УДК 004.02

Методи розпізнавання рухів, дій людей на відео послідовностях

Солдатов Д. В.
e-mail kingit@bk.ru
КПІ ім. Ігоря Сікорського
Київ, Україна

***Анотація*—В статті розглянуто постановку проблеми розпізнавання рухів об'єктів на відеопослідовностях, етапи її вирішення, проведено аналіз основних методів кожного з етапів. Розглянуто ключові складнощі, що виникають при вирішенні задачі. Наведено способи порівняння різних методів. Проаналізовано існуючі підходи до розпізнавання рухів на відеопослідовностях, виявлено особливості, сильні та слабкі сторони та обмеження різних методів виявлення ознак та їх класифікації. Обрано методи для подальшого дослідження та вдосконалення.**

Ключові слова — розпізнавання рухів; відеопослідовність; оптичний потік; SVM; CNN.

I. Вступ

Вхідними даними задачі розпізнавання рухів є відео з людьми, що рухаються, вихідними – промарковані області де відбувається рух та пояснення, який саме рух чи дія (ходьба, біг, стрибки, різноманітні жести). Широке коло застосування, наприклад у сфері розваг – керування комп'ютером за допомогою жестів, у сфері безпеки – автоматичне відслідковування сутичок між людьми, в медичній сфері – аналіз порушень в опорно-руховому апараті, та зростаючі вимоги до якості розпізнавання зумовлює актуальність дослідження. Мета цієї статті розглянути та проаналізувати існуючи методи та підходи розпізнавання рухів та обрати перспективний метод для подальшого дослідження. Задача розпізнавання рухів об'єктів на відеопослідовностях є комплексною. Її можна умовно розділити на дві частини:

- представлення вхідної відео послідовності у вигляді особливих ознак;
- класифікація виявлених ознак на відповідні класи дій.

Зі зростанням популярності та кількості різних алгоритмів розпізнавання дій, важливо виявляти сильні та слабкі сторони кожного з підходів та мати засоби для порівняння. Одним з таких засобів є кількісна оцінка кожного підходу з одним набором тестових даних за відповідним, одним і тим же протоколом.

II. Набори тестових даних

Бази тестових даних відрізняються:

- характеристиками самого відео – тривалість, частота кадрів, якість, розширення;
- змістом – кількість учасників, дій, умови;
- протоколами тестування та призначенням (виявлення та / або розпізнавання).

Таким чином, важливо проаналізувати набори даних та їх ключові особливості, щоб розуміти можливості та обмеження кожного випробуваного підходу. Набір тестових даних зазвичай розділяєтеся на три непересічні набори (навчання, перевірка, тестування), що дозволяє дослідникам налаштувати свою систему і оцінити помилку одночасно. [1]

Одними з перших стандартних наборів тестових даних є KTH [2] і Weizmann[3]. Обидва набори містять відео зняті на статичну камеру зі статичним простим фоном. KTH база являє собою набір з 600 відео на яких 25 акторів виконують 6 дій(ходьба, біг, біг підтюпцем, боксування, хвиля, оплески) в 4 різних умовах (усього 600 відео зразків). База Weizmann містить 90 відео зразків, 10 акторів, 9 дій. Однак через прості умови, що рідко зустрічаються в реальному житті ці бази цікаві лише для знайомства з розпізнаванням, висока точність на цих базах не є чимось значимим.

Щоб задовольнити відсутність природних параметрів у наборах даних KTH та Weizmann, зокрема, чистота фону, наступним кроком було тестування алгоритмів на відео з динамічним фоном. Бази тестових даних CMU(Carnegie Mellon University) Crowded Videos dataset [4] та MSR(Microsoft Research Group) Action Dataset I, II. [5] Динамічний фон був отриманий шляхом запису відео в середовищі з рухомими автомобілями і людьми.

Наступні набори UC Berkeley Sports Dataset, UCF Sports dataset, Olympic Dataset, Sports-1M, являють собою відео з різних спортивних змагань, умови розширені рухомою камерою та різними кутами зйомки.

У спробах створити набір даних, що відповідає вимогам додатків у реальному світі для розпізнавання дій, необхідно було зібрати відео, що не обмежується рухами камери, контекстом сцени, просторовою сегментацією та точками зору. Початок набору даних

для відео розпочався з колекції осіб, які «п'ють» у фільмах «Кава і сигарети», а також «Море любові» [6]. Аналогічно, відео зі восьми різних фільмів було задіяно для збору 92 зразків «поцілунків» і 112 зразків «удару» [7]. Два найпоширеніших набору даних з фільмів – Hollywood1 [8] і Hollywood2 [9].

Набір даних UCF101[10] був одним з найбільш складних і великих наборів даних для виявлення та розпізнавання дій. Велика кількість відео і дій, розділених на п’ять категорій: взаємодія людини з об'єктом, рухи тіла, взаємодія людини з людиною, гра на музичних інструментах і спорт. Пізніше набір даних ActivityNet[11] взяв на себе цю роль і став одним з найскладніших зі своїм масштабом та різноманітністю умов. Набори даних UCF101 і ActivityNet містять відео, які дуже нагадують відео, які можна знайти в реальному світі. Таким чином, алгоритми, які добре працюють на цих наборах даних, мають великий потенціал для використання в реальних умовах.

III. Представлення зображення: особливі ознаки

Для того, щоб класифікувати дію ефективним і точним способом, особливі ознаки, які надають змістовну інформацію, повинні бути виявленні і передані для класифікації. В ідеалі, модель представлення повинна бути стійкою до зміни зовнішнього вигляду актора(ів), фону, точки зору і якості відео, зберігаючи достатню інформацію для точної класифікації дії. Для подолання цього бар'єру було запроваджено велику кількість моделей представлення (дескрипторів ознак).

Дескриптори ознак поділяються на два типи[12]:

- загальні примітивні ознаки – ознаки які можуть бути отримані безпосередньо з необроблених вхідних відео, які потім можуть бути використані безпосередньо в модулі класифікації;
- спеціальні примітивні ознаки – ознаки які виявляються з необроблених вхідних відеозаписів і вимагають додаткової обробки в допоміжні ознаки, перш ніж вони перейдуть на етап класифікації;

A. Дескриптори загальних примітивних ознак

Дескриптори на основі фільтрів. Підходи на основі фільтрів можна розділити на два типи: дескриптори на основі градієнта та смугових фільтрів. Методи, засновані на градієнтах, покладаються на припущення, що локальний вигляд і форми об'єкта можуть бути представленні їх локальними градієнтами або напрямками країв. Підходи на основі смугового фільтра використовують орієнтовані фільтри для розкладання відео на основні компоненти з використанням локальної орієнтації та масштабу.

Найбільш поширені дескриптори на основі градієнта: гістограма орієнтованих градієнтів (HOG) [13], HOG3D [14], кубоїдний дескриптор [15], масштабно-інваріантне перетворення ознак [SIFT] [16], градієнтна гістограма розташування орієнтації (GLOH) [17], локальні тринальні структури (LTP) [18] і просторово-часові (ST) фрагменты [19].

HOGs визначають просторово орієнтований градієнт для виявлення інформації про вигляд дії. HOG3D розширює дескриптори HOG, зберігаючи просторово-часові орієнтовані градієнти для визначення інформації про форму і рух разом. Кубоїдний дескриптор об'єднує три градієнтні канали (*Gx, Gy, Gt*) в один вектор, щоб сформувати єдиний вектор ознак для кожної околиці. SIFT, який поєднується з детектором масштабно інваріантної області, DoG, використовує тривимірні гістограми для представлення градієнтних розташувань та орієнтацій.

Хоча багато з цих дескрипторів на основі орієнтованого градієнта забезпечують обчислювальну ефективність для збору важливої інформації, такої як зовнішній вигляд і / або рух, вони дуже чутливі до змін освітлення. Часто ці дескриптори не дають достатньої інформації і повинні використовуватися паралельно з іншими дескрипторами, які мають відмінні ознаки (наприклад, HOG часто зустрічаються з HOF) для подолання цих обмежень.[12]

Просторово-часові орієнтовані смугові фільтри можуть розкладати послідовність зображень на основні компоненти, використовуючи розмірність локальної орієнтації та масштабу (тобто кутові та радіальні частоти). Різні типи орієнтованих фільтрів були застосовані до ряду завдань для розуміння динамічного зображення, таких як розпізнавання і виявлення дій [20]. Ці моделі представлення, як правило, здатні характеризувати динаміку зображення без явної потреби відновлення потоку або сегментації відео [21]. Загальноприйнятими є два особливих підходи просторово-часової орієнтованої фільтрації застосованих для розпізнавання дій: 3D фільтри Габора [20, 21] і гауссові похідні фільтри [67]. Обидва похідні фільтри 3D Gabor і Gaussian зазвичай застосовуються в квадратурних парах і комбінуються для отримання деяких локальних вимірювань енергії. Часто бере участь подальша обробка, така як нормалізація та / або комбінація виходів фільтрів. Процес нормалізації забезпечує стійкість до фотометричних варіацій [21], а комбінування виходів фільтрів допомагає зібрати інформацію про динаміку зображення, яка інваріантна до просторового вигляду. Виходи фільтрів також можуть бути об'єднані для отримання явних оцінок руху [22].

Представлення, що базуються на орієнтованих просторово-часових смугових фільтрах, є стійкими до змін освітленості, варіацій у класі та оклюзії (перетинання). Однак деякі реакції фільтрів (наприклад, смугові фільтри) створюють чутливість до нерелевантних атрибутів вигляду. Крім того, ці фільтри, як правило, є чутливими до змін масштабу, що є проблематичним, оскільки розмір актора / дії не узгоджується між і в межах кожного відео.

Дескриптори на основі оптичного потоку. Алгоритми *на основі* оптичного потоку часто з'являються в різних алгоритмах розпізнавання дій. Оптичний потік забезпечує даними, які можна використовувати двома способами: визначати інформацію про рух і для цілей відстеження.

Оптичний потік можна використовувати для розпізнавання дій, описуючи рух актора. Стандартний алгоритм оптичного потоку може бути застосований для захоплення руху, створеного різними частинами

тіла [23]. Розділяючи оптичний потік на горизонтальні та вертикальні компоненти а потім розмиваючи їх за допомогою Гаусіана, створюється штучний набір каналів руху [23, 24].

Гістограми оптичного потоку (HOF) фіксують локальний рух структури шляхом квантування орієнтації оптичних векторів потоку. Хоча така характеристика руху є достатньою для розрізнення дуже чітких дій (наприклад, "ходьба" проти "хвилі" в наборі даних KTH), вона не розрізняє точні відмінності в діях (наприклад, "бокс" проти "оплесків" у наборі даних KTH). Таким чином, простий опис руху в поєднанні з інформацією про зовнішній вигляд (наприклад, HOG) може дати більш точні результати розпізнавання.

Гістограма граничного руху (MBH) є дескриптором, який використовує похідні оптичного потоку для кожного горизонтального і вертикального напрямків, відповідно [25, 26]. Обчислюючи просторові похідні для кожного поля потоку, можна знайти локальні орієнтації і величини градієнта для побудови гістограми локальної орієнтації. Оскільки MBH обчислює градієнт оптичного потоку, пригнічується постійний рух і зберігається тільки інформація щодо змін у полі потоку. Таким чином, MBH забезпечує простий спосіб придушення постійного руху (наприклад, руху камери) при збереженні місцевого відносного руху пікселів (наприклад, меж руху / руху переднього плану Це приваблива особливість, особливо для розпізнавання дій в реалістичних відео, оскільки вони, як правило, містять серйозний рух камери [26]. Більше того, більшість текстурної інформації з статичного фону усувається при розгляді похідних траєкторій.

Дескриптори ознак щільної траєкторії (DT) [26] були введені як ще одна форма дескрипторів, що відстежують шлях руху, які часто з'являлися в області розпізнавання та виявлення дії [26]. Ознаки DT в першу чергу вимагають щільної вибірки особливих точок на кожному кадрі. Потім відстежується кожна з точок вибірки з використанням оптичного потоку для отримання траєкторії. Дескриптор траєкторії отримують шляхом зчеплення нормованих векторів зміщення. Ці ознаки часто поєднуються з іншими ознаками (наприклад, HOG, HOF, MBH), агрегованими вздовж траєкторій. Були запропоновані різні моделі щільної траєкторії, які б підвищили вихідну модель DT [27]. Одним з підходів було кластеризувати щільні траєкторії для виявлення домінуючого напрямку руху і розглянути відносний руху між траєкторіями для збору інформації про об'єкт-фон і об'єкт-об'єкт. Інший підхід полягав у тому, щоб явно оцінити рух камери [27] шляхом зіставлення точок характеристик між кадрами за допомогою дескрипторів SURF [28] і щільного оптичного потоку [29]. Ця особлива характеристика траєкторії руху з компенсацією руху камери згадується як поліпшена функція щільної траєкторії (iDT) [30].

Дескриптори, засновані на згортковій нейромережі. В останні роки з'явилися велика кількість алгоритмів, що спираються на згорткові нейроні мережі (CNNs) у широкому спектрі проблем на основі штучного інтелекту, включаючи розпізнавання дій. Як випливає з назви, CNNs засновані на нейронних мережах, які є системою, що складається з послідовності шарів з набором штучних «нейронів» у кожному шарі. Перший шар мережі, вхідний шар, зазвичай складається з необроблених пікселів зображення / відео [30], але попередньо оброблені дані, такі як оптичний потік зміщення поля також можуть бути використані. Останній шар мережі, вихідний шар, зазвичай інтерпретується як softmax / логістична регресія. Альтернативно, результат вихідного шару може бути поданий в класифікатор (наприклад, SVM) для отримання оцінки класу. Архітектура CNN може характеризуватися локальними зв'язками в проміжних, прихованих, шарах. Приховані шари часто чергуються між операціями згортки, випрямлення і об'єднання, з додатковим шаром нормалізації. В окремих випадках об'єднання об'єктів не враховується. У поєднанні з глибоким навчанням, ваги мережі отримуються за допомогою зворотного поширення зі спільними вагами в шарі. В даний час CNNs домінують в емпіричних оцінках у багатьох задачах розпізнавання на основі образів, включаючи розпізнавання дій [31].

Мотивовані сучасною діяльністю з різних завдань класифікації зображень, CNNs використовувалися також різними способами у завданнях класифікації відео. Метод включення часового простору або інформації про рух до 2D архітектури CNN був головною точкою розгалуження багатьох алгоритмів класифікації відео. Найбільш інтуїтивним підходом було б замінити 2D-операцію згортки та / або об'єднання з 3D-елементами для обліку додаткового (часового) домену у відео [31, 32]. Альтернативно, часова інформація у відео може бути узагальнена в одне RGB-зображення так, що стандартні 2D CNN можуть бути застосовані для розпізнавання дій [30]. Рекурентні нейронні мережі (RNN), які здатні вивчати тимчасову динаміку шляхом явного розгляду послідовностей активацій CNN у повторюваному вигляді, є іншим підходом, що враховує часовий вимір у відео [31, 30]. Для врахування нездатності RNN враховувати далекі часові зв'язки, численні алгоритми пропонують додавання тривалої короткострокової пам'яті (LSTM) в архітектуру, щоб дозволити мережі виявляти і синтезувати часову динаміку [33]. Останні методи вдаються до CNN, щоб отримати вектори ознак зображень [34] або дескриптори iDT поєднані з дескрипторами ознак HOG, HOF і MBH в якості входів до LSTM-RNN.

Інші загальні дескриптори примітивних функцій. Не всі дескриптори, можуть бути класифіковані як моделі представлення на основі фільтрів, оптичних потоків або згорткових нейроних мереж. Наприклад, eSURF [28], фільтр MACH [35].

Розширені прискорені надійні характеристики (eSURF) є дескриптором, заснованим на відгуках Хаара (*dx, dy, dt*) вздовж трьох осей на основі SURF [28]. Вектор ознак будується шляхом підсумовування зважених відгуків вейвлет Хаара перетворення як однорідних зразків по кожній точці інтересу. Хаар-вейвлет-відгук зважується гаусіаном для врахування геометричних деформацій і помилок локалізації [28].

Фільтр максимальної середньої кореляції (MACH) [35] є одним з небагатьох алгоритмів, які розглядають ущільнення набору даних в один шаблон. Внутрішньокласові варіації дій узагальнюються в один шаблон шляхом оптимізації чотирьох показників продук-

тивності: середньої кореляції (ACH), середньої кореляційної енергії (ACE), середньої міри подібності (ASM) і дисперсії вихідного шуму (ONV). Він використовує просторово-часовий потік регулярності (SPREF), щоб отримати напрямок, що найкраще відображає загальну закономірність (тобто напрямок, в якому інтенсивність пікселів змінюється найменше), замість інших оцінок руху, щоб уникнути труднощів, що виникають через розриви руху, проблеми апертури, і великі варіації освітленості. Об'єм поля потоку SPREF кожного з прикладів перетворюється за допомогою перетворення Кліфорда Фур'є (CFT) для його ефективності, що використовується для синтезу фільтра MACH. Композитний відео шаблон отримують шляхом об'єднання середнього значення CFT, коваріаційної шумової матриці, середньої спектральної щільності потужності і середньої матриці подібності для мінімізації ACE, ASM і ONV при максимізації ACH.

B. Спеціалізовані примітивні та допоміжні ознаки

Деякі алгоритми вимагають вилучення примітивних ознак і подальшого уточнення в допоміжні ознаки, перш ніж вони можуть бути передані класифікатору. Прикладами спеціалізованих примітивних ознак є методи основані на силуетах / контурах та методи на основі відстеження об'єктів.

Моделі, засновані на силуеті / контурі. Численні когнітивні дослідження показали, що люди здатні витягувати різноманітну корисну інформацію з силуетів, таких як розпізнавання об'єктів, маркування частин і порівняння подібності з іншими формами. Таким чином, відео силует може надавати достатню інформацію для розпізнавання, при цьому він є стійким до умов освітлення і інваріантним до зовнішнього вигляду людини. Після вилучення силуетів акторів інформація може бути описана в різних формах. Силуети можуть бути або безпосередньо перетворені в одновимірні сигнали, перетворені в бінарні або скалярні зображення, описані за допомогою моментів, або вони можуть бути складені для формування просторово-часових обсягів, далі розглянуто приклади цих варіантів.

R-перетворення є дескрипторами форм, які перетворюють силуетні зображення в 1D-сигнали. Виходячи з квадрата суми перетворення Радона, що зазвичай використовується для виявлення ліній на зображеннях, над змінними радіусами, визначається інваріантне перетворення Радона, що робить вирівнювання відео, для виявлення позиції актора, непотрібним. Крім того, для вирішення задачі чутливості до масштабу перетворень Радона R нормалізується. Це вдосконалене розширення перетворення Радона, перетворення *R*, було використано в деяких алгоритмах розпізнавання дій.

Двійкові зображення силуетів, що також називаються зображеннями енергії руху (MEI), можуть бути побудовані шляхом накопичення різниці між силуетами в послідовних кадрах і масштабованого зображення, яке називається зображенням історії руху (MHI), і може бути побудовано для зберігання останнього руху по кожному пікселі. MEIs та MHIs разом дають інформацію про місце розташування та часову історію руху відповідно. Ці зображення були додатково описані за допомогою Hu моментів для подальших порівнянь з іншими рухами. Багато алгоритмів, основаних на силуеті, показали чутливість до переміщення об'єкта та орієнтації його на камеру. Цю проблему можна вирішити, замінивши функцію індикації руху силуету функцією заповнення силуету, щоб створити обсяг історії руху (MHV) замість MHI. Хоча MHVs мають таку привабливу особливість інваріантності точки зору з використанням функції заповнення, великою проблемою є отримання точної функції, яка б точно моделювала координати об'єкта інтересу, особливо у відео з неконтрольованими умовами.

Послідовність силуетів або їх контури / межі можуть бути об'єднані уздовж часової осі, щоб створити ознаки зображення, які фіксують взаємозв'язок між простором і часом, що називається томами простору-часу (STV). Інформація про розташування загальних частин тіла (наприклад, голови, тулуба і кінцівок) може бути отримана шляхом обчислення середнього часу, необхідного кожній точці всередині STV, щоб досягти контуру за допомогою процесу випадкового блукання[3] або диференціальної геометрії. Рівняння Пуассона може бути використано для ідентифікації руху окремих частин та їх орієнтацій [3]. Хоча MHV і STV здаються подібними, MHV ілюструють функцію новизни за допомогою 3D реконструкції, в той час як часова інформація не може спостерігатися в STV.

Хоча силуети / контури надають корисну інформацію, отримання точної сегментації актора не гарантується, особливо в ситуаціях, коли фон не є статичним, оскільки виділення фону залишається не повністю вирішеною проблемою в сфері комп'ютерного зору. Більш того, кут зору може різко змінити силует людини, а особливості всередині контуру не можуть бути виявлені, оскільки людина представлена як окрема область.

Моделі на основі відстеження. Алгоритми відстеження, трекінгу можуть бути використані в розпізнаванні дій шляхом відстеження траєкторії всього актора у відео, щоб відрізнити актора від фону [23] або шляхом відстеження частин тіла [36] або областей локального інтересу .

Методи на основі відстеження потенційно є стійкими до змін у зовнішньому вигляді кожного актора або локального регіону і показали, що вони дають вражаючі результати на відео з низькою роздільною здатністю [23]. Проте, незважаючи на значний прогрес, відстеження залишається невирішеною проблемою в комп'ютерному зорі, оскільки ініціалізація відстеження може бути складною, як і підтримка відстеження протягом тривалого періоду часу, особливо в сценах з переповненими або динамічними фонами. Більш того, оскільки трекери об'єктів часто припускають постійний вигляд ділянок зображення з плином часу, це припущення може створювати проблеми, коли зовнішній вигляд об'єкта змінюється, особливо коли два об'єкти зливаються або роз’єднуються. Крім того, вихід трекера є чутливим до змін

дрейфу та освітлення, що призводить до проблем у наступних кроках при представленні дії.

Історично сфера розпізнавання дій наближалася до завдання розпізнавання дії за допомогою спеціалізованих примітивних ознак. Однак ознаки, на які спираються ці спеціалізовані примітивні особливості, вважалися несприятливими, оскільки виділення фону та відстеження залишаються невирішеними проблемами в сфері комп'ютерного зору. Деякі результати досягалися поєднанням алгоритмів на основі фільтрів і оптичних потоків. Сучасні показники розпізнавання дій досягаються за допомогою алгоритмів на основі CNN.

IV. Класифікатори

Після того, як необроблене відео було перетворено в набір ознак, що представляють дії, ці ознаки мають бути класифіковані. Набір тренувальних даних (маркованих або немаркованих) може використовуватися для класифікації тестових даних у певний попередньо визначений клас.

Враховуючи набір тренувальних даних, класифікатори, створюють модель, яка б узагальнила дані. Модель може бути сформована шляхом розбиття простору ознак на набір областей прийняття рішень [1]. Ці області надають рекомендації для класифікації ознак в один з класів. Регіональні області розділені межами рішення, які можуть бути описані набором дискримінантних функцій. Прикладами алгоритмів навчання, які зазвичай використовуються в алгоритмах розпізнавання і виявлення дій, є — машина опорних векторів (SVM), AdaBoost і штучні нейронні мережі (ANNs).

Машина опорних векторів (SVM) є одним з найбільш поширених інструментів класифікації, що використовуються в розпізнаванні і виявленні дій, наприклад, [8, 28, 26, 27, 18]. SVM навчається знаходити гіперплощину (або межу рішення), яка відокремлює позначені дані двох класів на відповідні групи. Найкращою гіперплощиною є та, яка розділяє два класи з найбільшою відстанню між найближчою точкою від кожного класу до гіперплощини. Оскільки розпізнавання дій передбачає класифікацію відеозаписів на численні дії (класи), необхідно використовувати багатокласовий SVM, що може бути зроблено шляхом застосування підходу "один проти всіх" [8]. Підхід "один проти всіх" приймає дані тренування з класу *k*, позначені як позитивні, а решта - як негативні приклади для підготовки *k*-ої моделі. Існують два типи ядер SVM: лінійні, і нелінійні. Для визначення того, що буде відповідним ядром для алгоритму, слід вивчити співвідношення між кількістю ознак і навчальними даними. Лінійне ядро є кращим, коли кількість ознак є великою (тобто багатовимірний просторім ознак) (наприклад, DT / iDT - ознаки) відносно кількості навчальних вибірок для запобігання перенавчання в просторі ознак. Коли є декілька ознак з великою кількістю зразків, нелінійне ядро буде кращим вибором. Хоча нелінійні ядра, як правило, досягають меншої частоти помилок, лінійні SVM мають меншу обчислювальну складність і вимагають менше памяті, ніж нелінійні SVM, що полегшує їх застосування в реальному часі [25].

Adaptive Boosting (AdaBoost) - це алгоритм навчання, який приймає кілька слабких класифікаторів, які трохи краще, ніж випадкові вгадування, і створює мета-класифікатор. Призначаючи різні ваги навчальним зразкам, класифікатори прорізному реагують на різні зразки. Ваги окремого класифікатора присвоюються залежно від його точності [23]. Цей підхід з деяким успіхом застосовувався в різних алгоритмах розпізнавання дій [24, 6]

Інший широко використовуваний алгоритм класифікації - штучні нейронні мережі (ANNs). Штучний нейрон (персептрон) у кожному шарі обчислює зважену суму його входів. Якщо сума перевищує деякий заданий поріг, персептрон виводить значення [1]. Персептрон моделює лінійну дискримінантну функцію, що розбиває простір ознак, використовуючи межу рішення. Використовуючи багатошарову мережу, можна навчитися відокремлювати лінійно не роздільні ознаки). Мережа навчається методом зворотного поширення, що передбачає повторне подання навчальних даних в мережу та коригування ваг у мережі для отримання бажаного виходу [1, 33]. Кількість нейронів у прихованих шарах регулює виразну потужність мережі [33]. Достатньо невеликої кількості прихованих нейронів для добре відокремлених або лінійно відокремлюваних моделей, але великі вкраплення зі складними розподілами вимагають більше прихованих нейронів. Хоча велика кількість прихованих нейронів створює дискримінаційну мережу, що зменшує похибку навчання, навчання стає надзвичайно трудомістким. Крім того, це може призвести до перенавчання даних, що призводить до моделювання випадкових шумів у даних тестування та поганого узагальнення даних [1]. ANN із занадто малим числом прихованих нейронів не матиме достатньо параметрів, щоб відповідати навчальним даним, що дає погані результати класифікації на тестових даних. Таким чином, пошук проміжної кількості прихованих нейронів є ключем до отримання хороших результатів класифікації за допомогою такого потужного інструменту.

Як і у випадку виявлення ознак, алгоритми на основі CNN також залишаються популярним вибором у розпізнаванні дії у часі. [32]. Останні дослідження локалізують дії у часі використовуючи ковзаюче часове вікно, для визначення схеми дії та її клас [32], або використовуючи довготривалу короткострокову память в поєднанні рекурентних нейроних мереж LSTM-RNNs [34]. Багато високопродуктивних алгоритмів розпізнавання дій спираються на CNNs для представлення особливих ознак і на LSTM-RNN для моделювання часового переходу дій, які дозволяють часове розпізнавання [34]. Проте LSTM-RNN не обмежуються локалізацією дій або об'єктів, що представляють інтерес. Вони можуть використовуватися для послідовного вдосконалення виявленого результату, що особливо корисно для виявлення послідовностей коротких рухів. Високі результати в розпізнаванні рухів показує метод, що використовує просторово-часову сгорткову нейронну мережу для розпізнавання рухів на основі скелету об’єкта[38].

Висновки

Проаналізувавши існуючі підходи до розпізнавання рухів на відеопослідовностях, виявлено особливості, сильні та слабкі сторони та обмеження різних методів виявлення ознак та їх класифікації. Передові методи, що показують кращі результати, широко використовують згорткові нейронні мережі. Один із таких методів – просторово-часова сгорткова нейронна мережа для розпізнавання рухів на основі скелету об'єкта[38], його і було обрано, як перспективний для подальшого дослідження. Також планується дослідити вплив динамічного фону та освітленості на виявлення ознак за допомогою CNN та її структуру.

Перелік посилань

[1] Y. Du, F. Chen, and W. Xu. Human Interaction Representation and Recognition Through Motion Decomposition. IEEE Signal Processing Letters, 14(12):952-955, 2007.

[2] C. Schuldt, I. Laptev, and B. Caputo. Recognizing Human Action: A Local SVM Approach. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 32-36, 2004

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. In IEEE International Conference on Computer Vision (ICCV), pages 1395-1402, 2005.

[4] Y. Ke, R. Sukthankar, and M. Hebert. Event Detection in Crowded Videos. In IEEE International Conference on Computer Vision (ICCV), pages 1-8, 2007.

[5] J. Yuan, Z. Liu, and Y Wu. Discriminative Subvolume Search for Ecient Action Detection. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2442-2449, 2009.

[6] I. Laptev and P. Perez. Retrieving Actions in Movies. In IEEE International Conference on Computer Vision (ICCV), pages 1-8, 2007.

[7] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions from Movies. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[9] M. Marszalek, I. Laptev, and C. Schmid. Actions in Context. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2929-2936, 2009.

[10] K. Soomro, A.R. Zamir, and M. Shah. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. Technical Report CRCV-TR-12-01, University of Central Florida, 2012.

[11] C. Snoek, B. Ghanem, J.C. Niebles, F.C. Heilbron, W. Barrios, V. Escorcia, and P. Mettes. ActivityNet: A Large-Scale Activity Recognition Challenge.

[12] Soo Min Kang and Richard P. Wildes. Review of Action Recognition and Detection Methods, York University, pages 49-61 Toronto, Ontario Canada

[13] J. Wang, P. Liu, M.F.H. She, A. Kouzani, and S. Nahavandi. Supervised Learning Probabilistic Latent Semantic Analysis for Human Motion Analysis. Neurocomputing, 100:134-143, 2013.

[14] A. Klaser, M. Marszalek, and C. Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. In British Machine Vision Conference (BMVC), 2008.

[15] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition via Sparse Spatiotemporal Features. In Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pages 65-72, 2005.

[16] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision (IJCV), 60(2):91-110, 2004.

[17] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 27(10):1615-1630, 2005.

[18] L. Yefet and L. Wolf. Local Trinary Patterns for Human Action Recognition. In 12th IEEE International Conference in Computer Vision (ICCV), pages 492-497, 2009.

[19] E. Shechtman and M. Irani. Space-Time Behavior-Based Correlation - OR - How to Tell If Two Underlying Motion Fields are Similar without Computing Them? IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 29:2045-2056, 2007.

[20] H. Ning, T. Han, D. Walther, M. Liu, and T. Huang. Hierarchical Space-Time Model Enabling Efficient Search for Human Actions. In IEEE Transactions in Circuits and Systems for Video Technology, volume 19, pages 808-820, 2006.

[21] O. Chomat and J. Crowley. Probabilistic Recognition of Activity Using Local Appearance. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1999.

[22] J.M. Gryn, R.P. Wildes, and J.K. Tsotsos. Detecting Motion Patterns via Direction Maps with Application to Surveillance. Computer Vision and Image Understanding (CVIU), 113(2):291-307, 2009.

[23] A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing Action at a Distance. In IEEE International Conference on Computer Vision (ICCV), pages 726-733, October 2003.

[24] A. Fathi and G. Mori. Action Recognition by Learning Mid-Level Motion Features. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[25] N. Dalal, B. Triggs, and C. Schmid. Human Detection Using Oriented Histograms of Flow and Appearance. In European Conference on Computer Vision (ECCV), volume 3952, pages 428-441, 2006.

[26] H. Wang, A. Klaser, C. Schmid, and C.L. Liu. Action Recognition by Dense Trajectories. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3169-3176, 2011.

[27] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In IEEE International Conference on Computer Vision (ICCV), 2013.

[28] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In European Conference on Computer Vision (ECCV), pages 404-417, 2006.

[29] J. Shi and C. Tomasi. Good Features to Track. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 593-600, 1994.

[30] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic Image Networks for Action Recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[31] X. Wang, A. Farhadi, and A. Gupta. Actions Transformations. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2658-2667, 2016.

[32] Z. Shou, D. Wang, and S.F. Chang. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1049-1058, 2016.

[33] J.Y.H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond Short Snippets: Deep Networks for Video Classication. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR),pages 4694-4702, 2015.

[34] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2678-2687, 2016.

[35] A. Basharat, A. Gritai, and M. Shah. Learning Object Motion Patterns for Anomaly Detection and Improved Object Detection. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1-8, 2008.

[36] C. Fanti, L. Zelnik Manor, and P. Perona. Hybrid Models for Human Motion Recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1166-1173, 2005.

[37] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. In IEEE

[38] Sijie Yan, Yuanjun Xiong, Dahua Lin, Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, Department of Information Engineering, The Chinese University of Hong Kong. arXiv:1801.07455v2, 25 Jan 2018.

УДК 004.02

Методы распознавания движений, действий людей на видеопоследовательностях

Солдатов Д. В.
e-mail kingit@bk.ru
КПИ им. Игоря Сикорского
Киев, Украина

***Аннотация*—В статье рассмотрено постановку проблемы распознавания движений объектов на видеопоследовательности, этапы ее решения, проведен анализ основных методов каждого из этапов. Рассмотрены ключевые сложности, возникающие при решении задачи. Приведены способы сравнения различных методов. Проанализировано существующие подходы к распознаванию движений на видеопоследовательностях, выявлены особенности, сильные и слабые стороны, и ограничения различных методов выявления признаков и их классификации. Избран метод для дальнейшего исследования и усовершенствования.**

Ключевые слова - распознавание движений; видеопоследовательность; оптический поток; SVM; CNN.

UDC 004.02

Action and Movements Recognition Methods

D. V. Soldatov
e-mail kingit@bk.ru
Igor Sikorsky Kyiv Polytechnic Institute
Kiev, Ukraine

***Abstract*—The article describes the formulation of the problem of recognition of the movements of objects in a video sequence, the stages of its solution, the analysis of the basic methods of each of the stages. A wide range of applications and growing requirements on the quality of recognition determines the relevance of the study. The process of action recognition and detection begins with extracting useful features, from the input video sequence. Features are then processed through a classier to identify the action class (for example, running, walking, jumping, various gestures). The article describes the main feature descriptors, in the filter-based category: histogram of oriented gradients, cuboid descriptor, scale-invariant feature transform, gradient location-orientation histogram, local trinary patterns, and spatiotemporal patches, optical flow-based descriptors: histograms of optical flow, the motion boundary histogram, dense trajectory, convolutional neural network-based descriptors. Some algorithms require the extraction of primitive features and further refinement of the auxiliary features before they can be passed to the classifier. Examples of the use of specialized primitive features are methods based on silhouettes / contours and methods based on object tracking. There are methods for classifying extracted features, including the following: support vector machines, adaptive boost, artificial neural networks, convolutional neural networks. The key difficulties arising in solving the problem are considered. There are ways to compare various methods. One of the ways to draw comparisons is to quantitatively evaluate each approach on the same database with the same protocol. From simple KTH datasets and Weizmannnd to Carnegie Mellon University Crowded Videos dataset and Microsoft Research Action Group dataset to more complex video conditions and large-scale UCF101 and ActivityNet datasets. Existing approaches to recognition of motion in video sequences are analyzed. The article reveals characteristics, strengths and weaknesses of the various methods of detecting features and their classification. Leading methods that show the best results widely use convolutional neural networks. One of such methods is a spatio-temporal graph convolutional neural network for action recognition based on the object's skeleton. A method for further research and improvement was chosen.**

Keywords - motion recognition; video sequence; optical flow; SVM; CNN.