

УДК 004.852

**О.Г. Руденко¹, О.О. Безсонов², Д.Г. Смерчинський³**¹ ХНЭУ, г. Харьков, Украина, oleg.rudenko@hneu.ua² ХНЭУ, г. Харьков, Украина, oleksandr.bezsonov@hneu.ua³ ХНЭУ, г. Харьков, Украина, xxdiviatorxx@gmail.com

РОЗПІЗНАВАННЯ ЕМОЦІЙНОГО СТАНУ ЛЮДИНИ ЗА ДОПОМОГОЮ ЗГОРТКОВИХ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

Запропоновано метод розпізнавання емоційного стану людини за допомогою згорткових штучних нейронних мереж, заснований на виділенні локальних ознак. Результати тестування підтверджують перспективність використання розробленого методу для комп'ютерного моніторингу виразів обличчя задля перевірки реакції людини та різноманітні події.

РОЗПІЗНАВАННЯ, ЗГОРТКОВА НЕЙРОННА МЕРЕЖА, ЕМОЦІЙНИЙ СТАН, ЛОКАЛІЗАЦІЯ

Вступ

Розпізнавання візуальних образів є одним з найважливіших компонентів систем управління та обробки інформації, автоматизованих систем та систем прийняття рішень. Завдання, пов'язані з класифікацією і ідентифікацією предметів, явищ і сигналів, що характеризуються кінцевим набором деяких властивостей і ознак виникають в таких галузях як робототехніка, інформаційний пошук, моніторинг та аналіз візуальних даних, дослідження штучного інтелекту. Головним завданням штучного інтелекту є побудова інтелектуальних інформаційних систем, які володіли б рівнем ефективності рішень неформалізованих задач, які можна порівняти з людськими можливостями або таких, що переважають їх.

При взаємодії з іншими людьми обличчя людини є важливим джерелом інформації. Вираз обличчя, міміка, жестикуляція при розмові, рухи головою є зручним, природним і способом передачі інформації. Нездатність комп'ютера з одного боку сприйняти, а з іншого – відтворити природні для людини способи спілкування ускладнює передачу і сприймання інформації при роботі з ЕОМ. Для досягнення мети розпізнавання комп'ютером рухів голови, міміки, зміни виразу обличчя потрібно реалізувати стійкі алгоритми аналізу і класифікації цифрового зображення обличчя людини.

Такі алгоритми можуть виконувати широкий спектр завдань комерційного і некомерційного характеру. Прикладом застосування таких алгоритмів може слугувати моніторинг виразів обличчя задля перевірки реакції людини на різні оголошення, рекламні анонси кінострічок, книг, телешоу. Спеціалісти вважають, що використання подібних програм для роботи з фокус-групами значно дешевше і ефективніше, ніж соціальні опитування.

На сьогоднішній день ще не існує алгоритму, який міг би ідентифікувати міміку та емоції людини з 100% точністю. Тим не менше, вже зараз можна реалізовувати програми для детектування міміки, які можуть з успіхом використовуватися у сфері освіти, медицини, охорони та забезпечення

безпеки, комп'ютерних ігор, реклами та інших областей. Користувачами сканерів людських емоцій також можуть стати люди з аутизмом, які насилу розпізнають міміку.

Метою цієї роботи є розробка і тестування методу розпізнавання емоційного стану людини, заснованого на застосуванні згорткових нейронних мереж.

1. Задача розпізнавання міміки людини

Головною метою математичної моделі є визначення домінуючих ознак об'єкта локалізації і розробка моделі для аналізу розпізнавання міміки людини. Для цього повинні бути вирішені наступні задачі:

1) Захоплення та відстеження особи.

Задача захоплення вимагає визначення: чи присутня особа на зображенні. Якщо присутня, то алгоритм повинен отримати розмір особи і координати її положення на зображенні. А відстеження вимагає подальшого визначення розміру і положення, що змінюються з плином часу.

2) Розпізнавання і опис особливостей особи.

Це завдання вимагає визначення положення органів (очі, ніс, рот і т.д.) на обличчі, а також визначення форми цих органів.

3) Класифікація.

Що стосується першої задачі, то слід виділити два її різновиди:

1. Локалізація обличчя (Face localization);

2. Відстеження переміщення обличчя (Face tracking).

Задача пошуку особи на зображенні може бути вирішена існуючими засобами. На сьогодні є кілька методів виявлення та локалізації обличчя на зображенні, які можна розділити на дві категорії:

1. Емпіричне розпізнавання;

2. Моделювання зображення обличчя.

До першої категорії відносяться методи розпізнавання «зверху-вниз» на основі інваріантних властивостей (invariant features) зображень обличчя, спираючись на припущення, що існують деякі ознаки присутності обличчя на зображенні, які

інваріантні щодо умов зйомки. Дані методи можна розділити на наступні підкатегорії:

- виявлення елементів і особливостей (features), які характерні для зображення особи (краю, яскравість, колір, характерна форма рис обличчя та ін.);
- аналіз виявлених особливостей, винесення рішення про кількість і розташування облич (емпіричний алгоритм, статистика взаємного розташування ознак, моделювання процесів візуальних образів, застосування жорстких і деформованих шаблонів тощо).

Для реалізації технології розпізнавання руху елементів обличчя необхідно взяти за основу спрощений варіант задачі виявлення особи. Це означає, що пошук обличчя можна буде проводити порівняно рідко (близько 10 кадрів / сек.), а оцінка руху органів обличчя повинна проводитися з більшою інтенсивністю.

Для коректної роботи алгоритму необхідно створення бази даних особливостей особи з подальшим тестуванням. Для більш точної реалізації емпіричних методів можуть бути використані моделі, які дозволяють врахувати можливості трансформації обличчя, а, отже, мають або розширений набір базових даних для розпізнавання, або механізм, що дозволяє моделювати трансформацію на базові елементи. Складнощі з побудовою бази даних класифікатора орієнтованих на самий різний спектр користувачів з індивідуальними особливостями, рисами обличчя і так далі, сприяє зниженню точності розпізнавання даного методу [1].

До другої категорії відносяться методи математичної статистики і машинного навчання. Методи цієї категорії спираються на інструментарій розпізнавання образів, розглядаючи задачу виявлення особи, як окремий випадок задачі розпізнавання. Зображенню ставиться у відповідність деякий вектор ознак, що використовується для класифікації зображень на два класи: обличчя / не обличчя. Найпоширеніший спосіб отримання вектора ознак — це використання самого зображення: кожен піксель стає компонентом вектора, перетворюючи зображення $n \times m$ в вектор простору $R^{(n \times m)}$, де n і m — цілі позитивні числа [2]. Недоліком такого подання є надзвичайно висока розмірність простору ознак.

Для адаптації активного контуру до зображення міміки необхідно провести відповідну бінаризацію досліджуваного об'єкта, тобто його перетворення в різновид цифрових растрових зображень, а потім вже слід проводити відповідну оцінку параметрів активного контуру і обчислення вектора ознак.

Активна контурна модель визначається як:

1. Множина точок N .
2. Внутрішня область енергії шару (internal elastic energy term).
3. Зовнішня область енергії шару (external edge based energy term).

Для поліпшення якості розпізнавання виділяються два кольорових класи — шкіра і губи.

Функція належності кольоровому класу має значення в діапазоні від 0 до 1. Також для простоти припустимо, що область рота диктора вже виділена якимось чином. У цьому випадку основні операції по обробці отриманого зображення, які нам необхідно виконати, представлені на рис. 1.

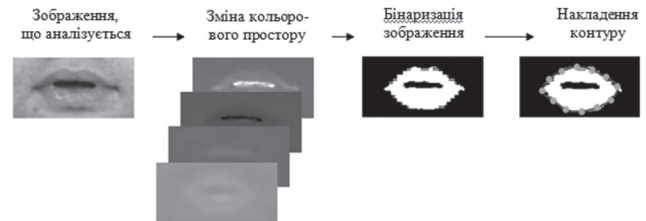


Рис. 1. Основні етапи обробки зображення

2. Нейромережові методи розпізнавання міміки

Нейромережові методи пропонують більш гнучкий підхід для вирішення задачі розпізнавання образів [3]. Навчаються нейронні мережі на наборі навчальних прикладів. Під час навчання мережі її вагові параметри налаштовуються за допомогою обраного алгоритму навчання. Навчена НМ може успішно застосовуватися для розпізнавання людини в різних умовах та при наявності різноманітних завад.

Типовий метод розпізнавання зображення складається з наступних компонентів:

1. Перетворення вихідного зображення в початкове уявлення (може включати в себе як предобробку, так і математичні перетворення, наприклад обчислення головних компонентів);
2. Виділення ключових характеристик (наприклад, береться перші n головних компонент або коефіцієнтів дискретного косинусного перетворення);
3. Механізм класифікації (моделювання): кластерна модель, метрика, нейронна мережа тощо.

3. Архітектура згорткової нейронної мережі

Згорткова нейронна мережа (ЗНМ) відноситься до багатошарових мереж і складається з двох частин: виділяючих ознак і класифікатора (рис. 2). Виділення ознак здійснюється частиною мережі, яка складається за допомогою чергування шарів згортки (С-шари) і шарів усереднення (S-шари). Пластини з нейронами в кожному шарі сприймають інформацію з попереднього шару за допомогою свого рецептивного поля. Рецептивним полем є квадратне вікно з нейронами на попередньому шарі, яке пов'язане за допомогою параметрів, що настроюються, з нейроном в поточному шарі. Класифікатором зазвичай буває двошаровий перцептрон [4].

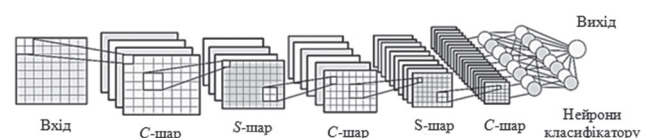


Рис. 2. Архітектура згорткової нейронної мережі

Особливість *C*-шару полягає в тому, що для всіх нейронів використовуються одні й ті ж параметри, що налаштовуються (ваги). Завдяки цьому, такі ваги, які називаються пов'язаними, дозволяють реагувати на деяку ознаку в будь-якому місці карти. Кожна карта *C*-шару може бути пов'язана з декількома картами попереднього шару, що дозволяє реагувати на інформацію з різних модальностей. У рецептивних полів є крок перетину, чим менше цей крок, тим більше рецептивних полів буде накладено на карту і тим більше нейронів буде в наступному шарі. Таким чином, виділення ознак відбувається за рахунок настроювання загальних ваг і розміру рецептивного поля.

Реакція нейрона визначається функцією активації. Функція активації береться однаковою для всіх елементів мережі і являє собою функцію, яка безперервно диференціюється в кожній точці і симетричну щодо осі абсцис. На вхід цієї функції подається згортка вхідних даних і настроюються ваги нейрона p . Найбільш поширеною функцією активації, яка б показала найменшу помилку узагальнення і високу швидкість збіжності помилки навчання, є гіперболічний тангенс.

Основним завданням *S*-шару є полегшення виділення загальних ознак за рахунок зменшення масштабу відфільтрованого відображення, отриманого на мапі попереднього *C*-шару. Рецептивні поля нейронів *S*-шару не перетинаються, і кожна карта *S*-шару пов'язана тільки з однією картою попереднього *C*-шару.

В кінцевому підсумку, карти шарів вироджуються в поодинокі нейрони, при цьому втрата точності відображень компенсується зростанням кількості карт і пов'язаністю між ними. Поодинокі нейрони вважаються входом для класифікатора, який складається з шару звичайних нейронів, в якому кожен нейрон пов'язаний з усіма нейронами входу. Кількість нейронів у класифікаторі відповідає кількості класів, до яких слід віднести вхідне зображення. Параметри для кожного нейрона в класифікаторі унікальні.

Навчання згорткової нейронної мережі відбувається методом зворотного поширення помилки, відповідно якому здійснюється коригування ваг за рахунок отриманого вектора помилки - різницею між отриманими і відомими даними. Тобто, якщо нейронна мережа неправильно класифікує подане на вхід зображення, синаптичні ваги оновлюються відповідно до помилки, і при наступному проході того ж зображення, мережа вірно класифікує об'єкт [5].

4. Використання ЗНМ для класифікації емоцій

Згорткові мережі являють собою варіацію архітектури багатозарового перцептронів і включають в себе згорткові шари, шари підвибірки (субдискретизація), і повнозв'язні шари. Архітектура згорткової мережі використовує переваги двовимірної

структури вхідних даних — зображень за допомогою методу локальної зв'язності, обмежуючи кількість зв'язків між нейронами прихованого згорткового шару і вхідними даними. Конкретно, кожен нейрон прихованого шару пов'язаний тільки з обмеженою локальною (які не мають розривів) ділянкою зображення.

Крім цього, нейронна мережа використовує загальні, або розділяючи ваги, накладаючи штучне обмеження на алгоритм навчання зворотним поширенням помилки, так, щоб кожен нейрон прихованого шару мав набір ваг, спільний з іншими нейронами цього шару. При прямому поширенні така мережа здійснює математичну операцію згортки вхідного зображення набором фільтрів, які подаються вагами нейронів прихованого шару.

Проміжними результатами мережі є так звані «карти ознак» - двовимірні матриці, що представляють собою результат згортки окремим фільтром.

Шар підвибірки або субдискретизація, виконує операцію угруповання карт ознак, розглядаючи регіони розміру $p \times p$ та агрегуючи значення, отримані в результаті згортки. Основне призначення субдискретизації - знизити варіативність даних, забезпечуючи стійкість до трансляцій локальної ознаки в межах окремого регіону. Таким чином, в умовах, коли одна і та сама ознака виявляється зміщеною на деяке значення ($\delta x, \delta y$) в межах, що не перевищують P , відповідний нейрон, що інкапсулює локальну субдискретизовану ознаку зображення, буде як і раніше активний. Таким чином мережа забезпечує інваріантність до просторових перетворень.

Однією з основних особливостей згорткової мережі є те, що така модель не володіє інформацією про локалізацію зображеного об'єкта - його конкретного місцезнаходження та орієнтацію в просторі. При цьому в рішенні прикладних задач управління і обробки інформації, знання параметрів локалізації є необхідною умовою - в залежності від розташування або пози об'єкта система обробки інформації може класифікувати зображення по-різному відповідно до покладених на неї задач. Крім того, без знання параметрів локалізації, згорткова мережа вразлива до узагальненого класу проблем цілісності, тобто модель, навчена деяким локальним ознакам зображеного об'єкта, буде позитивно класифікувати штучні зображення, де відповідні елементи розташовані в хаотичному порядку як помилку, яку з легкістю уникає людина.

Операція згортки забезпечує інваріантність по відношенню до трансляції локальних ознак (зрушенню по осях x та y). Архітектура згорткових мереж не передбачає стійкості до інших афінних перетворень, таких як обертання, дзеркальне відображення і масштабування. Для вирішення цієї проблеми, як правило, використовуються евристичні методи (вирівнювання зображення по лінії горизонту, використання просторових пірамід і різних відображених копій оригіналу). Перетворення,

які не є афінними по відношенню до плоского зображення, такі як обертання об'єкта в тривимірній площині, аналогічним чином не можуть бути оброблені за допомогою операції згортки. Для об'єктів, представлених в різних орієнтаціях, моделі доводиться навчатися різним, непересічним наборам ознак.

Згорткові нейронні мережі та інші підходи, які використовують локальні інваріантні ознаки, в процесі розпізнавання періодично вимушені навчатися ознакам, які максимізують інваріантність в рамках представленої вибірки, але не забезпечують дискримінантність представлення об'єкту в цілому.

Альтернативні підходи до виділення локальних ознак включають в себе методи класичного комп'ютерного зору, які не використовують навчальні моделі. Ці методи здійснюють пошук на зображенні характерних ділянок, які відповідають алгоритмічно явно заданим умовам. Серед них виділяються:

1. Виявлення країв / меж. Краєм називається ділянка зображення, яка представляє собою межу між двома контрастними регіонами, помітну людським оком. Математично точки, складові такої ділянки, визначаються як точки, де градієнт зображення має локальний максимум. Крім цього, виділення меж дозволяє знизити вплив деяких факторів, таких як освітлення і тіні, які не впливають на розпізнавання. Виділення меж (за допомогою фільтрів Кенні, Собеля, або згортки вейвлетами Габора) часто використовується як попередній етап обробки зображень в інших алгоритмах розпізнавання, в тому числі - згорткових мережах.

2. Виявлення кутів або «точок інтересу». До цієї групи належать алгоритми, які виділяють локальні ділянки зображення, максимально чутливі до змін. Відповідна умова формулюється наступним чином: якщо розглянути функцію ковзного вікна $w(x, y)$, то для будь-яких комбінацій значень зсуву (u, v) в двовимірній площині зображення знайдемо такі x і y , які максимізують

$$E(x, y) = \sum_{u, v} w(x, y) [I(x + u, y + v) - I(x, y)]^2, \quad (1)$$

де I — інтенсивність зображення в межах, заданих функцією вікна. Традиційно ця група алгоритмів (що включає в себе детектор Харріса, детектор Ши-Томасі та інші) використовувалася для відшукування кутів між прямими лініями, але в даний момент розглядаються також будь-які точки з високим значенням кривизни.

3. Виявлення ділянок неоднорідності. Під ділянками неоднорідності, на відміну від кутів, розуміються деякі безперервні регіони зображення, які відрізняються за значеннями кольору або інтенсивності від навколишнього фону, і при цьому схожі між собою. Як правило, такі ділянки відповідають локальним екстремумам зображення.

Перераховані локальні ознаки широко використовуються в задачах візуального трекінгу і стеження за об'єктом, але в чистому вигляді непридатні для задачі розпізнавання в силу своєї недискримінантної природи - такі методи не надають можливості відрізнити один кут (або ділянку неоднорідності) від іншого і визначити відмінність або схожість в числовому еквіваленті. Цим вимогам, проте, задовольняють підходи, які використовують ідею відшукування точок інтересу з використанням локальних дескрипторів, і представлені такими алгоритмами як SIFT, SURF і ORB [6].

Дескриптор є композицією ділянок зображення, локалізованих спільно, де для кожної ділянки або блоку розраховуються параметри орієнтації, масштабу, і деякі інші, що дозволяють з деякою вірогідністю ідентифікувати місце розташування ознаки, яке відповідає даному дескриптору. З урахуванням використання в дескрипторах параметрів орієнтації та масштабу, такі ознаки виявляються інваріантними по відношенню до обертання зображення, зміни масштабу і яскравості / контрасту. Такі алгоритми як SIFT і ORB, крім того, забезпечують можливість зіставлення зображень, співвідносячи однакові локальні ознаки один з одним.

Ознаки, що використовують локальні дескриптори, можуть ефективно використовуватися для розпізнавання зображень одного і того ж об'єкта під впливом афінних перетворень в тривимірному просторі. Локальність ознак дає можливість справлятися з проблемою оклюзії, забезпечуючи можливість зіставляти об'єкти по частинах. Основні проблеми методів сімейства SIFT - слабка стійкість до варіативності, що не дозволяє алгоритму відносити до одного класу об'єкти, які візуально відрізняються формою або текстурою.

Методи виділення локальних ознак дозволяють справлятися з деякими класами проблем розпізнавання зображень, забезпечуючи стійкість до оклюзії, знижуючи обчислювальне навантаження при обробці зображень високої розмірності і дозволяючи формувати інваріантні ознаки для виявлення об'єктів під дією інваріантних перетворень. При цьому нездатність здійснювати локалізацію об'єктів на зображенні і вразливість до хаотичного розташування локальних ознак призвела до появи методів, що враховують просторові відносини між локальними ознаками при розпізнаванні [7].

4. Моделювання

З використанням сучасних технологій штучного інтелекту та комп'ютерного зору було розроблено програмний комплекс для розпізнавання емоційного стану людини. Було розглянуто декілька варіантів розпізнавання міміки, перший з яких виявився невдалим, що було виявлено вже після створення та навчання мережі.

Початковий алгоритм включав в себе вирішення задачі класифікації за допомогою багатопарового

персептрону із ста тридцяти шістьма вхідними нейронами, які приймали шістдесят вісім точок антропометричної моделі обличчя. Експеримент виявився невдалим через те, що нейронна мережа не змогла знайти закономірностей.

На цьому етапі алгоритм було змінено на використання глибинної згорткової нейронної мережі, яка вважається однією з найкращих мереж для розпізнавання та класифікації графічних зображень. Слабким місцем ЗНМ є висока чутливість до різних перетворень зображень, наприклад, афінних, через що було вирішено стандартизувати вхідні зображення афінними перетвореннями.

Таким чином реалізацією алгоритму є програма на мові програмування Python з використанням сучасних бібліотек комп'ютерного зору dlib, OpenCV, та бібліотеки машинного навчання TensorFlow від Google. Нейронна мережа була навчена з використанням маркірованого набору даних розміру 20000 зображень. Мережа здібна розпізнавати 7 базових мімічних проявів, серед яких злість, радість, сум, здивованість, нейтральність, страх та відраза.

Після 100 епох навчання точність розпізнавання складала близько 80%, а для зберігання усіх параметрів мережі знадобилося 445 МБ пам'яті на жорсткому диску. Слід зазначити, що розпізнавання страху має більш високу імовірність помилки, на відміну від розпізнавання посмішки, яку можна розпізнати майже з 99% точністю. Приклади розпізнавання емоцій людини за допомогою ЗНМ наведено на рис. 3. На рисунках наведено фотографію людини, вірогідність кожної емоції у вигляді стовбової діаграми та смайлик, що відображає найбільш вірогідну емоцію. Графік зміни вірогідності правильного розпізнавання емоції протягом процесу навчання наведено на рис.4.



Рис. 3. Приклади розпізнавання емоцій людини за допомогою ЗНМ

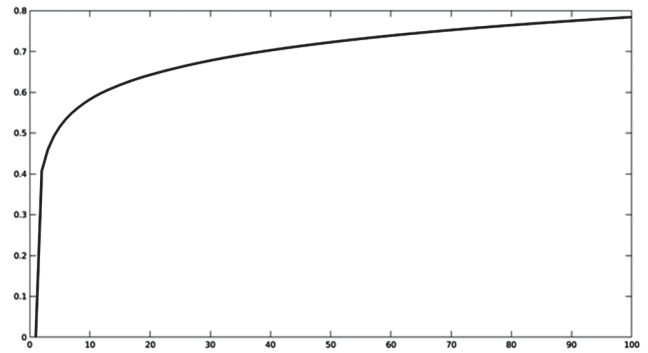


Рис. 4. Графік зміни вірогідності правильного розпізнавання емоції

Висновки

Як свідчать результати моделювання, використання ЗНМ є досить ефективним при вирішенні задач розпізнавання образів, зокрема задачі виявлення емоційного стану людини. Слід, однак, зазначити, що час навчання мережі при цьому є досить великим (кілька діб). Також істотною є кількість параметрів мережі, що значно ускладнює застосування цієї технології на мобільних пристроях.

Чутливість ЗНМ до різних перетворень зображень потребує деякої її модифікації, наприклад, використання афінних перетворень або інших методів нормалізації зображень.

Потребує також подальшого вивчення питання робастності алгоритмів розпізнавання в умовах різного роду завад.

Список литературы:

1. Droniou A. Gated autoencoders with tied input weights / A. Droniou, O. Sigaud // International Conference on Machine Learning. — 2013. — Pp. 17–24.
2. Гиренко А.В. Методи кореляційного виявлення об'єктів / А.В. Гиренко, В.В. Ляшенко, В.П. Машталір, Є.П. Путятін // Харків: АТ «БізнесІнформ», 1996. 112 с.
3. Krizhevsky A. ImageNet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, G. Hinton // Advances in Neural Information Processing Systems (Proc. of NIPS). — 2012. — P. 1106–1114.
4. 7. Deng L. Deep Learning: Methods and Applications / L. Deng, D. Yu // Fundamentals and Trends in Signal Processing. — 2013. — v. 7. - №3. — P. 197–387.
5. Sun Y. Deep convolutional network cascade for facial point detection / Y. Sun, X. Wang, X. Tang // Proc. of IEEE Conference on Computer Vision and Pattern Recognition. — 2013. — P. 3476–3483.
6. Rublee E. ORB: an efficient alternative to SIFT or SURF / E. Rublee // Computer Vision (ICCV), 2011 IEEE International Conference. — 2011. — Pp. 2564–2571.
7. Ding Ch. Multi-task Pose-Invariant Face Recognition / Ch. Ding, Ch. Xu, D. Tao // IEEE Trans. on Image Processing. — 2015. — v. 24. — v.3. — P.980–986.

Надійшла до редколегії 19.04.2017