шуму класифікатора, нівелюючи внутрішні помилки[11].

Неважко помітити, що підхід до інтеграції як до завдання побудови функції F, що класифікує, можна звести до завдання побудови функції R інтегрування результатів, застосувавши наявний метод класифікації одиночних зображень об'єктів. Альтернативними підходами є, наприклад, техніки «супер-роздільна здатність», що здійснюють піксельне зіставлення безлічі вхідних зображень I() і будують єдине «ідеальне» зображення об'єкта, яке згодом класифікується. Однак варто зауважити, що з огляду на особливості найбільш точного існуючого в даний момент методу класифікації зображень - згорткових нейронних мереж - а саме, його нестійкості до випадкових піксельних спотворень, в рамках цієї роботи будуть розглядатися методи побудови функції інтеграції R результатів розпізнавання одиночних зображень.

У літературі завдання об'єднання результатів класифікації одиночних об'єктів зазвичай у контексті методів отримання більш точної класифікації шляхом об'єднання результатів кількох різних класифікаторів[12]. Залежно від використовуваної моделі результату класифікації об'єкта та від інтерпретації оцінок класифікатора розглядаються різні методи комбінування.

Завдання комбінування результатів класифікації об'єктів можна

розглядати як завдання колективного прийняття рішення. Введемо поняття предиктора достовірності результату класифікатора як речовиннозначну функцію ($I()$, $f\hat{}$), що відображає ступінь впевненості в тому, що результат класифікації зображення $I()$ функцією $f\hat{}$ буде вірним[13]. Як предиктор має сенс використовувати обчислювані характеристики виразів, що явно впливають на точність класифікації, такі як оцінка змащування та рівня фокусування, оцінка рівня шуму, артефактів оцифровки та ін. (такі предиктори можна вважати апріорними, оскільки вони спираються безпосередньо на характеристики вхідних зображень). Інший клас предикторів обумовлюються значеннями оцінок класифікації (апостеріорні предиктори), пов'язані з поняттям оцінки достовірності результату розпізнавання. Прикладом широко використовуваного апостеріорного предиктора достовірності є значення оцінки першої (максимальної) альтернативи:

$$p(I(x), \hat{f}) = \max \hat{f}(I(x)). \tag{2.7}$$

Нехай поставлений певний предиктор достовірності. Тоді задача інтеграції результатів розпізнавання послідовності I(x) ={$I1(x)$, . . . , IN (x)} потужності N може бути розглянута як завдання колективного прийняття рішення з N експертами, оцінки рівнів компетентності яких є функціями від значень предиктора достовірності. Варто зауважити, що рівні компетентності експертів у даній моделі є відображенням вхідних даних – так як саме характеристики окремих спостережень (тобто окремі зображення $I1()$, . . . , IN ()) необхідні для оцінки важливості експертів.

Важливим питанням у рамках цього завдання є питання доцільності використання голосування кількох експертів замість використання думки самого компетентного експерта[14]. Переходячи до задачі це питання формулюється так: за яких моделей вхідних даних у задачі комбінування результатів розпізнавання слід вибирати ту чи іншу стратегію комбінування?

Для відповіді на це запитання пропонується провести експериментальне

дослідження. Було підготовлено чотири набори даних, характеристики яких наведені в таблиці 2. Набори даних MRZ-MSEGM і MRZ-CLEAN містять відеопослідовність результатів розпізнавання символів машино-читаної зони міжнародних документів. Набори даних ICN-MSEGM та ICN-CLEAN містять відеопослідовність результатів розпізнавання символів поля «Номер» платіжних банківських карток, виконаного за допомогою індент-друку. Зображенням символів у аналізованих тестових наборах властивий широкий спектр спотворень: нерівномірна або недостатня освітленість, цифровий шум, розфокусування та «змазаність» через рух оптичного сенсора щодо носія, відблиски від зовнішнього джерела світла та перешкоди, створювані голографічним захисним шаром документа[15]. Розпізнавання кожного окремого образу символу отримано за допомогою згорткових нейронних мереж, навчених окремо для символів машини зони та для символів поля «Номер» платіжних банківських карток, на окремих навчальних наборах зображень із застосуванням методу ауґментації даних. Набори даних MRZ-MSEGM та ICN-MSEGM містять помилки, викликані некоректною або недостатньо точною роботою алгоритмів локалізації документа та алгоритмів сегментації текстових рядків. Набори MRZ-CLEAN та ICN-CLEAN є підмножинами відповідних на борів MRZ-MSEGM та ICN-MSEGM, що не містять подібних помилок. Таким чином, у наборах даних MRZ-CLEAN та ICN-CLEAN кожна відеопослідовність містить образи строго одного і того ж символу, без будь-яких дефектів сегментації.

Таблиця 2 - Характеристики тестових наборів даних MRZ-MSEGM, MRZ-CLEAN, ICN-MSEGM і ICN-CLEAN.

Характеристики набору даних	**MRZ-MSEGM**	**MRZ-CLEAN**
Сила множини класів C	37	
Загальна кількість зображень символів	637874	631530

Точність розпізнавання окремих зображень, %	96.7357	96.8994
Кількість відеорядів	7581	7508
Мінімальна довжина I(x)	3	

Таблиця 2 - Характеристики тестових наборів даних MRZ-MSEGM, MRZ-CLEAN, ICN-MSEGM і ICN-CLEAN.

Максимальна довжина I(x)	223	
Середня довжина I(x)	21	
Характеристики набору даних	**ICN-MSEGM**	**ICN-CLEAN**
Сила множини класів C	10	
Загальна кількість зображень символів	31580	29166
Точність розпізнавання окремих зображень, %	90.9816	96 8936
Кількість відеорядів	1898	1748
Мінімальна довжина I(x)	3	
Максимальна довжина I(x)	25	
Середня довжина I(x)	12	

На представлених тестових наборах даних проведено порівняння базових стратегій комбінування класифікаторів, представлених у оглядовому розділі: правило добутку (1.2), суми (1.3), мінімуму (1.4), максимуму (1.5) та медіани (1.6). Точність розпізнавання відеопослідовності символу є відносна частка відеопослідовностей, для яких ідеальна відповідь збігається з класом, який отримав максимальну оцінку згідно з тим чи іншим правилом комбінування. Додатково проведено порівняння базових правил комбінування з методом голосування (1.1), узагальненим таким чином:

$$\mathrm{Vote}(\alpha)(\hat{f}(\mathbf{I}(x)))(c) = \\ = \alpha \cdot \frac{1}{N}\sum_{i=1}^{N} 1_{\mathbf{I}_c(x)}(I_i(x)) + (1-\alpha)\cdot \max_{i=1}^{N}\left(1_{\mathbf{I}_c(x)}(I_i(x))\cdot p(I_i(x),\hat{f})\right), \quad (2.8)$$

де $\mathrm{I}c(x) = \{I(x) \in \mathrm{I}(x) \mid f(I(x)) = c\}$ – підмножина елементів відеопослідовності, для яких вибором класифікатора є клас c, $1_{\mathrm{I}}c(x)(I(x))$ – індикаторна функція приналежності образу $I(x)$ до підмножини $\mathrm{I}c(x)$, а $p(I(x),\hat{f})$ – предиктор достовірності. Як предиктор достовірності використовувався апостеріорний предиктор «правило першої альтернативи» (2.7)

На малюнку 2.8 представлені порівняльні значення точності розпізнавання відеопослідовностей з використанням правил комбінування на тестових наборах даних MRZ-MSEGM, MRZ-CLEAN, ICN -MSEGM та ICN-CLEAN. Горизонтальна вісь графіків відповідає значенням параметра правила комбінування. Точність розпізнавання з використанням інших правил комбінування представлені горизонтальними лініями.

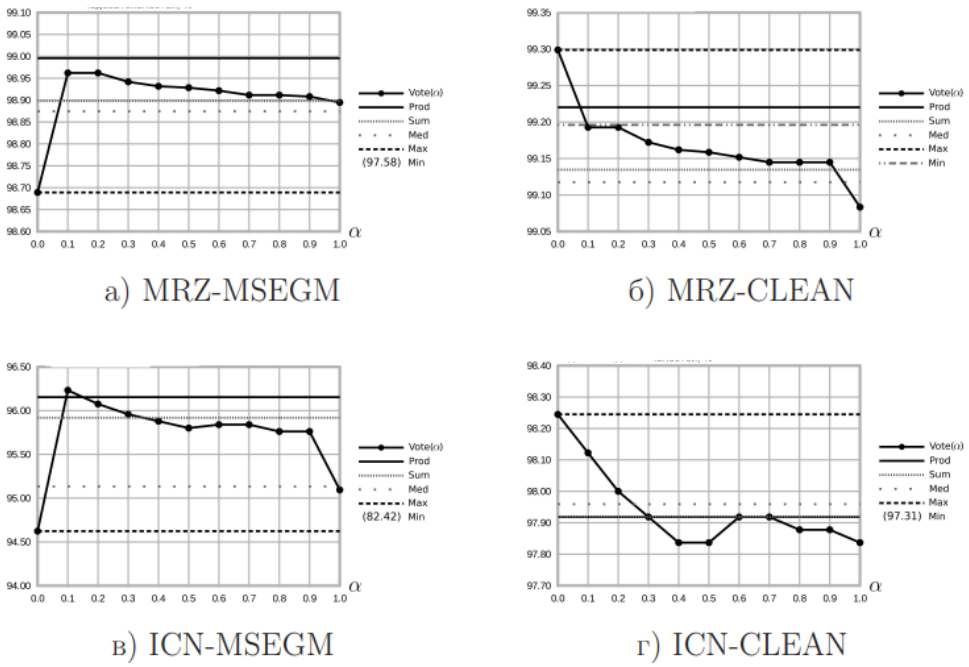


Рисунок 2.8 - Порівняння точності розпізнавання відеопослідовностей символів з використанням базових стратегій комбінування

На малюнку 2.8 продемонстрована значна різниця в оптимальному виборі

стратегії комбінування залежно від моделі вхідних даних: на тестових наборах, в яких зустрічаються помилки локалізації та сегментації символів, більш високу точність розпізнавання відеопослідовностей забезпечують правило добутку, голосування та правило суми (рис. 2.8а, 2.8в). При цьому на тестових наборах, в яких такого типу помилки були виключені (рис. 2.8б, 2.8г), більш високу точність розпізнавання забезпечує правило максимуму. Іншими словами, при розгляді даного завдання як завдання колективного прийняття рішення, у разі більш суворої моделі вхідних даних (з відсутністю помилок локалізації та сегментації символів) вигідніше довіряти єдиному компетентному експерту, ніж колективній думці кількох експертів[17].

За наявності помилок локалізації та сегментації символів стійкість предикторів достовірності зменшується, що у свою чергу збільшує різницю між оцінками компетентності експертів (які конструюються на основі значень предикторів) та дійсними значеннями компетентності експертів (які відповідають апостеріорним ймовірностям прийняття правильного рішення). У такому разі вибір експерта з максимальним рівнем компетентності частіше буває помилковим і таким чином різниця між рівнем дійсної компетентності обраного експерта та рівнями компетентності інших експертів скорочується. Таким чином оптимальність вибору найкращого (з точки зору стійкого предиктора відеопослідовності достовірності) відсутні покадрові помилки результату локалізації і у разі, коли в сегментації символів, відповідає ширшому положенню теорії колективного прийняття рішення, згідно з яким порушення першої частини затвердження Кондорсе (при збільшенні кількості експертів ймовірність колективного прийняття правильного рішення збільшується, якщо для кожного експерта ймовірність прийняття правильного індивідуального рішення вище, ніж ймовірність прийняття неправильного рішення) відбувається при збільшенні різниці між рівнями компетентності максимально компетентного експерта та інших.

З результатів проведеного експерименту можна зробити висновок, що у разі побудови системи розпізнавання об'єкта у відеопотоці для вибору стратегії

комбінування результатів необхідно керуватися не тільки моделлю результатів розпізнавання об'єкта, але й моделлю шуму вхідних даних. При цьому, у разі фіксованої моделі шуму вхідних даних для інтеграції результатів класифікації одиночних об'єктів можна користуватися результатами досліджень, які були спрямовані на комбінування різних класифікаторів з метою максимізації точності розпізнавання одного об'єкта. У той самий час, пряме застосування розглянутих правил комбінування неможливе у разі, якщо модель результату розпізнавання об'єкта складніша, ніж простий результат класифікації. Як приклад такого об'єкта можна назвати текстовий рядок, на який проводиться класифікація кожного символу незалежно.

2.4 Задача зупинки

Модель системи розпізнавання об'єкта у відеопотоку (див. рис. 2.6) не передбачає обмеження на кількість вхідних зображень, а оскільки основною метою системи розпізнавання об'єктів є автоматизація введення, важливим параметром є абсолютний час (тобто час з точки зору оператора) , необхідний для отримання остаточного результату розпізнавання. На відміну від процесу зйомки фотографії, відеопотік природно не обмежений у часі. Звідси випливає завдання зупинки, яке полягає у прийнятті рішення про те, що знову отриманий результат $f(\{I0(x), IT1(0)(x), IT2(0)(x), \ldots, It(x)\})$ в момент часу $T()$ можна вважати остаточним і цикл захоплення зображень можна припинити. При розпізнаванні складних об'єктів, які складаються з безлічі незалежно розпізнаваних об'єктів, рішення про зупинення розпізнавання окремих об'єктів впливає на час Δ, необхідний для розпізнавання складеного об'єкта, а значить і на кількість інформації, що обробляється в рамках загальної системи. Таким чином, завдання зупинки (тісно пов'язане із завданням інтеграції) є важливим аспектом системи розпізнавання у відеопотоку, особливо в рамках взаємодії з іншими підсистемами, об'єктом розпізнавання яких у сукупності є складовий об'єкт, такий як текстове поле або документ в цілому.

У найпростішому вигляді правило зупинки можна представити у вигляді предиката, що діє на відеопослідовності: : I * → {0, 1}. Істинність предикату тягне за собою зупинку процесу захоплення і розпізнавання зображень:

$$P\left(\{I_1(x), I_2(x), \ldots, I_n(x)\}\right) = \begin{cases} 1: & \text{Рішення про зупинку} \\ 0: & \text{Продовження роботи} \end{cases}$$

Нехай ь I(x) = {$I_1(x)$, $I_2(x)$, . . . , I_N (x)}— послідовність зображень об'єкта ∈ X, а $I^{(n)}(x)$ = {$I_1(x)$, $I_2(x)$, . . . , $I_n(x)$} ⊆ I(x)— префікс цієї послідовності, що має довжину ≤ N . Позначимо через D_P(I(x)) кількість зображень, які будуть оброблені системою розпізнавання до спрацювання правила зупинки (2.9):

$$D_P(\mathbf{I}(x)) = \min\left[N, \min\left\{|\mathbf{I}^{(n)}(x)| \;\middle|\; n \in \{1, 2, \ldots, N\} \wedge P(\mathbf{I}^{(n)}(x))\right\}\right]. \quad (2.9)$$

З урахуванням правила зупинки при обробці відеопослідовності I() на розпізнавання подаються тільки зображення з підпослідовності I(P)() = $I^{(P)}(x)$ = $I^{(DP(I(x)))}(x)$, і вихідний набір відеопослідовностей набуває вигляду $B^{(P)}$ = {$I_1^{(P)}$ (x_{b1}), $I_2^{(P)}$(x_{b2}), . . . , $I_H^{(P)}$ (x_{bH})}.

Для формалізації задачі зупинки скористаємося загальною моделлю[18] взаємодії системи розпізнавання з користувачем, яка використовується в задачах визначення достовірності результату розпізнавання об'єкта та для оцінки ефективності роботи системи використовує функціонал, описаний в економічних термінах. Нехай W_c - вартість введення коректного результату розпізнавання об'єкта, W_e - вартість введення помилкового результату, і W_f - вартість розпізнавання одного зображення об'єкта. Тоді функція ефективності правила зупинки може бути записана у вигляді середньої вартості роботи системи:

$$W_{F,P}(B) = W_c \cdot V_F(B^{(P)}) + W_e \cdot \left(1 - V_F(B^{(P)})\right) + W_f \cdot \frac{1}{H}\left(\sum_{h=1}^{H} D_P(\mathbf{I}(x))\right), \quad (2.10)$$

де $V_F(B^{(P)})$ — точність розпізнавання відеопослідовностей з урахуванням зупинки за правилом P (аналогічно у разі інтеграції результатів розпізнавання одиночних об'єктів точність обчислюється відповідно).

Спрощуючи вираз і зважаючи на константність W_e приходимо до загальної постановки задачі зупинки як до завдання пошуку правила зупинки, що оптимізує функціонал ефективності:

$$W_{F,P}(B) = V_F(B^{(P)}) \cdot (W_c - W_e) + W_f \cdot \frac{1}{H}\left(\sum_{h=1}^{H} D_P(\mathbf{I}(x))\right) \to \min_P . \quad (2.11)$$

Аналогічний функціонал ефективності будується з урахуванням функціоналу точності в рамках завдання інтеграції результатів розпізнавання одиночних об'єктів.

У контексті розпізнавання об'єктів у відеопотоку завдання зупинення процесу розпізнавання є досить новим і маловивченим.

2.5 Висновки з розділу

У цьому розділі були показані характеристики завдання розпізнавання об'єкта в видеопотоці. Представлені різні способи формалізації системи розпізнавання у відеопотоку та побудована модель динамічної системи з модулем комбінування покадрових результатів розпізнавання та модулем зупинки. Були показані властивості динамічної системи розпізнавання об'єктів у відеопотоці та запропоновано постановки задачі інтеграції результатів розпізнавання кількох спостережень одного і того ж об'єкта та завдання зупинки в контексті таких систем.

Завдання інтеграції (комбінування) результатів розпізнавання кількох спостережень одного й того ж об'єкта розглянуто як завдання колективного прийняття рішення. Показано, що модель вхідних даних може проводити вибір оптимальної стратегії комбінування. Відповідно до проведеного експериментального дослідження, на тестових наборах даних, у яких

зустрічаються дефекти попередньої обробки зображення, такі правила комбінування як голосування та правило добутку. У той же час на тестових наборах, в яких такого типу помилки були виключені, більш високу точність розпізнавання відеопослідовності забезпечує правило максимуму. У термінах колективного прийняття рішення, у разі більш суворої моделі вхідних даних (тобто при розпізнаванні множини зображень з мінімальним внеском дефектів попередньої обробки) вигідніше довіряти єдиному найбільш компетентному експерту, ніж колективній думці кількох експертів.

Описано завдання зупинення процесу розпізнавання об'єкта у відеопотоці, що виникає через відсутність обмеження на кількість одержуваних спостережень у часі. Завдання є новим стосовно систем оптичного розпізнавання об'єктів.

3. ІНТЕГРАЦІЯ РЕЗУЛЬТАТІВ РОЗПІЗНАВАННЯ РЯДКОВОГО ОБ'ЄКТА У ВІДЕОПОТОЦІ

3.1 Вступ

Розпізнавання таких об'єктів як параграфи тексту, текстові рядки, поля документів тощо, пов'язане з набором складнощів, якщо джерелом зображення є камера мобільного пристрою. У подібних умовах зйомки зображень характерні спотворення, такі як дефокусування, змазування, відблиски на світловідбивних поверхнях, недостатня роздільна здатність для достатньої точності алгоритмів розпізнавання символів та інші. На малюнку 3.1 представлений приклад відблиску на документі та його впливу на зображення текстових полів, витягнутих із послідовних кадрів відеопотоку.

Рисунок 3.1 — Фрагмент кадру з відблиском на поверхні, що відбиває документ (ліворуч) і витягнуті зображення текстових полів на кадрах відеопотоку (праворуч). Зображення з пакета даних MIDV-500

Однією з переваг використання відеопотоку при розпізнаванні об'єктів є можливість обробки множини кадрів у реальному часі, тобто розпізнавання одного і того ж об'єкта багаторазово, таким чином збільшуючи фінальну

точність розпізнавання[19]. Варто також зазначити, що вибір єдиного найкращого результату в деяких випадках може бути прийнятною стратегією, оскільки у відеопотоці документа може бути кадру з цілком видимим об'єктом. Таким чином, з'являється необхідність у вивченні методом комбінування кількох результатів розпізнавання.

Цілями даного розділу є побудова моделі результату розпізнавання рядкового об'єкта, що враховує альтернативні варіанти класифікації одиночних об'єктів, і на її основі побудова алгоритму інтеграції результатів розпізнавання рядкових об'єктів. У розділі 3.2 буде описана модель результату розпізнавання одиночного та рядкового об'єктів, яка використовується в подальшому для побудови алгоритму. У розділі 3.3 наведено встановлення завдання інтеграції результатів розпізнавання рядкових об'єктів. У розділі 3.4 наводиться запропонований алгоритм, і розділ 3.5 представлено його експериментальне дослідження.

3.2 Модель результату розпізнавання рядкового об’єкта

Розглянемо модель результату розпізнавання одиночного об’єкта. Нехай відбувається класифікація зображення I деякого об’єкта c на один із K класів з множини $C = \{c1, c2, \ldots, cK\}$ за допомогою модуля класифікації f . У класичній постановці результатом класифікації є один із класів в $f(I) = c_f$, де $c_f \in C$, і завдання розпізнавання одиночного об’єкта полягає в максимізації апостеріорної ймовірності збігу класу c_f з істинним значенням c. У більш спільній постановці модуль класифікації $f\hat{}$ ставить вхідному зображенню I у відповідність безліч пар $f(I) = \{(c_1, q_1),(c_2, q_2), \ldots ,(c_K, q_K)\}$, де q – оцінка належності об’єкта до класу c. Фінальним результатом розпізнавання є клас, що відповідає максимальній оцінці приналежності:

$$f(I) = \arg\max\{\hat{f}(I)\} \in \left\{ c_f \;\middle|\; \left((c_f, q_f) \in \hat{f}(I)\right) \wedge \left(q_f = \max_{(c,q)\in\hat{f}(I)} q\right) \right\}. \quad (3.1)$$

Якщо існує кілька пар (c_{f1}, q_f),(c_{f2}, q_f),… з рівним максимальним значенням оцінки належності, як відповідь для береться один з класів згідно з прийнятою конвенцією (наприклад, клас з мінімальним індексом у множині C). Модель результату розпізнавання одиночного об'єкта є варіантом моделі результату Алгоритмів Обчислення Оцінок (АОО) і є найбільш широко використовуваною моделлю в методах оптичного розпізнавання зображень за допомогою згорткових нейронних мереж[20].

Для визначення результату розпізнавання рядкового об'єкта необхідно ввести поняття порожнього класу λ, що позначає відсутність одиночного об'єкта. Розширеним результатом класифікації одиночного об'єкта вважатимемо відображення $a : C \cup \{\lambda\} \to [0, 1]$ з множини класів, об'єднаної з міткою порожнього класу λ у множину оцінок власності. Кожна оцінка власності є речовим числом від 0 до 1 і сума оцінок власності дорівнює одиниці. Таким чином задається безліч усіляких результатів розпізнавання одиночного об'єкта $\hat{C}$:

$$\hat{C} \stackrel{\text{def}}{=} \left\{ a \in [0,1]^{C \cup \{\lambda\}} \;\middle|\; \sum_{c \in C \cup \{\lambda\}} a(c) = 1 \right\}. \tag{3.2}$$

На безлічі результатів розпізнавання одиночного об'єкта $\hat{C}$ можна задати метрику таким чином:

$$\rho_{\hat{C}}(a,b) \stackrel{\text{def}}{=} \frac{1}{2} \sum_{c \in C \cup \{\lambda\}} |a(c) - b(c)|, \quad \forall a, b \in \hat{C}. \tag{3.3}$$

Легко переконатися, що функція $\rho_{C\wedge}(a, b)$ має властивості дійсної метрики:

1. $\rho_{C\wedge}(a, b) = 0 \Leftrightarrow \forall c \in C \cup \{\lambda\} : a(c) = b(c) \Leftrightarrow a = b$, отже, аксіома тотожності виконується;

2. $\forall c \in C \cup \{\lambda\} : |a(c) - b(c)| = |b(c) - a(c)| \Rightarrow \rho C\wedge(a, b) = \rho C\wedge(b, a)$, отже, аксіома симетрії виконується;

3. $\forall x, y \in \mathrm{R} : |x + y| \leq |x| + |y| \Rightarrow \forall c \in C \cup \{\lambda\} : |a(c) - d(c)| \leq |a(c) - d(c)| + |d(c) - b(c)| \Rightarrow \rho C\wedge(a, b) \leq \rho C\wedge(a, d) + \rho C\wedge(d, b)$, отже, нерівність трикутника також виконується.

Варто зазначити, що метрика $\rho_{C\wedge}(a, b)$ відповідає манхеттенській метриці у просторі векторів над упорядкованою множиною $C \cup \{\lambda\}$. Оскільки для a і b сума значень при всіх $c \in C \cup \{\lambda\}$ дорівнює одиниці, безліч значень функції $\rho_{C\wedge}(a, b)$ є відрізком [0, 1].

Позначимо через ˆλ «порожній результат»:

$$\hat{\lambda} \stackrel{\text{def}}{=} \{(\lambda, 1), (c_1, 0), (c_2, 0), \ldots, (c_K, 0)\}. \qquad (3.4)$$

Результатом X розпізнавання рядкового об'єкта називатимемо рядок над безліччю $C\hat{}\backslash\{\hat{}\lambda\}$, тобто елементом $X \in \mathrm{X}$, де $\mathrm{X} = (C\hat{}\backslash\{\hat{}\lambda\})^*$. Рядок X є послідовністю результатів розпізнавання одиночних об'єктів $X = \mathrm{x_1x_2\ldots x_n}$, де $x_i \in C\hat{}\backslash\{\hat{}\lambda\}$, довжиною рядка $|X| =$ називається кількість елементів у цій послідовності. Позначення $X_{i\ldots\mathrm{j}}$ відноситься до підстроювання рядка X, що включає елементи x_ix_i+1 . . . x_j−1x_j для $1 \leq \mathrm{i} \leq \mathrm{j} \leq \mathrm{n}$. При i > j підрядок $X_{\mathrm{i}\ldots\mathrm{j}}$ відповідає порожньому рядку ˆλ нульової довжини.

Введемо поняття елементарної редакційної зміни T як пари $(a, b) \neq (\hat{}\lambda, \hat{}\lambda)$, де $a, b \in C\hat{}$. Редакційна зміна $T = (a, b)$, стосовно рядка X, відповідає:

1. заміні елемента $x_i = a$ у рядку X на елемент b, якщо $b \neq \hat{}\lambda$;
2. видаленню елемента $x_i = a$ з рядка X, якщо $b = \hat{}\lambda$;
3. вставці елемента b у рядок X, якщо $a = \hat{}\lambda$.

Розглянемо два довільні рядки $X, Y \in \mathrm{X}$ кінцевої довжини. Редакційним приписом називається послідовність елементарних редакційних змін $T_{X,Y} = T_1T_2\ldots T_L$, що переводить рядок X в рядок Y . Вагою редакційного розпорядження вважатимемо суму відстаней (у термінах метрики $\rho_{C\wedge}$) між парами об'єктів, що беруть участь в елементарних редакційних змінах $T = (a, b)$ розпорядження $T_{X,Y}$:

$$w(T_{X,Y}) \stackrel{\text{def}}{=} \sum_{i=1}^{L} \rho_{\hat{C}}(a_i, b_i). \tag{3.5}$$

Метрика на безлічі результатів розпізнавання рядкових об'єктів X задається як мінімальна вага редакційного припису, що переводить один рядок в інший:

$$\rho_{\mathbb{X}}(X, Y) = \min\{w(T_{X,Y})\}. \tag{3.6}$$

Метрика ρ_X може розглядатися як одна з реалізацій Узагальненої Відстані Левенштейна (Generalized Levenshtein Distance), і має властивості дійсної метрики за умови, що $\rho_{C^\wedge}$ ними володіє також. Для розрахунку відстані між двома результатами розпізнавання рядкових об'єктів $\rho_X(X,Y)$ можна скористатися наступною рекурентною схемою. Нехай $d(i,j)=\rho_X(X_{1\ldots i}, Y_{1\ldots j})$ – відстань між префіксами рядків X та Y , що мають відповідні довжини. Тоді:

$$\begin{aligned}
d(0,0) &= 0,\\
d(i,0) &= \sum_{k=1}^{i} \rho_{\hat{C}}(x_k, \hat{\lambda}),\\
d(0,j) &= \sum_{k=1}^{j} \rho_{\hat{C}}(\hat{\lambda}, y_k),\\
d(i,j) &= \min\left\{\begin{array}{l}
\rho_{\hat{C}}(x_i, \hat{\lambda}) + d(i-1, j),\\
\rho_{\hat{C}}(\hat{\lambda}, y_j) + d(i, j-1),\\
\rho_{\hat{C}}(x_i, y_j) + d(i-1, j-1)
\end{array}\right\},
\end{aligned} \tag{3.7}$$

та значенню яке шукаємо метрики $\rho_X(X, Y)$ відповідає значення $d(|X|, |Y|)$.

Варто зазначити, що максимальним можливим значенням метрики $\rho_X(X, Y)$ є максимум довжин рядків X і Y (при використанні $\rho_{C^\wedge}$ як метрика на множині результатів розпізнавання одиночних об'єктів). При цьому, оскільки ρ_X є окремим випадком Узагальненої відстані Левенштейна, існує спосіб побудувати нормалізований варіант цієї метрики, із збереженням аксіом тотожності, симетрії та нерівності трикутника:

$$\bar{\rho}_{\mathbb{X}}(X,Y) \stackrel{\text{def}}{=} \frac{2 \cdot \rho_{\mathbb{X}}(X,Y)}{\alpha \cdot (|X| + |Y|) + \rho_{\mathbb{X}}(X,Y)}, \tag{3.8}$$

де α – максимально можлива вага елементарної вставки чи видалення. Для випадка метрики $\rho_{C\wedge}$: α = max{$\rho C\wedge$ (a, ˆλ), $\rho C\wedge$ (ˆλ, b), a, $b \in C$ˆ} = 1.

Крім Узагальненої відстані Левенштейна існують інші підходи до порівняння рядкових об’єктів, такі як алгоритм динамічної трансформації тимчасової шкали (Dynamic Time Warping, DTW)[21]. У класичній постановці, однак, алгоритм DTW передбачає відповідність граничних компонентів рядкових об’єктів, не передбачає штрафу за вставку та видалення компонентів, і не має властивостей метрики (не гарантує виконання нерівності трикутника).

3.3 Завдання інтеграції результатів розпізнавання рядкового об'єкта

Розглянемо задачу розпізнавання рядкового об'єкта у відеопотоці. На вхід до системи подається послідовність зображень I_1, I_2,...,I_N рядкового об'єкта ν ∈ C*. За допомогою модуля Fˆ розпізнавання рядкового об'єкта на одиночному зображенні кожному з зображень ставиться у відповідність результат розпізнавання Fˆ(I) ∈ X. У рамках аналізованої моделі будемо вважати, що у вихідному результаті розпізнавання рядкового об'єкта оцінки належності, що відповідають порожньому класу λ дорівнюють нулю:

$$\begin{gathered}\hat{F}(I_i) = X_i, \quad X_i \in \mathbb{X}, \\ X_i = x_1^i x_2^i \dots x_{n_i}^i, \\ x_j^i(\lambda) = 0, \quad \forall j \in \{1, \dots, n_i\}.\end{gathered} \tag{3.9}$$

Завдання полягає у комбінуванні результатів X_1, X_2,...,X_N з деякими вагами w_1, w_2,...,w_n в єдиний результат X ∈ X, що мінімізує відстань за деякою метрикою до істинного значення ν. Оскільки X ∈ X є рядком над безліччю Cˆ\{ˆλ}, а ν – рядком над безліччю класів C, для визначення відстані між ними необхідно

провести додаткову конвертацію. Найбільш природним способом є приведення справжнього значення ν у вигляд рядка ν ∈ X:

$$\begin{aligned}
&\boldsymbol{\nu} = \boldsymbol{\nu}_1\boldsymbol{\nu}_2\ldots\boldsymbol{\nu}_{n_{\boldsymbol{\nu}}}, \quad \boldsymbol{\nu}_j \in C \\
&\hat{\boldsymbol{\nu}} = \hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_2\ldots\hat{\boldsymbol{\nu}}_{n_{\boldsymbol{\nu}}}, \quad \hat{\boldsymbol{\nu}}_j \in \hat{C} \setminus \{\hat{\lambda}\}, \\
&\hat{\boldsymbol{\nu}}_j \stackrel{\text{def}}{=} \{(\lambda,0),(c_1,0),(c_2,0),\ldots,(\boldsymbol{\nu}_j,1),\ldots,(c_K,0)\},
\end{aligned} \tag{3.10}$$

і як відстань від інтегрованого результату X до істинного значення ν використовувати відстань $\rho_X(X, \nu\hat{})$, або його нормалізований варіант $\rho\tilde{}_X(X, \nu\hat{})$.

Проте, з погляду практичного застосування, важливою є також можливість отримати фінальний результат розпізнавання рядкового об'єкта (за аналогією з фінальним результатом для одиночного об'єкта). Для отримання фінального результату можна скористатися наступною двоетапною процедурою:

1. На першому етапі кожному компоненту $x_j \in C\hat{}\setminus\{\hat{}\lambda\}$ інтегрованого результату $X = x_1x_2...x_{nx}$ ставиться у відповідність або клас $c_{xj} \in C$ з максимальною оцінкою приналежності $x_j(c_{xj})$, або порожній клас λ, якщо його оцінка (λ) перевищує деякий поріг θ:

$$\bar{x}_j = \begin{cases} \arg\max\limits_{c\in C} x_j(c), & \text{якщо } x_j(\lambda) < \theta, \\ \lambda, & \text{якщо } x_j(\lambda) \geqslant \theta. \end{cases} \tag{3.11}$$

2. На другому етапі з отриманого рядка $x_1\bar{}x_2\bar{}...x_{n\bar{}X}$ видаляються всі компоненти $x_j\bar{}$ = λ. Результуючий рядок $X_\theta\bar{} \in C^*$ можна використовувати як фінальний результат розпізнавання рядкового об'єкта.

Як відстань від інтегрованого результату X до істинного значення ν тепер можна використовувати відстань Левенштейна $\text{levenshtein}(X_\theta\bar{}, \nu)$ або його нормалізований варіант:

$$\rho_L(\bar{X}_\theta, \nu) = \frac{2 \cdot \text{levenshtein}(\bar{X}_\theta, \nu)}{|\bar{X}_\theta| + |\nu| + \text{levenshtein}(\bar{X}_\theta, \nu)}. \quad (3.12)$$

Завдання інтеграції рядкових об'єктів було розглянуто у роботі у контексті розпізнавання мови. Замість інтеграції результатів розпізнавання кількох зображень $I_1,I_2,...,I_N$ за допомогою єдиного модуля розпізнавання $F^\wedge$, розглядається інтеграція результатів розпізнавання одного «зображення» I різними системами розпізнавання $F_1,F_2,...,F_N$. Дані постановки задач можна вважати схожими з точністю до моделі шуму: інтеграція результатів розпізнавання рядкового об'єкта у послідовності спрямована на фільтрацію компоненти шуму у вихідних зображеннях $I_1,I_2,...,I_N$ (обумовленою неточністю вхідних даних, помилками попереджувальної обробки та ін.) та її вплив на результат роботи модуля розпізнавання $F^\wedge$, тоді як інтеграція результатів різних модулів розпізнавання спрямована на фільтрацію шуму, привнесеного самими модулями розпізнавання $F_1F_2...,F_N$.

Крім розпізнавання мови, підхід, представлений, також застосовувався для комбінування безлічі класифікаторів у завданнях оптичного розпізнавання друкованих та рукописних текстів. Підхід, описаний зветься ROVER (Recognizer Output Voting Error Reduction)[24] і передбачає двомодульну схему, представлену малюнку 3.2. На першому етапі модуль вирівнювання призводить всі вхідні рядкові об'єкти до виду рядків однакової довжини, виробляючи відповідні вставки порожнього класу λ оптимальним чином. На другому етапі модуль голосування вибирає клас для кожного компонента результуючого рядка на основі лінійної комбінації частоти виникнення та оцінки достовірності, породженої модулем розпізнавання.

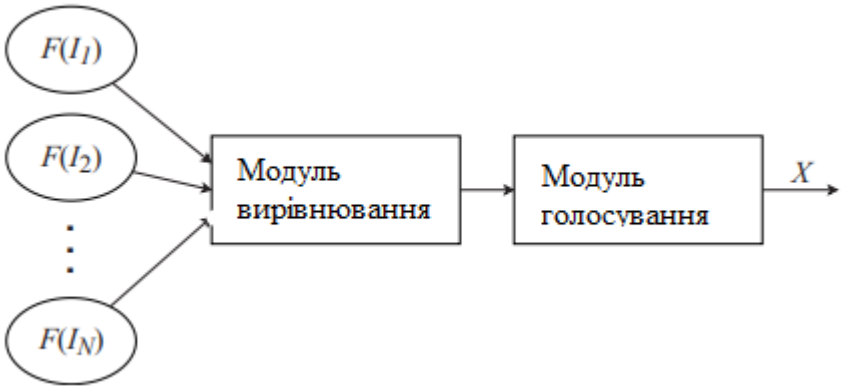


Рисунок 3.2 - Двомодульна схема підходу ROVER

Модель одиночного результату розпізнавання рядка у підході ROVER являє собою пару з рядка над безліччю класів розпізнавання одиночних об'єктів оцінки достовірності модуля розпізнавання, тобто об'єкт з множини $C^* \times R$. Для побудови алгоритму інтеграції результатів розпізнавання рядкового об'єкта з розширеною моделлю одиночного результату, розглянемо становлення задачі вирівнювання рядків.

Нехай задані рядки $X_1,...,X_N$, де $X_i \in X$, і $|X_i| = n_i > 0$:

$$X_1 = x_1^1 x_2^1 \dots x_{n_1}^1$$
$$X_2 = x_1^2 x_2^2 \dots x_{n_2}^2$$
$$\dots$$
$$X_N = x_1^N x_2^N \dots x_{n_N}^N$$

Під вирівнюванням заданої множини рядків розумітимемо функцію align : $\{1,...,N\} \times \{1,...,\max n_i\} \rightarrow \{1,...,\sum_{i=1}^{N} n_i\}$. Функція align$(i, j)$ задає номер компонента вихідного «інтегрованого» рядка, значення якого вносить вклад компонент x^i_j. Для кожного вхідного рядка значення функції align для окремих компонентів рядка різні і зберігають порядок: $\forall i \in \{1,...,N\}$, $\forall j \in \{1,...,n_i - 1\}$: align(i, j) < align$(i, j + 1)$.

Введемо також функцію match: $\{1, \dots, N\} \times \{1, \dots, \sum_{i=1}^{N} n_i\} \rightarrow C\hat{}$, що задається таким чином:

$$\text{match}(i,k) \stackrel{\text{def}}{=} \begin{cases} x_j^i, \text{ якщо } \text{align}(i,j) = k, \\ \hat{\lambda}, \text{ якщо } \nexists j : \text{align}(i,j) = k \end{cases}. \tag{3.13}$$

Завдання вирівнювання полягає у пошуку функції вирівнювання align такою, щоб досягти мінімального значення штрафного функціоналу:

$$\sum_k \sum_{i_1<i_2} \rho_{\hat{C}}(\text{match}(i_1,k), \text{match}(i_2,k)) \to \min, \tag{3.14}$$

відображає сумарну попарну відстань між результатами розпізнавання одиночних об'єктів, що вносять внесок в одні й ті ж компоненти інтегрованого результату.

Для узагальнення модуля голосування (див. рис. 3.2), який вибирає клас для кожного компонента результуючого рядка, введемо сімейство функцій комбінування результатів розпізнавання одиночних об'єктів (N):

$$r^{(N)} : \hat{C}^N \times (\mathbb{R}_0^+)^N \to \hat{C} \setminus \{\hat{\lambda}\}. \tag{3.15}$$

Функція $r^{(N)}$ приймає на вхід N результатів розпізнавання одиночних об'єктів a_1, a_2, . . . , a_N таких, що ∃i : $a_i \neq \hat{\lambda}$, і набір асоційованих з ними невід'ємних ваг w_1, w_2, . . . , w_N, що відображають значимість результату, таких, що $\sum^N_{i=1} > 0$.

Тоді функція інтеграції результатів розпізнавання рядкових об'єктів $R^{(N)}$ набуває вигляду:

$$R^{(N)}(X_1, X_2, \ldots, X_N, w_1, w_2, \ldots, w_N) = r_1^{(N)} r_2^{(N)} r_2^{(N)} \ldots r_{n_R}^{(N)}, \tag{3.16}$$

де $n_R = \max_{(i,j)}\text{align}(i, j)$, а кожен компонент результуючого рядка обчислюється з використанням функції комбінування та відповідно до результату вирівнювання:

$$r_j^{(N)} = r^{(N)}\left(\text{match}(1,j)), \text{match}(2,j), \ldots, \text{match}(N,j), w_1, w_2, \ldots, w_N\right). \tag{3.17}$$

У загальному випадку точне вирішення задачі передбачає розрахунок схеми динамічного програмування (за аналогією зі схемою розрахунку Узагальненої Відстані Левенштейна) з трудомісткістю, що експоненційно залежить від кількості вхідних рядків N (оскільки при розрахунку необхідно використовувати результат підзадач вирівнювання рядків $X_{11}...i_1$, $X_{21}...i_2$,..., $X_{N1}...i_N$) для всіх кортежів (i_1, i_2,...,i_N) ∈ {1, . . . , n_1}× {1, . . . , n_2}×. . .× {1, . . . , n_N}). При розрахунку даної схеми можна використовувати евристичні алгоритми по позову найкоротшого шляху, такі як A*-пошук. У наступному розділі буде представлений алгоритм інтеграції результатів розпізнавання рядкових об'єктів, з апроксимацією функціоналу вирівнювання методом, який використовується у підході ROVER.

3.4 Алгоритм інтеграції результатів розпізнавання рядкового об'єкта

При розрахунку інтегрованого результату розпізнавання рядкового об'єкта породжується набір проміжних інтегрованих результатів $R^{(1)}(X_1, w_1)$, . . . , $R^{(i-1)}(X_1, \ldots, X_{i-1}, w_1, \ldots, w_{i-1})$, де результат $R^{(i-1)}$ використовується для вирішення завдання вирівнювання на кроці i. На першому етапі алгоритму:

$$R^{(1)}(X_1, w_1) = X_1. \tag{3.18}$$

На кожному наступному i-му кроці алгоритму будується оптимальне вирівнювання рядків X_i і $R^{(i-1)}(X_1, \ldots, X_{i-1}, , \ldots, w_{i-1})$ за допомогою схеми динамічного програмування. Нехай $d(l,m)$ = $\rho_X(X_{i1}...l, R^{(i-1)}(X_1, \ldots, X_{i-1}, w_1, \ldots w_{i-1}) 1...m)$, а $P_p(l, m)$ – допоміжні функції для p ∈ {1,2,3}. Розрахунок $d(l,m)$ та $P_p(l, m)$ проводиться відповідно до наступної процедури:

$$
\begin{aligned}
d(0,0)=0, \quad d(l,0)=\sum_{k=1}^{l}\rho_{\hat{C}}(x_k^i,\hat{\lambda}), \quad d(0,m)=\sum_{k=1}^{m}\rho_{\hat{C}}(\hat{\lambda},r_k^{(i-1)}),\\
P_1(l,m)=\rho_{\hat{C}}(x_l^i,\hat{\lambda})+d(l-1,m),\\
P_2(l,m)=\rho_{\hat{C}}(\hat{\lambda},r_m^{(i-1)})+d(l,m-1),\\
P_3(l,m)=\rho_{\hat{C}}(x_l^i,r_m^{(i-1)})+d(l-1,m-1),\\
d(l,m)=\min\{P_1(l,m),P_2(l,m),P_3(l,m)\}.
\end{aligned}
\tag{3.19}
$$

Для розрахунку результату інтеграції на i-му кроці $R^{(i)}(X_1, \ldots, X_i, w_1, \ldots, w_i)$ введемо дві допоміжні функції $t_X : \{0, \ldots, n_i + n_{Ri-1}\} \to \{1, \ldots, n_i\}$ і $tR : \{0, \ldots, n_i + n_{Ri-1}\} \to \{1, \ldots, n_{Ri-1}\}$, розрахунок яких проводиться за наступною рекурентною процедурою:

$$
\begin{aligned}
&t_X(0)=n_i,\\
&t_R(0)=n_{R_{i-1}},\\
&t_X(k+1)=\begin{cases} t_X(k), \text{ якщо } P_2(t_X(k),t_R(k))=d(t_X(k),t_R(k))\wedge \\ \qquad\qquad \wedge P_1(t_X(k),t_R(k))\neq d(t_X(k),t_R(k)) \\ t_X(k)+1, \text{ в інших випадках}\end{cases}\\
&t_R(k+1)=\begin{cases} t_R(k), \text{ якщо } P_1(t_X(k),t_R(k))=d(t_X(k),t_R(k)) \\ t_R(k)+1, \text{ в інших випадках}\end{cases}
\end{aligned}
\tag{3.20}
$$

Інтегрований результат на i-му кроці розраховується так:

$$
\begin{aligned}
&(X_1,\ldots,X_i,w_1,\ldots,w_i)=r_1^{(i)}r_2^{(i)}\ldots r_{n_{R_i}}^{(i)},\\
&=\begin{cases} r^{(2)}\left(r_{t_R(t(k))+1}^{(i-1)},\hat{\lambda},W_{i-1},w_i\right), \text{ якщо } t_X(t(k))=t_X(t(k)-1),\\ r^{(2)}\left(\hat{\lambda},x_{t_X(t(k))+1}^{i},W_{i-1},w_i\right), \text{ якщо } t_R(t(k))=t_R(t(k)-1),\\ r^{(2)}\left(r_{t_R(t(k))+1}^{(i-1)},x_{t_X(t(k))+1}^{i},W_{i-1},w_i\right), \text{ в інших випадках}\end{cases}
\end{aligned}
\tag{3.21}
$$

де $W_i = \sum_{k=1}^{i} w_k$, допоміжна функція $t(k) = n_{Ri} - k + 1$ а функція $r^{(2)}$ – функція інтеграції двох результатів розпізнавання одиночних об'єктів. Слід зазначити, що рамках запропонованого алгоритму від функції інтеграції $r^{(N)}$ потрібна наступна властивість:

$$r^{(N)}(a_1,\ldots,a_N,w_1,\ldots,w_N) = \\ = r^{(2)}(r^{(N-1)}(a_1,\ldots,a_{N-1},w_1,\ldots,w_{N-1}),a_N,w_1+\ldots+w_{N-1},w_N). \quad (3.22)$$

У випадку, якщо функція r не володіє властивістю, процедура вирівнювання залишається незмінною, а інтегрований результат на кроці i необхідно обчислювати для кожного компонента результуючого рядка за формулою, попередньо відновивши функції align і match у явному вигляді.

В рамках даної роботи як функція r пропонується використовувати виважене середнє, що має властивість:

$$r^{(N)}(a_1,\ldots,a_N,w_1,\ldots,w_N)(c) = \frac{1}{W_N}\sum_{i=1}^{N} a_i(c)\cdot w_i, \quad \forall c \in C \cup \{\lambda\}. \quad (3.23)$$

У формі псевдокоду процедура інтеграції результатів розпізнавання рядкового об'єкта представлена як Алгоритм 1. Трудомісткість обчислення функцій $\rho_{C^\wedge}$ і r становить $O(K)$, де K – кількість класів, на яких відбувається класифікація одиночного об'єкта. Оскільки верхня оцінка на довжину результуючої строки R після виконання i-ї ітерації алгоритму становить $O(\sum^{i}_{j=1}|X_i|) \leq O(i \cdot \max^{i}_{j=1}|X_i|)$, трудомісткість кожної ітерації алгоритму можна оцінити як $O(M^2NK)$, де $M = \max^{N}_{i=1}|X_i|$, та загальну трудомісткість Алгоритму 1 як $O(M^2N^2K)$.

3.5 Експериментальні результати

У даному розділі будуть продемонстровані результати експериментального дослідження роботи алгоритму інтеграції результатів розпізнавання рядкових об’єктів, представленого у розділі 3.4. Як об’єкт розпізнавання розглядалося текстове поле документа, що засвідчує особу.

Експериментальне дослідження проводилося на відкритому пакеті даних MIDV-500, що містить відеоролики 50 документів, що засвідчують особу, різних

типів (по 10 відеороликів для кожного документа, по 30 кадрів у відеоролику) з розміченими ідеальними позиціями та значеннями текстових полів. Були проаналізовані 4 групи полів: дати, записані цифрами та розділовими знаками, номер документа, рядки машиночитаної зони (MRZ, Machine-Readable Zone) та компоненти імені власника документи, записані латинським алфавітом.

Розглядалися лише кадри, на яких документ цілком присутній у кадрі (отже відеопослідовності в аналізованій підмножині пакета даних мали різну довжину, від 1 до 30 кадрів). Для того, щоб мінімізувати ефекти нормалізації та забезпечити більш ясне представлення результатів, кожен кліп був доповнений до 30 кадрів шляхом повторення кліпу з початку (отже, всі аналізовані кліпи мали таку ж саму довжину 30).

Кожне поле вирізалося з початкового зображення за допомогою проективного перетворення, відповідно до спільної розмітки ідеальних меж документа та координат текстового поля, з доданими відступами, рівними 30% від найменшої сторони текстового поля. Розмір вирізуваних зображень текстових полів відповідав роздільній здатності 300 точок на дюйм. Кожне вирізане текстове поле розпізнавалася за допомогою мобільного додатка, що відповідає за розпізнавання одиничного текстового рядка, з розширеною моделлю результату.

Як відстань між інтегрованим результатом розпізнавання текстового поля та його істинним значенням використовувалась нормалізована відстань Левенштейна ρ_L між істинним значенням та текстовим рядком, отриманим за допомогою процедури, описаної в розділі 3.3.

Усі порівняння значень символів проводилися незалежно від регістру, і навіть латинська буква «O» вважалася ідентичною цифрі «0».В рамках даного експериментального дослідження Алгоритм 1, що працює в рамках розширеної моделі результату розпізнавання рядкового об’єкта, був порівняний з аналогом, що працює в рамках класичної моделі. Для кожної групи текстових полів і для кожної відеопослідовності проводилася інтеграція методом ROVER, де як вхідні дані використовувалися прості текстові рядки, сформовані процедурою, при зміненні до покадрових результатів розпізнавання. Поріг θ значення оцінки

порожнього символу і для контрольного методу ROVER і для Алгоритму 1 дорівнював 0.6.

На малюнку 3.3 представлені результати роботи порівнюваних алгоритмів для чотирьох груп текстових полів набору даних MIDV-500. Можна відзначити, що для кожної групи полів інтеграція Алгоритмом 1 повних результатів розпізнавання (тобто з урахуванням альтернативних варіантів розпізнавання кожного одиночного символу) досягає меншого значення помилки ніж інтеграція методом ROVER (що враховує тільки перші альтернативи розпізнавання кожного символу), незалежно від довжини послідовності інтегрованих результатів.

На малюнку 3.4 представлені результати роботи алгоритмів для всіх чотирьох груп полів. Досягнуті середні значення відстані між інтегрованим результатом розпізнавання текстового поля та його істинним значенням для різних довжин інтегрованого префікса відеопослідовності представлені в таблиці 3.

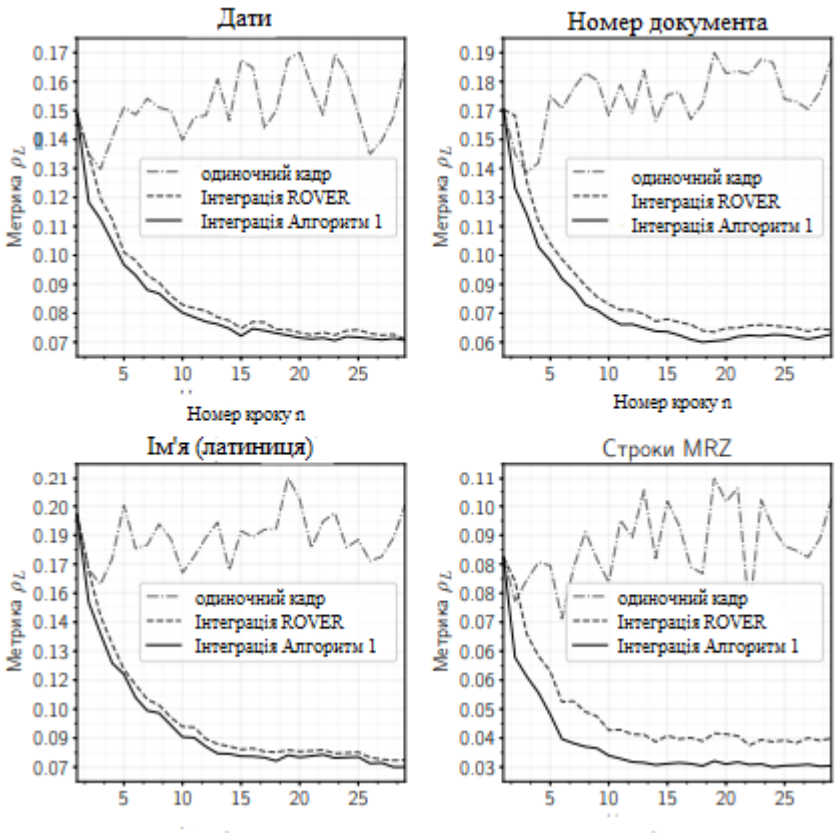


Рисунок 3.3 – Результати роботи алгоритмів інтеграції для чотирьох груп текстових полів набору даних MIDV-500

Таблиця 3 — Досягнута відстань між інтегрованим результатом розпізнавання та істинним значенням без інтеграції, методом ROVER та за допомогою Алгоритму 1

Метод інтеграції	Номер кадру (довжина послідовності інтегрованих результатів)								
	3	6	9	12	15	18	21	24	27
Без інтеграції	0.136	0.154	0.160	0.157	0.168	0.159	0.165	0.166	0.150
Інтеграція методом ROVER	0.125	0.096	0.083	0.075	0.070	0.069	0.069	0.069	0.067
Інтеграція Алгоритмом 1	**0.115**	**0.089**	**0.078**	**0.071**	**0.066**	**0.065**	**0.066**	**0.066**	**0.064**

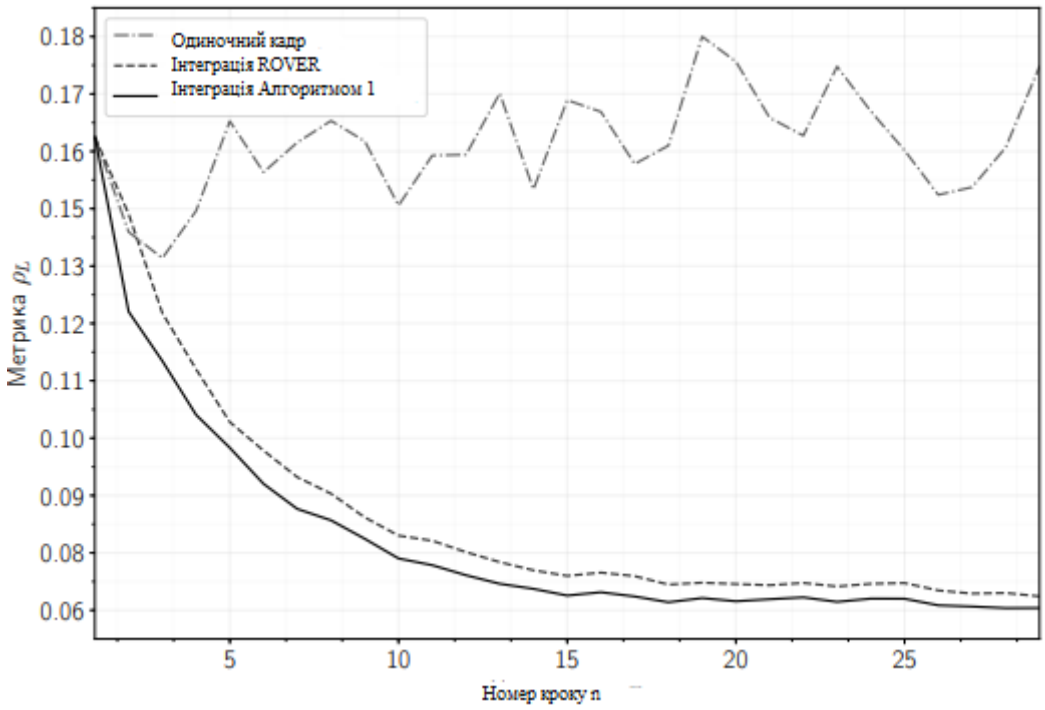


Рисунок 3.4 – Результати роботи алгоритмів інтеграції для текстових полів набору даних MIDV-500

3.6 Висновки розділу

У розділі було розглянуто завдання комбінування кількох результатів розпізнавання рядкового об'єкта з метою збільшення точності фінального результату розпізнавання. Була описана модель результату розпізнавання рядкового об'єкта, що враховує альтернативні варіанти класифікації одиночних об'єктів і був представлений алгоритм інтеграції результатів розпізнавання рядкових об'єктів у рамках описаної моделі.

За результатами проведеного експериментального дослідження можна зробити такі висновки:

1. Методи інтеграції результатів розпізнавання рядкових об'єктів дозволяють досягти значного збільшення точності фінального результату розпізнавання об'єкта під час аналізу відеопослідовності.
2. Метод ROVER, в оригіналі призначений для комбінування результатів розпізнавання одного і того ж образу об'єкта декількома алгоритмами розпізнавання, застосуємо також для комбінування результатів розпізнавання різних образів одного і того ж об'єкта з використанням єдиного модуля розпізнавання.
3. І метод ROVER, що приймає на вхід результати розпізнавання рядкових об'єктів у вигляді рядків над безліччю класів значення одиночних об'єктів, так і Алгоритм 1, що приймає на вхід результати розпізнавання рядкових об'єктів у розширеній моделі, показують значне поліпшення точності інтегрованого результату зі збільшенням кількості використаних кадрів. У задачі розпізнавання текстових полів документів, що засвідчують особу, Алгоритм 1 показує більш високу точність роботи, ніж пряме застосування алгоритму ROVER.

4. За формою графіків залежності відстані між інтегрованим результатом і істинним значенням від кількості використаних кадрів (див. рис. 3.3 і 3.4) можна судити про те, що інтеграція має властивість спадної прибутковості (згідно з термінологією алгоритмів «anytime»). Ця властивість є важливою для вирішення завдання зупинення розпізнавання об'єкта у відеопотоці.

ВИСНОВОК

Основні результати роботи полягають у наступному:

1. Побудовано математичну модель системи розпізнавання об'єкта з відеопотоку з блоком комбінування покадрових результатів розпізнавання та з блоком прийняття рішення про зупинення. Як функціонал ефективності системи була розглянута лінійна комбінація відстані від інтегрованого результату розпізнавання до істинного значення об'єкта і штрафної функції від часу від моменту початку процесу зйомки до зупинки. Дана модель дозволяє розглядати систему розпізнавання об'єкта у відеопотоку як ітераційний обчислювальний процес, який здатний видати у будь-який час найкраще на даний момент рішення, та припинити захоплення нових зображень згідно із заданим правилом зупинки.

2. Виконано оригінальне дослідження впливу характеристик вхідних даних на вибір оптимальної стратегії комбінування покадрових результатів класифікації в рамках завдання розпізнавання одиночного символу відеопотоку. Показано, що якщо в послідовності оброблюваних зображень одиночного зображення відсутні помилки попередньої обробки (такі, як помилки локалізації та сегментації символів), більш високу точність фінального результату забезпечує правило максимальної оцінки. Для відеопослідовностей, у яких зустрічаються помилки локалізації та сегментації символів, вищу точність фінального результату забезпечують правила проведення оцінок, правило голосування і правило суми оцінок.

3. Розроблено новий алгоритм комбінування результатів розпізнавання рядкового об'єкта, який враховує альтернативні варіанти класифікації окремих символів (компонентів рядкового об'єкта). Експериментально показано, що запропонований алгоритм здатний забезпечити більш високу точність інтегрованого результату порівняно з методом інтеграції результатів розпізнавання як рядків над безліччю значень компонентів, стосовно завдання розпізнавання текстового рядка у відеопотоку.

Таблиця 2 - Досягнуті найкращі значення середньої відстані від інтегрованого результату до ідеального значення в момент зупинки; результати розпізнавання інтегровані за допомогою Алгоритму 1

Метод зупинки	Найкраща точність про обмеженні середнього числа кадрів					
	$\leqslant 3$	$\leqslant 4$	$\leqslant 5$	$\leqslant 6$	$\leqslant 7$	$\leqslant 8$
N_{CX}	$\varnothing$	0.083	0.080	0.078	0.073	0.072
N_{CR}	0.096	0.084	0.080	0.077	0.074	0.072
N_K	0.115	0.104	0.097	0.089	0.084	0.082
Алг. 2	**0.092**	**0.082**	**0.076**	**0.073**	**0.072**	**0.070**

4. Метод розроблений виходячи з припущення про те, що завдання зупинки процесу розпізнавання стає монотонною починаючи з деякого кроку. На основі розробленого методу запропоновано новий алгоритм зупинки процесу розпізнавання рядкового об'єкта у відеопотоку, у якому оцінка обчислюється шляхом моделювання наступного інтегрованого результату з використанням накопичених спостережень. Було продемонстровано, що завдання розпізнавання текстових рядків запропоноване правило зупинки є більш ефективнішим, ніж граничне відсічення кількості оброблених кадрів або порогове відсікання розмірів максимального кластеру ідентичних результатів.

5. Спільне використання розроблених алгоритмів (Алгоритм 1 комбінування результатів розпізнавання рядкових об'єктів, що враховує альтернативні варіанти класифікації символів, та Алгоритм 2 зупинки процесу розпізнавання рядка) дозволяє досягти більшої точності розпізнавання при тому ж середній кількості оброблених зображень. У таблиці 7 показані досягнуті найкращі значення середньої відстані від результату до істинного значення за різних обмежень на середню кількість оброблених кадрів з інтегрування результатів Алгоритмом 1 і при використанні розглянутих алгоритмів зупинки, включаючи Алгоритм 2.

ПЕРЕЛІК ПОСИЛАНЬ

1. Як працює комп'ютерний зір// [Електронний ресурс]. – Режим доступу: https://blog.algorithmia.com/introduction-to-computer-vision/
2. Виявлення кордонів Кенні крок за кроком у Python - комп'ютерне бачення// [Електронний ресурс]. – Режим доступу: https://towardsdatascience.com/canny-edge-detection-step-by-step-npython-computer-vision-b49c3a2d8123
3. OpenCV: Detecting Edges, Lines, and Shapes // [Електронний ресурс]. – Режим доступу: https://hub.packtpub.com/opencv-detecting-edges-linesshapes/
4. Smart IDReader: Document Recognition in Video Stream / K. Bulatov[et al.] // 2017 14th IAPR International Conference on Document Analysisand Recognition (ICDAR). Vol. 06. — 2017. — P. 39–44.
5. Sourvanos N., Tsatiris G. Challenges in Input Preprocessing for MobileOCR Applications: A Realistic Testing Scenario // 9th International Conference on Information, Intelligence, Systems and Applications (IISA). —07/2018. — P. 1–5.
6. Zilberstein S. Using Anytime Algorithms in Intelligent Systems // AI Magazine. — 1996. — Sept. — Vol. 17. — P. 73–83.
7. Tamaki M. On the optimal stopping problems with monotone thresholds //Journal of Applied Probability. — 2015. — Vol. 52, no. 4. — P. 926–940.
8. Ferguson T. S., Klass M. J. House-hunting without second moments //Sequential Analysis: Design Methods and Applications. — 2010. — Vol.29. — P. 236–244.
9. Sung Cheol Park, Min Kyu Park, Moon Gi Kang. Super-resolution image reconstruction: a technical overview // IEEE Signal Processing Magazine. — 2003. — Vol. 20, no. 3. — P. 21–36.
10. A Survey: The Methods & Techniques of Super-Resolution Image Reconstruction / A. Semwal [et al.] // International Journal for Scientific Research& Development. — 2017. — Vol. 4, no. 12. — P. 243–249.

11. International standard ISO/IEC 14496-12: Information technology – Coding of audio-visual objects – Part 12: ISO base media file format. —ISO/IEC, 2005. — 94 p.
12. Schwenk H., Gauvain J.-L. Combining multiple speech recognizers usingvoting and language model information // IEEE International Conferenceon Speech and Language Processing. — 2000. — P. 915–918.
13. Ye P., Doermann D. Document Image Quality Assessment: A Brief Survey // 2013 12th International Conference on Document Analysis andRecognition. — 2013. — P. 723–727.
14. Berend D., Kontorovich A. Consistency of Weighted Majority Votes //Proceedings of the 27th International Conference on Neural InformationProcessing Systems. — Montreal, Canada : MIT Press, 2014. — P. 3446–3454. — (NIPS'14)
15. Ding I. J., Yen C. T., Hsu Y. M. Developments of Machine LearningSchemes for Dynamic Time-Wrapping-Based Speech Recognition. // Mathematical Problems in Engineering. — 2013. — P. 542680–1–10.
16. Cazenave T. Overestimation for Multiple Sequence Alignment // CIBCB2007: IEEE Symposium on Computational Intelligence and Bioinformaticsand Computational Biology. — 2007. — P. 159–164.
17. Method of determining the necessary number of observations for videostream documents recognition / V. V. Arlazarov [et al.] // Proc. SPIE.Vol. 10696. — 2018. — P. 10696-10696–6.
18. Smith R. An Overview of the Tesseract OCR Engine // Proceedings of theNinth International Conference on Document Analysis and Recognition -Volume 02. Vol. 2. — IEEE Computer Society, 2007. — P. 629–633. —(ICDAR '07).
19. Merz C. J. Using Correspondence Analysis to Combine Classifiers // Mach.Learn. — Hingham, MA, USA, 1999. — Vol. 36, no. 1/2. — P. 33–58
20. Ting K. M., Witten I. H. Issues in Stacked Generalization // J. Artif. Int.Res. — USA, 1999. — Vol. 10, no. 1. — P. 271–289.

21.. Rogova G. Combining the Results of Several Neural Network Classifiers //Neural Netw. — Oxford, UK, UK, 1994. — Vol. 7, no. 5. — P. 777–781.
22. On Combining Classifiers / J. Kittler [et al.] // IEEE Trans. Pattern Anal.Mach. Intell. — Washington, DC, USA, 1998. — Vol. 20, no. 3. — P. 226–239
23. Rokach L. Ensemble-based classifiers // Artificial Intelligence Review. —2010. — Vol. 33, no. 1. — P. 1–39.
24. Fiscus J. G. A post-processing system to yield reduced word error rates:Recognizer Output Voting Error Reduction (ROVER) // 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings. — 1997. — P. 347–354.

ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інженерії програмного забезпечення

Магістерська робота

«Розробка інформаційної системи для розпізнавання об'єктів з відеопотоку за допомогою технології комп'ютерного зору»

Виконав: студент групи ПДМ-61 Треньов Микита Георгійович

Керівник: доктор філософії, доцент кафедри Гребенюк В.В.

Київ - 2021

Слайд 1

МЕТА, ОБ'ЄКТ ТА ПРЕДМЕТ ДОСЛІДЖЕННЯ

Мета роботи: оптимізація процесу розпізнання об'єктів у відеопотоці за рахунок комбінування множини результатів спостережень

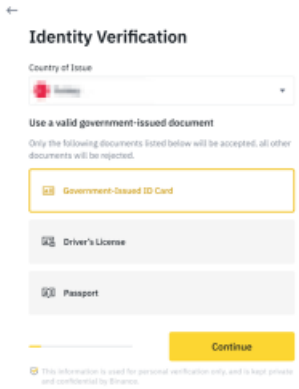
Об'єкт дослідження: мобільна система для розпізнавання документів, що засвідчують особу

Предмет дослідження: моделі та алгоритми розпізнавання та класифікації строкових даних

Слайд 2

Сфери застосування

Мобільні системи оптичного розпізнавання та автоматичного введеня даних можуть використовуватися у таких сферах:



- корпоративне та державне управління
- віддалена ідентифікація особистості
- підтвердження фінансових операцій та ін.

3

Слайд 3

Розпізнавання на мобільних пристроях

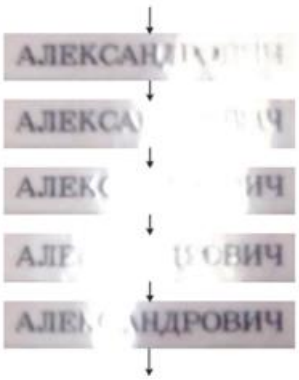
Проблеми мобільного розпізнавання:

- неконтрольовані умови зйомки
- недостатнє та/або нерівномірне освітлення
- змазаність
- фокусування
- відблиски
- низька роздільна здатність

4

Слайд 4

Розпізнавання даних у відеопотоці



Використання відеопотоку дозволяє збільшити точність результату за рахунок багаторазового розпізнавання різних зображень того самого об'єкта.

При розпізнаванні відеопотоку виникає 2 завдання:
- Як комбінувати покадрові результати
- Коли процес слід зупинити

5

Слайд 5

Існуючі підходи

Використання множини спостережень

- Метод супер-роздільної здатності

- Метод ансамблів групової класифікації

- Метод вибору найкращого кадру

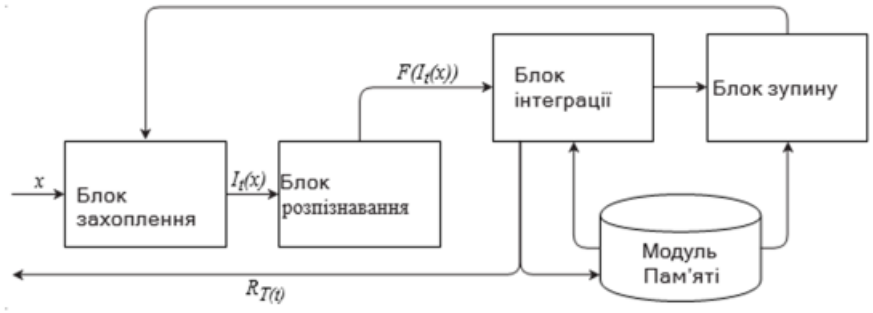
Комбінування результатів розпізнавання строкових даних

- ROVER (Recognizer Output Voting Error Reduction) - метод, що ґрунтується на голосуванні класифікаторів та застосовується в системах розпізнавання мовлення за допомогою комбінування кількох алгоритмів, а також для комбінування результатів розпізнавання текстових рядків

6

Слайд 6

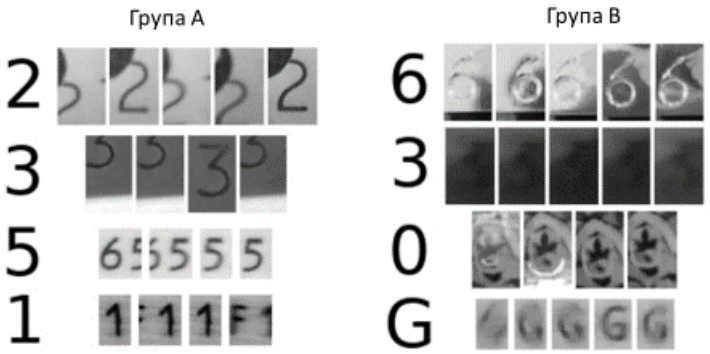
Модель системи розпізнавання у відеопотоці



7

Слайд 7

Стратегії комбінування покадрових результатів класифікації

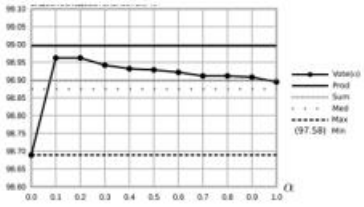


8

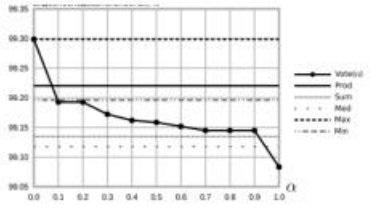
Слайд 8

Інтеграція одиночних символів

Група А



Група В



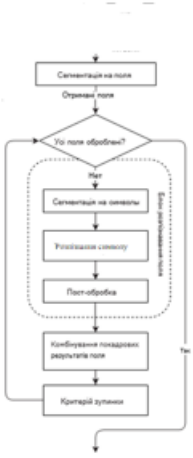
$$\mathrm{Prod}(X)(\sigma) = P(\sigma|X) = \frac{1}{P(\sigma)^{N-1}} \prod_{i=1}^{N} C_{\Sigma}(X_i)(\sigma); \quad (1)$$

$$\mathrm{Max}(X)(\sigma) = \left(\max_{i=1}^{N} C_{\Sigma}(x_i)(\sigma) \right) \cdot \left(\sum_{k=1}^{K} \max_{i=1}^{N} C_{\Sigma}(x_i)(\sigma_k) \right)^{-1}; \quad (2)$$

9

Слайд 9

Схема обробки кадру у системі розпізнавання документів у відеопотоці



10

Слайд 10

Приклад розпізнавання документу

11

Слайд 11

ВИСНОВКИ

1.Розроблено модель системи розпізнавання об'єкта у відеопотоку з блоком комбінування покадрових результатів розпізнавання та з блоком прийняття рішення про зупинення.

2.Дана модель дозволяє розглядати систему розпізнавання об'єкта у відеопотоку як ітераційний обчислювальний процес, який здатний видати в будь-який час найкраще на даний момент рішення, та припинити захоплення нових зображень згідно з заданим правилом зупинки.

3.Показано, що якщо в послідовності оброблюваних зображень одиночного зображення відсутні помилки попередньої обробки (такі, як помилки локалізації та сегментації символів), більш високу точність фінального результату забезпечує правило максимальної оцінки. Для відеопослідовностей, у яких зустрічаються помилки локалізації та сегментації символів, більш високу точність фінального результату забезпечують правила проведення оцінок, правило голосування та правило суми оцінок

12

Слайд 12

АПРОБАЦІЯ РОБОТИ

Статті:

1. Треньов М.Г. Дослідження властивостей двовимірних дискретних перетворень у комп'ютерному зорі // Зв'язок. №3 (139), 2021. Стаття подана до друку

Тези доповідей:

1. Треньов М.Г. Комп'ютерний зір та області його застосування. // Міжнародна конференція для країн Європи «Цифрова трансформація на основі інновацій у сфері ІКТ для розвитку цифрової економіки». – Київ: ДУТ, 2021.

Слайд 13

ДЯКУЮ ЗА УВАГУ!

Слайд 14