Київський національний університет імені Тараса Шевченка

**Лабораторна робота №1**

**З навчального курсу «Розпізнавання жестів з використанням нейромереж»**

**«** **gesture recognition»**

**“1506.01911v3.pdf”**

Виконав:

студент 4 курсу

факультету кібернетики

спеціальність «Комп’ютерні науки»

групи ТТП-42

Чебан Богдан Володимирович

**Київ 2025**

**Вступ**

Звіт аналізу статті «Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video». Спочатку подано версію українською мовою, а потім – англійською.

**UA. Звіт**

1. Постановка задачі та загальна мета роботи

Стаття спрямована на пошук нових шляхів для покращення розпізнавання жестів у відео. Основна ідея полягає у створенні моделі, яка б змогла не лише «бачити» окремі кадри, але й правильно розуміти послідовність рухів у часі. Для цього дослідники інтегрували два ключових компоненти: часові згортки (temporal convolutions) та двосторонню рекуренцію (bidirectional recurrence). Це дозволяє моделі враховувати як локальні зміни зображення, так і довгострокові залежності між кадрами. Метою є досягнення високої точності у визначенні моментів початку та завершення жесту, що є важливим у практичних застосуваннях, наприклад, при взаємодії людини з комп’ютером.

2. Огляд існуючих підходів до розпізнавання жестів

**-Одноразовий аналіз кадру (Single-Frame CNN):** Цей метод базується на аналізі окремих кадрів за допомогою згорткових мереж. Хоча такий підхід добре працює для загальної класифікації об’єктів, він не може врахувати динаміку руху.

**-Часове пулінгування (Temporal Feature Pooling):** Тут ознаки з послідовності кадрів агрегуються методом середнього або максимального пулінгу. Такий метод дозволяє інтегрувати інформацію з кількох кадрів, проте втрачається порядок подій, що часто є критичним для точного розпізнавання жестів.

**-Рекурентні мережі (RNN):** Використання RNN, особливо в двосторонній архітектурі з LSTM, дозволяє моделі обробляти інформацію з минулих та майбутніх кадрів, що є важливим для розуміння контексту руху. Це підхід дозволяє більш точно вловлювати моменти, коли жест починається або завершується.

**-Комбінований підхід (Temporal Convolutions + RNN):** Найцікавішим є поєднання часових згорток і рекурентних шарів. Часові згортки витягують локальні особливості руху, а рекурентна частина допомагає «запам'ятовувати» послідовності та моделювати їх еволюцію.

3. Технологія розпізнавання та ключові блоки системи

Модель складається з кількох основних компонентів:

**-CNN-шар:** Основним завданням є вилучення просторових ознак з кожного кадру. Завдяки цій частині мережі модель розпізнає основні об’єкти, форми та контури.

**-Часові згортки:** Тут виконується операція, яка схожа на стандартну згортку, але застосовується не тільки в просторі, а й у часовій площині. Завдяки факторизації операції у двовимірну (просторову) та одномірну (часову) згортки, досягається менша кількість параметрів та можливість вбудовувати більше нелінійності.

**- Двостороння рекуренція:** Цей блок дозволяє моделі аналізувати послідовність як вперед, так і назад. Такий підхід особливо важливий, коли потрібно точно визначити межі жесту, адже інколи ключові сигнали з’являються лише після певного затримання.

**- Softmax-класифікатор:** На завершальному етапі виходи з рекурентних шарів підсумовуються та проходять через класифікатор, який формує остаточний прогноз для кожного кадру.

4. Підготовка даних та їх обробка

У дослідженні використовується датасет Montalbano, який є одним із найбільших у галузі розпізнавання жестів. Особливості підготовки даних включають:

-**Кадрування за допомогою інформації про скелет:** Це дозволяє зосередитися на зображенні користувача, виключаючи зайві деталі заднього фону.

**-Зміна розміру зображень:** Після кадрування зображення змінюють розмір до 64×64 пікселів, що значно зменшує обчислювальні витрати при збереженні достатньої інформації для розпізнавання.

**-Аугментація даних:** До даних застосовуються випадкові трансформації – горизонтальні та вертикальні зсуви, обертання, масштабування та навіть зміни в часі (temporal scaling). Це сприяє покращенню узагальнюючої здатності моделі та дозволяє їй стійко працювати у реальних умовах.

5. Проведення експериментів та тренування моделі

Автори статті докладно описують процес тренування:

**-Енд-ту-енд навчання:** Модель тренується повністю «з нуля», що дозволяє оптимізувати всі шари одночасно. Важливим аспектом є використання зворотного розповсюдження помилки через час (BPTT) для рекурентних частин.

**-Оптимізація за допомогою Adam:** Використання сучасного алгоритму оптимізації Adam допомагає швидше знаходити оптимум серед високорозмірного простору параметрів.

**-Регуляризація та dropout:** Щоб уникнути перенавчання, на всіх повнозв’язаних шарах застосовується dropout із параметром 0.5, а також використовується активація leaky ReLU, що дозволяє уникнути проблеми з «зникненням» градієнтів.

**-Оцінка якості:** Для вимірювання успішності застосовували метрики, такі як Jaccard Index, точність (precision), відтворення (recall) та error rate. Найкращий результат (Jaccard Index ≈ 0.906) демонструє, що комбінований підхід забезпечує значне покращення порівняно з базовими моделями.

6. Порівняльний аналіз з іншими дослідженнями

Один із важливих аспектів статті – це порівняння отриманих результатів з роботами інших дослідників:

-Переваги моделі: Розроблена модель перевершує інші підходи, такі як метод Neverova et al. (2014), навіть якщо використовувати лише RGB-D дані.

-Висновок: Ці результати свідчать про те, що інтеграція часових згорток і двосторонньої рекуренції дозволяє моделі краще розуміти динаміку руху та коректно визначати межі жестів. Навіть у випадку, коли відсутня додаткова інформація (наприклад, дані про позу або глибину), модель демонструє високу ефективність.

7. Опис обладнання та тестового середовища

У дослідженні використовувалися сучасні графічні процесори (NVIDIA GPU), що значно пришвидшували процес тренування моделі. Вхідні дані отримувалися за допомогою пристрою Microsoft Kinect, який забезпечує не лише кольорове зображення (RGB), але й інформацію про глибину (D) та позицію користувача. Це дозволило створити багатоканальний підхід до розпізнавання жестів.

**Загальні враження та висновки**

Стаття дуже детально описує шлях від базових підходів до інноваційних рішень у галузі розпізнавання жестів. Розширене використання часових згорток разом із двосторонньою рекуренцією показало, що навіть для задач з відносно короткими послідовностями можна досягти високої точності у визначенні меж та класифікації жестів. Аналіз експериментальних результатів демонструє, що поєднання різних методів дозволяє не тільки покращити якість розпізнавання, але й зробити систему більш стійкою до шумів та варіацій у даних. Це дослідження може бути надзвичайно корисним як для академічних досліджень, так і для практичних застосувань у сфері інтерактивних технологій.

**ENG. Report**

1. Problem Statement and Overall Objective

The article focuses on finding new ways to improve gesture recognition in videos. Its core idea is to build a model that can not only analyze individual frames but also understand the sequence of movements over time. To achieve this, the researchers integrate two key components: temporal convolutions and bidirectional recurrence. This combination allows the model to capture both local motion features and long-term dependencies between frames. The ultimate goal is to accurately detect the beginning and end of gestures—a critical requirement for real-world human-computer interaction applications.

2. Overview of Existing Approaches to Gesture Recognition

The article reviews and compares several approaches:

**Single-Frame Analysis (Single-Frame CNN):** This method uses standard convolutional neural networks to analyze individual frames. Although effective for basic object recognition, it lacks the capability to capture motion dynamics.

**Temporal Feature Pooling:** Here, features from a sequence of frames are aggregated using mean or max pooling. This approach integrates temporal information but loses the sequential order, which is vital for precise gesture detection.

**Recurrent Neural Networks (RNNs):** By employing bidirectional RNNs—especially with LSTM cells—the model processes data in both forward and backward directions. This allows for a more nuanced understanding of temporal context, essential for determining when a gesture starts or ends.

**Combined Approach (Temporal Convolutions + RNN):** The most compelling strategy merges temporal convolutions to extract local motion features with recurrent layers to model long-term dependencies. This integrated approach significantly boosts performance.

3. Technology and Key Components of the Recognition System

The architecture consists of several integral components:

**CNN Layer:** This part is responsible for extracting high-level spatial features from each video frame. It identifies basic shapes, edges, and objects in the image.

**Temporal Convolutions:** Unlike standard convolutions, temporal convolutions operate over the time dimension. By factorizing the operation into 2D spatial and 1D temporal convolutions, the network efficiently extracts motion cues while keeping the parameter count manageable.

**Bidirectional Recurrence:** This component processes the video sequence in both directions, which is especially useful for accurately marking the start and finish of gestures.

**Softmax Classifier:** Finally, the outputs from the recurrent layers (both forward and backward) are summed and passed to a softmax layer, which produces the final class predictions for each frame.

4. Data Preparation and Preprocessing

The researchers used the Montalbano Gesture Recognition Dataset, one of the largest available datasets for gesture recognition. Key steps in data preparation include:

**Cropping Using Skeleton Information:** The images are cropped based on skeletal data to focus on the person performing the gesture, thus removing irrelevant background details.

**Image Rescaling:** After cropping, each frame is rescaled to a 64×64 resolution. This reduction in size helps in decreasing computational load while preserving essential details.

**Data Augmentation:** A series of random transformations such as translations, rotations, scaling, and even temporal scaling are applied. This step greatly enhances the model’s robustness and its ability to generalize to real-world conditions.

5. Experimental Setup and Model Training

The article provides an in-depth discussion of the training process:

**End-to-End Training:** The entire model is trained from scratch using backpropagation through time (BPTT), ensuring that all layers are optimized jointly.

**Adam Optimization:** The use of the Adam optimizer accelerates the convergence process by effectively navigating the high-dimensional parameter space.

**Regularization Techniques:** To avoid overfitting, dropout (with a probability of 0.5) is applied at the input of every fully connected layer, and leaky ReLU activations are used to mitigate issues like vanishing gradients.

**Evaluation Metrics:** The performance of the models is assessed using the Jaccard Index, precision, recall, and error rate. Notably, the combined model (temporal convolutions with LSTM) achieves an impressive Jaccard Index of approximately 0.906.

6. Comparative Analysis with Other Works

A significant portion of the article is dedicated to comparing the proposed model with previous approaches:

**Strengths of the Proposed Model:** The combined approach not only outperforms traditional methods like single-frame CNNs and simple temporal pooling but also surpasses sophisticated models such as those by Neverova et al. (2014), even when using only RGB-D data.

**Key Insight:** The results clearly indicate that integrating temporal convolutions with bidirectional recurrence allows for a more refined capture of motion dynamics, even when additional inputs like depth or pose data are omitted.

7. Description of Hardware and Test Environment

The experiments were conducted on systems equipped with NVIDIA GPUs, which significantly reduced training times. The video data was captured using a Microsoft Kinect, providing both RGB and depth information along with skeletal tracking. This multi-modal input is crucial for robust gesture recognition.

**Final Thoughts and Conclusions**

This article provides a comprehensive journey from traditional approaches to an innovative, integrated method for gesture recognition. The extended use of temporal convolutions combined with bidirectional recurrence shows that even in tasks involving relatively short sequences, it is possible to achieve remarkable accuracy in detecting gesture boundaries and classifying actions. The experimental results underscore that a thoughtful fusion of different techniques not only boosts performance but also enhances the robustness of the system in noisy, real-world scenarios. Such advancements are valuable both for academic research and for practical applications in interactive technology.