Київський національний університет імені Тараса Шевченка

**Лабораторна робота №1**

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

**проф. Крак Ю.В.**

Виконав:

студент 4 курсу

факультету кібернетики

спеціальність «Комп’ютерні науки»

групи ТТП-42

Ходаков Максим

**Київ 2025**

**"Stochastic Fine-grained Labeling of Multi-state Sign Glosses for Continuous Sign Language Recognition"**.

**1. Постановка задачі / Мета роботи**  
У статті розглядається проблема розпізнавання безперервної жестової мови (Continuous Sign Language Recognition, CSLR). Основною метою роботи є розробка нової архітектури, яка дозволяє більш детально розбивати кожен жест (gloss) на підкомпоненти, що відповідають окремим фазам руху. Для цього автори пропонують концепцію стохастичного fine-grained маркування (Stochastic Fine-grained Labeling, SFL), де кожен жест моделюється як послідовність декількох станів. При цьому кількість станів не є фіксованою, а визначається як випадкова змінна за допомогою навчальної ймовірнісної моделі. Таке моделювання дозволяє отримати додатковий нагляд у часовій області під час тренування, що сприяє більш точному вирівнюванню послідовності вхідних кадрів із відповідними жестовими описами. Метою роботи є покращення якості розпізнавання жестової мови за рахунок гнучкішої і адаптивнішої розбивки на підстани, що в свою чергу допомагає знизити показники помилок розпізнавання.

**2. Існуючі підходи до розпізнавання жестів (огляд літератури)**  
У розділі огляду літератури автори розглядають історію та еволюцію підходів до розпізнавання жестової мови:

* **Класичні методи:**  
  Ранні роботи зосереджувались на застосуванні традиційних алгоритмів, таких як Gaussian Mixture Models (GMM) у поєднанні з Hidden Markov Models (HMM). Ці підходи зазвичай використовували ручну розробку ознак, що суттєво обмежувало їхню здатність узагальнювати дані та робило їх не надто стійкими до змін у відео.
* **Гібридні підходи:**  
  Після впровадження методів глибокого навчання з’явилася можливість поєднувати автоматичне вилучення ознак за допомогою конволюційних нейронних мереж (CNN) з HMM. Такі методи дозволили оптимізувати вилучення ознак спільно з процесом вирівнювання, але часто вимагали складної етапної процедури перенавчання, оскільки параметри CNN не завжди навчались ефективно в кінцевій системі.
* **CTC-підходи:**  
  Ще одним важливим кроком стала інтеграція Connectionist Temporal Classification (CTC), що дозволяє здійснювати «м’яке» вирівнювання послідовностей без необхідності проводити явну попередню обробку (реалізацію алгоритму Viterbi для HMM). Це значно спрощує тренувальний процес, адже вирівнювання виконується автоматично після кожного міні-батчу, але водночас виникає проблема недостатнього навчання нижнього рівня вилучення ознак.
* **Seq2Seq та трансформери:**  
  Також були запропоновані підходи на основі архітектури Sequence-to-Sequence (Seq2Seq), де вирівнювання виконується через глобальне зважене сумування вхідної послідовності. Проте такі методи часто страждали від експозиційного зміщення (exposure bias). Сучасні дослідники почали використовувати трансформери, що завдяки своїм механізмам уваги дозволяють ефективно вирішувати задачу розпізнавання мовних послідовностей.

Таким чином, попередні роботи охоплювали різні підходи, проте залишалися певні недоліки, зокрема щодо точності вирівнювання та адаптивності вилучення ознак. Саме тому в даній статті пропонується інтегрувати стохастичне fine-grained маркування, яке дозволяє автоматично визначати оптимальну кількість станів для кожного жесту.

**3. Опис схеми технології розпізнавання**  
Запропонована архітектура побудована за класичною схемою «візуальна модель – контекстуальна модель – модель вирівнювання». Детальний опис компонентів:

* **Візуальна модель:**  
  Для вилучення ознак з окремих кадрів відео використовується 2D-конволюційна нейронна мережа, зокрема ResNet18, попередньо натренована на ImageNet. Цей підхід дозволяє ефективно витягувати просторові ознаки з зображень, що є важливим для подальшого аналізу.
* **Контекстуальна модель:**  
  Після вилучення ознак з кадрів, інформація передається у трансформер-енкодер. Застосування відносного позиційного кодування дозволяє моделювати часову залежність між послідовними кадрами, що є критично важливим для розпізнавання жестів, де послідовність рухів має вирішальне значення.
* **Модель вирівнювання (Alignment Model):**  
  Для вирівнювання послідовності ознак з послідовністю жестових описів використовується алгоритм CTC (Connectionist Temporal Classification). CTC дозволяє проводити вирівнювання без необхідності попереднього розбиття відео на окремі жести, оскільки алгоритм обчислює «м’яке» вирівнювання через послідовність прихованих станів, включаючи спеціальний символ «blank».
* **Стохастичне fine-grained маркування (SFL):**  
  Це ключовий внесок статті. Метод дозволяє для кожного жесту визначати змінну кількість підстанів (sub-gloss states). Функція розширення (extension function) приймає стандартний послідовний опис жесту та заданий вектор кількості станів, після чого генерує розширену послідовність підстанів. Алгоритм REINFORCE використовується для навчання ймовірнісного розподілу кількості станів, що дозволяє підсилити ті варіанти, які призводять до зниження CTC-помилки.
* **Додаткові стохастичні методи:**
  + **Stochastic Frame Dropping (SFD):** під час тренування випадковим чином відкидаються певні кадри відео. Це дозволяє не лише зменшити обчислювальне навантаження і споживання пам’яті, але й забезпечує додаткову аугментацію даних, сприяючи адаптації моделі до різних швидкостей та умов зйомки.
  + **Stochastic Gradient Stopping (SGS):** частина кадрів обробляється без зворотного розповсюдження градієнта. Такий підхід зменшує ризик перенавчання нижніх шарів мережі і допомагає знизити обчислювальні витрати під час тренування.

**4. Обробка даних та підготовка для розпізнавання**  
Розглядаються питання, пов’язані з підготовкою даних та їхньою аугментацією для ефективного тренування моделі:

* **Вибір датасетів:**  
  Для експериментальної перевірки використовуються два відомих датасети – PHOENIX-2014 та PHOENIX-2014-T. Обидва датасети містять записи німецької жестової мови, отримані з телепередач (наприклад, прогноз погоди). Ці датасети містять тисячі кадрів, що дозволяє перевірити ефективність моделі у реальних умовах.
* **Попередня обробка відео:**  
  Кадри з відео спочатку масштабовано до розміру 256×256, після чого проводиться обрізання до розміру 224×224. Під час тренування застосовується випадкове обрізання, що додає додаткову аугментацію, а для тестування – центроване обрізання, щоб забезпечити стабільність оцінки.
* **Аугментація даних за допомогою стохастичних методів:**  
  Метод SFD (випадкове відкидання кадрів) дозволяє моделі зустрічати різноманітні варіації у послідовностях відео, змінюючи таким чином ритм жестів. SGS, у свою чергу, забезпечує зниження обчислювального навантаження, оскільки частина кадрів не враховується під час зворотного розповсюдження градієнта. Це сприяє покращенню узагальнювальних властивостей моделі.

**5. Проведення експериментів**  
Автори детально описують експериментальне налаштування та результати:

* **Налаштування тренування:**  
  Модель тренується протягом 30 епох із batch size = 8. Для оптимізації використовують алгоритм Adam із стандартними параметрами (β₁ = 0.9, β₂ = 0.999), а швидкість навчання регулюється за допомогою схеми, де початкова швидкість η₀ = 1×10⁻⁴ з періодичним зменшенням.  
  Крім того, проводиться детальний аналіз впливу гіперпараметрів:
  + **pdrop (ступінь випадкового відкидання кадрів):** експерименти показали, що оптимальне значення (близько 50%) дозволяє моделі уникнути перенавчання, зберігаючи при цьому достатню кількість інформації для розпізнавання.
  + **pstop (відсоток кадрів, для яких зупиняється обчислення градієнта):** вибір оптимального значення допомагає збалансувати навчання візуальної моделі та знизити обчислювальні витрати.
* **Процес навчання SFL:**  
  При використанні стохастичного fine-grained маркування кількість підстанів для кожного жесту визначається динамічно за допомогою алгоритму REINFORCE. В процесі тренування для кожного прикладу відбирається декілька зразків кількості станів (Monte Carlo sampling), після чого обчислюється винагорода (reward) на основі зниження CTC-помилки. Для зменшення дисперсії оцінки градієнта вводиться базова функція (baseline).
* **Метрика оцінки:**  
  Як основна метрика використовується Word Error Rate (WER), що обчислюється як сума операцій заміни, вставки та видалення, необхідних для перетворення розпізнаного речення у правильне. Це дозволяє порівнювати результати з іншими роботами в області CSLR.
* **Результати експериментів:**  
  Аналіз експериментів показав, що застосування SFD та SGS сприяє зниженню перенавчання, а інтеграція SFL значно покращує результати порівняно з моделями, де кількість станів для жестів залишається фіксованою. Остаточна модель із застосуванням мовної моделі (LM) демонструє покращення WER на dev та test наборах (зменшення помилок на 1.1% та 0.7% відповідно) порівняно з іншими state-of-the-art методами.

**6. Співставлення результатів з іншими роботами**  
У статті проводиться порівняльний аналіз результатів запропонованого підходу з роботами, які використовують інші сучасні методи розпізнавання:

* **Порівняння на датасеті PHOENIX-2014:**  
  Запропонована модель демонструє кращі показники за метрикою WER порівняно з попередніми роботами, використовуючи як базову модель класичний підхід з CTC, так і різні варіанти гібридних архітектур. Інтеграція мовної моделі (LM) дозволяє досягти ще більшого покращення, зменшуючи помилки розпізнавання.
* **Порівняння на датасеті PHOENIX-2014-T:**  
  Для цього датасету, який є розширенням PHOENIX-2014 з дещо іншими межами речень, модель показує конкурентні результати навіть при використанні лише жестових анотацій (gloss annotations). Це свідчить про універсальність запропонованого підходу та його здатність адаптуватись до різних умов розпізнавання.

Таким чином, результати експериментів підтверджують ефективність використання стохастичного fine-grained маркування у поєднанні з SFD та SGS, що дозволяє моделі перевершувати або бути на рівні із сучасними state-of-the-art системами.

**7. Опис обладнання / обчислювальної платформи**  
Хоча стаття не містить докладного опису апаратного забезпечення, можна зробити кілька висновків на основі наданої інформації:

* **Використання попередньо натренованої моделі:**  
  Для вилучення ознак використовується ResNet18, натренована на ImageNet, що свідчить про застосування сучасних GPU або інших потужних обчислювальних ресурсів для попереднього тренування.
* **Ефективність розрахункових процесів:**  
  Застосування методів SFD та SGS дозволяє знизити обчислювальні витрати під час тренування, оскільки менша кількість кадрів обробляється та частина градієнтів не враховується. Це свідчить про оптимізацію процесу тренування для роботи на сучасних обчислювальних платформах.
* **Можливе використання GPU:**  
  Хоча конкретні моделі пристроїв не зазначені, типовий експеримент з використанням подібних архітектур проводиться на сучасних графічних процесорах (наприклад, NVIDIA Tesla або GeForce), що дозволяє обробляти великі обсяги відеоданих в режимі реального часу.

**8. Основні висновки та внесок роботи**  
Стаття має кілька ключових внесків у галузь розпізнавання безперервної жестової мови:

* **Новизна методології:**  
  Запропоновано новий підхід для детального розбивання жестів, який дозволяє автоматично визначати кількість підстанів для кожного жесту за допомогою стохастичного fine-grained маркування (SFL). Це відрізняє роботу від класичних підходів, де кількість станів фіксована.
* **Покращення якості розпізнавання:**  
  Завдяки інтеграції методів SFD та SGS вдалося значно знизити ризик перенавчання, покращити узагальнення моделі та зменшити обчислювальні витрати. Результати експериментів свідчать про зниження показника помилок (WER) порівняно з традиційними методами.
* **Сучасний підхід до вирівнювання послідовностей:**  
  Використання алгоритму CTC у поєднанні з адаптивним розширенням жестових анотацій дозволяє більш точно вирівнювати послідовності кадрів із відповідними жестами. Це сприяє покращенню навчального процесу і підвищенню стійкості системи до варіацій у відео.
* **Порівняльний аналіз:**  
  Результати роботи демонструють конкурентоспроможність запропонованого підходу як на датасеті PHOENIX-2014, так і на його розширеній версії PHOENIX-2014-T. Досягнуті показники перевершують або наближаються до state-of-the-art рішень у галузі розпізнавання жестової мови.
* **Підсумок:**  
  Загалом, стаття робить вагомий внесок у розробку ефективних систем розпізнавання безперервної жестової мови, пропонуючи інтегрований підхід, який поєднує сучасні методи вилучення ознак, контекстуальне моделювання за допомогою трансформерів та інноваційне стохастичне маркування жестів. Це відкриває нові перспективи для подальших досліджень у сфері комп’ютерного зору та обробки природної мови.