Себало М.М.
студент,
Науковий керівник: **Провотар О.І.**
доктор фізико-математичних наук, професор,
Київський національний університет імені Тараса Шевченка

ЗГЛАДЖЕННЯ РЕЗУЛЬТАТІВ РОЗПІЗНАВАННЯ ПОЗИ НА ВІДЕО

Зазвичай, задачі розпізнавання пози людини дається наступне визначення: пошук розташування частин тіла людини (або пози) на зображенні [1]. Вирішується задача розпізнавання пози людини на відео. Її ціллю є знайти позу у кожному кадрі відео при допущенні, що вона змінюється плавно між кадрами вздовж тимчасових осей. Тож, ми розраховуємо знайти позу у всій послідовності кадрів. Дається наступне визначення позі: розташування рамок, що обмежують людське тіло та її кінцівки. Тобто, щоб знайти позу, необхідно розмістити рамки навколо людини та частин її тіла у кожному кадрі відео.

Загалом, даний алгоритм розпізнавання пози людини на відео можна описати наступним чином. На першому кроці обираються рамки-кандидати з окремих кадрів, які можуть містити один із елементів пози – всю людину або її кінцівку. Після цього здійснюється виявлення ознак у рамках-кандидатах, які дозволять класифікувати зображення у рамці. Далі до виявлених ознак застосовується класифікатор, що показує ймовірність знаходження частини тіла у рамці-кандидаті. Обираються рамки з найкращим результатом. До них застосовується згладжувальний алгоритм, що дає фінальний результат – послідовність поз у кадрах, згладжених за тимчасовими осями.

Розглянемо детальніше останній етап описаного алгоритму. До цього були пройдені наступні кроки: обрані рамки-кандидати за допомогою селективного пошуку [2], обчислені ознаки для цих рамок за допомогою згорткових нейронних мереж [3; 4], присвоєно оцінки за допомогою класифікатора методом опорних векторів [5]. Якби ми розпізнавали позу на простих зображеннях без відео, класи були б незалежними та зображення б не мали тимчасових зв’язків. У такому випадку наступним кроком було б застосування пошуку локальних максимумів для рамок-кандидатів для кожного окремого класу на зображенні. Це і було б фінальним результатом.

Натомість у нашому випадку можна використати додаткову інформацію з відео ряду. Тому останній етап розпізнавання складається з двох кроків. На першому кроці на кожному кадрі обираються N рамок-кандидатів з класом «людина», що мають найвищі оцінки. Після цього положення кожної рамки згладжується вздовж тимчасових осей. В результаті ми отримуємо згладжений трек відслідковування людини на відео. Другим кроком є знаходження треків рук. Для здійснення тимчасового згладжування використовується інформація про оптичний потік між сусідніми кадрами.

Оптичний потік – це представлення видимого сліду руху об'єктів, поверхонь, і граней візуальної сцени, що спостерігається під час відносного руху між спостерігачем (наприклад, око людини або камера) і сцени. Оптичний

потік використовується для отримання швидкості руху об'єктів на відео, тому є дуже корисним у задачах розпізнавання руху чи діяльності. В нашому випадку оцінюється рух кінцівок та людини між сусідніми кадрами.

Ми використовуємо алгоритм DeepFlow у реалізації Вайнзапфела та інших [6]. Алгоритм використовує мультишарову архітектуру подібну до згорткової нейронної мережі, що дозволяє відслідковувати переміщення великих областей і ставити їх у відповідність на кадрах.

Після отримання оптичного потоку, можна розпочинати процедуру згладження. Завдання знаходження найкращого гладкого треку у послідовності кадрів може бути вирішено методами динамічного програмування [7]. Далі буде показано, що дане завдання можна розділити на послідовність підзавдань, що перетинаються, а це і є необхідною вимогою для застосування методів динамічного програмування.

Дано послідовність з F кадрів і необхідно знайти найкращий трек для певного класу. На кожному кадрі $t, t = 1 \dots F$, обираємо N_t рамок-кандидатів, найкращі за оцінкою знайденою методом опорних векторів. У роботі використано $N_t = 10, \forall t$. Якщо відомо, що в кожному кадрі присутній лише один об'єкт класу, то завдання знаходження треку зводиться до вибору рамок-кандидатів таким чином, щоб отримати найбільшу оцінку для треку в цілому. Потрібно, щоб рамки у сусідніх кадрах були близько одне до одного (з поправкою на відносний рух між кадрами) та були схожі. Ці дві характеристики представлені «попарною оцінкою», що містить компоненти близькості та схожості. Також використовується «унарна оцінка», що відповідає оцінці, отриманій методом опорних векторів.

Нехай дано набір значень $L_j^{F-1}, j = 1 \dots N_{F-1}$, що відповідає найкращій можливій оцінці, яка може бути отримана для послідовності з $F-1$ кадрів, при умові, що для кадру $F-1$ було обрано рамку під номером j. При такому визначенні вся оцінка послідовності кадрів залежить лише від вибору рамки в останньому кадрі. Подібним чином можна рекурсивно пройти до першого кадру, де оцінка буде залежати лише від унарної оцінки обраної рамки.

Дана проблема і її вирішення подібні до задачі пошуку найкоротшого шляху в ациклічному графі (але в нашому випадку ведеться пошук найдовшого шляху, тому що іде максимізація оцінки).

Описана процедура застосовується без змін для знаходження треку для рамок, що відповідають класу «людина». В цьому випадку лише один екземпляр класу присутній в кадрі.

Тепер необхідно застосувати алгоритм пошуку треку для класу «рука». У цьому випадку ми маємо два екземпляри класу у кожному кадрі. Тому процедура динамічного програмування подібна до випадку з людиною, проте має декілька важливих особливостей. По-перше, обробка рук відбувається після того, як була здійснена обробка людини для всієї послідовності. Тому всі рамки-кандидати для руки, що мають перетин з рамкою людини менший, ніж θ, відкидаються. По-друге, як вже зазначалось, необхідно знайти трек двох рук вздовж всієї послідовності.

Ми використовуємо унарну та попарну оцінку, що була описана для людини. Також ми додаємо внутрішньокадрову оцінку, що відповідає відношенню перетину до об'єднання між рамками лівої та правої руки.

Дані зміни пояснюються наступним чином: не можна просто обрати два найкращих треки, що знайдені методами динамічного програмування, тому що вони можуть належати одній руці. З іншого боку, не можна ввести жорстке обмеження, тому що цілком можливо, що в деяких позах рамки навколо рук можуть перетинатись або повністю накладатись одне на одного.

Також використовується наступна особливість пошуку. Селективний пошук застосовується до кожного кадру. В багатьох випадках рамка, що присутня в одному кадрі, не буде знайдена у наступному. Це ускладнює обробку послідовності кадрів, оскільки іноді пошук не знаходить прийнятних рамок-кандидатів і трек відхиляється в якомусь напрямку або повністю втрачає кінцівку. В той же час відео у нашому наборі даних не містять різких рухів. Тому можна вирішити проблему з відсутністю прийнятних рамок-кандидатів. Ми включаємо рамку з найвищою оцінкою L^{t-1} з попереднього кадру при пошуці L^t.

Таким чином, було представлено алгоритм розпізнавання пози людини на відео та детально розглянуто його останній етап, в якому інформація з відео використовується для покращення покадрового розпізнавання пози людини.

Список використаних джерел:

1. Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 104(2-3): 90–126, November 2006. URL: http://linkinghub.elsevier.com/retrieve/pii /S1077314206001263.

2. Uijlings J. R. R., van de Sande K. E. A., T. Gevers and Smeulders A. W. M. Selective search for object recognition. International Journal of Computer Vision, 104(2):154–171, 2013.

3. Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR, abs/1311.2524, 2013.

4. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc., 2012.

5. Corinna Cortes and Vladimir Vapnik. Support-vector networks. Machine Learning, 20(3):273–297, 1995.

6. Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. DeepFlow: Large displacement optical flow with deep matching. In IEEE Intenational Conference on Computer Vision (ICCV), Sydney, Australia, December 2013.

7. Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. Introduction to Algorithms. McGraw-Hill Higher Education, 2nd edition, 2001.