УДК 004.93, 004.89
DOI: 10.18523/2617-3808.2022.5.62-67

Афонін А. О., Оксюта І. М.

АЛГОРИТМ ВИЛУЧЕННЯ КЛЮЧОВИХ ФРАГМЕНТІВ ЗОБРАЖЕНЬ У СИСТЕМАХ ВІДЕОПОШУКУ

У статті описано алгоритм вилучення ключових кадрів фрагментів зображень у процесі оброблення відеозаписів для індексації у відеопошукових системах. Наведено дослідження сучасних методів машинного навчання у вирішенні задач детекції та кластеризації зображень для досягнення найвищої точності в процесі оброблення. Запропоновано метрики оцінки якості кадрів для визначення найкращих і ранжування. Результати роботи алгоритму може бути використано в системах розпізнавання облич для подальшого присвоєння міток у процесі відеопошуку.

Ключові слова: системи відеопошуку, оброблення відеозаписів, детекція облич, комп'ютерний зір, машинне навчання.

Вступ

У наш час біометричні технології, такі як сканування райдужки ока, відбитків пальців та обличчя, стають все більш важливими та є ключовими у сучасних системах безпеки та розпізнавання людей. Зокрема, оброблення облич здійснюють для ідентифікації, визначення віку, статі, етнічної належності або емоційного стану. Тому зазвичай обличчя вважають найважливішим біометричним ідентифікатором. Подібні рішення широко імплементують у системах відеоспостереження як для державних органів безпеки, так і для приватних цілей.

Основною перешкодою для досягнення високої точності у відеопошукових системах, на відміну від статичних зображень, є переповненість оточення. Рух, оклюзія зображення, зміна освітленості або роздільної здатності, наявність додаткових перепон чи рух камери заважають коректно відпрацьовувати стандартним методам, що пристосовані до фото. Через це розпізнавання облич на кожному кадрі відеофайлу буде надмірним і дуже ресурсозатратним. Крім того, результати детекції в цих кадрах можуть бути або некоректними, або просто низької якості, непридатними для розпізнавання. З іншого боку, процес збирання даних із кожного відеокадру займає багато часу. Тому, щоб вирішити проблему величезної кількості кадрів даних у відео, нам потрібно вибрати кілька кадрів, облич на яких достатньо для завдань розпізнавання обличчя.

У цій роботі розглянуто загальну структуру та підхід до побудови системи для вилучення найважливіших кадрів, запропоновано методи оцінювання якості обличчя в кадрі. Також подано інтерфейс для оброблення відеофайлів і розбиття на ключові зображення. Ця система може стати одним із модулів відеопошукової системи, що спрямована на роботу з людськими обличчями або застосунком для препроцесингу даних у набір зображень, що можуть бути використані для навчання моделей машинного навчання, відповідно до специфіки предметної галузі.

Основні підходи до побудови системи статичного узагальнення

Оскільки для створення статичного підсумку використовують лише візуальну інформацію, ця категорія відеозведення є більш простою, швидшою та придатною для індексування й пошуку у відеоспостереженні. Головним процесом є саме вилучення ключових кадрів, які якнайкраще описуватимуть відеофайл і можуть бути використані в системі розпізнавання образів для отримання текстуального опису файлу всередині системи відеопошуку.

Вибір ключових кадрів може бути оснований на конкретних об'єктах у кадрах (рух, фон, обличчя, кольори). Більшість методів вибору ключових кадрів базуються на глобальних функціях, ураховуючи елементи, які представляють кадр у глобальній манері, як-от колір, текстура, гістограма зображення тощо. З іншого боку, методи, базовані на локальних особливостях, використовують характерні точки або регіони для створення підсумку, що дає нам змогу зосередитися на конкретних об'єктах у фреймах, отримуючи ключові кадри відповідно до конкретного об'єк-

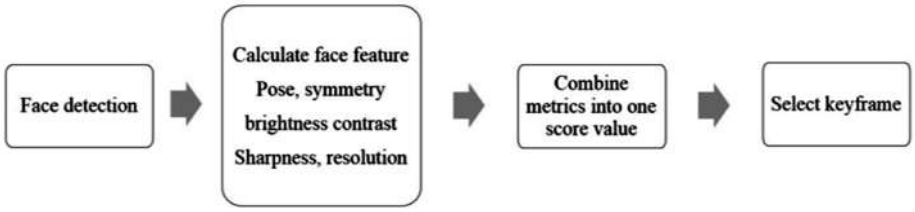

Рис. 1. Схема системи статичного узагальнення з використанням FQA

та. У рамках цієї роботи було конкретизовано застосування подібних систем безпосередньо для розпізнавання облич. Тому розглянемо ключові особливості оброблення відеоряду для такого завдання.

Методи вилучення ключових кадрів, основані на ключових точках, хоч і є важливими, однак не підходять для систем розпізнавання обличчя. По-перше, ці підходи не враховують труднощі, з якими стикаються під час оброблення зображень, як-от умови освітлення, зміна пози, вираз обличчя, оклюзія, відстань між обличчям і камерою тощо. По-друге, використовуючи ключові точки, ми не можемо гарантувати, що вилучимо найбільш нейтральне зображення обличчя, яке буде найкраще розпізнаватись. Із цих причин логічно зосередитись над виділенням ключового кадру на основі оцінки якості обличчя (face quality assessment, FQA), що формується на основі геометрії обличчя, зокрема пози і роздільної здатності зображення, фактора точності виявлення лиця, освітлення, виразу обличчя тощо.

У загальному вигляді систему статичного узагальнення для пошуку облич можна описати такими кроками (див. рис. 1):

1. Face Detection. На початку роботи з кожного кадру відеофайлу необхідно виокремити ділянки з обличчями. Найбільш поширеними алгоритмами детекції є моделі з використанням згорткових нейронних мереж. Такі алгоритми досить якісно обробляють зображення й дають змогу оцінити точність виявлення обличчя на знімку. В результаті цього кроку з відеофайлу отримують набір зображень виявлених облич.
2. Clustering. Для якісного процесингу й підрахунку якості для кожного обличчя, наявного у відеоряді (може бути багато людей на відео), потрібно відокремити їх одне від одного. Для цього застосовують алгоритми кластеризації, тобто зображення на основі їхнього вмісту розділяють на окремі групи з наданням міток за допомогою алгоритмів машинного навчання. В результаті цього кроку набір усіх облич розділяють на окремі класи (які стосуються різних людей).
3. FQA. Для оцінювання кожного зображення можуть бути застосовані різні метрики на розсуд розробника. Головною метою є різнобічна оцінка зображення та компонування загальної оцінки.
4. Ранжування. Відбір *N* найкращих кадрів для кожного обличчя залежно від результатів FQA.

До основних проблем, які можуть заважати коректній роботі такої системи, є похибки алгоритмів на кожному з кроків. І детектори, і кластеризатори обробляють вхідні дані з деякою похибкою, тому важливо підібрати оптимальні параметри для цих алгоритмів. Також важливим є час оброблення відеофайлів у такій системі. Важливо максимально оптимізувати процес, використовуючи швидкодійні мови програмування та принципи паралелізму.

Алгоритм вилучення ключових кадрів

Опис алгоритму

У процесі дослідження порівнювали різні методи машинного навчання для реалізації підзадач алгоритму. Для запропонованого рішення можна виділити такі кроки (рис. 2):

1. Вхідний відеофайл ділять на фрейми (кадри) за допомогою обробника.
2. Для кожного кадру застосовують детектор облич. Найкращим із погляду реалізації буде використання детектора MTCNN. Метод окрім виділення зон обличчя також надає додаткову інформацію про ключові точки обличчя, яку використовуватимуть для оцінки якості.
3. Вилучення ділянок з обличчями в окремі зображення. Результат роботи детекції виглядає як підмасив вхідного зображення, тому досить легко відокремити обличчя від загального фрейму.
4. Тепер для зображень в отриманому наборі необхідно провести кластеризацію, щоб відокремити обличчя, які належать різним людям. Для початку застосуємо дискретизацію зображень, тобто переведення фрейму у вектор,

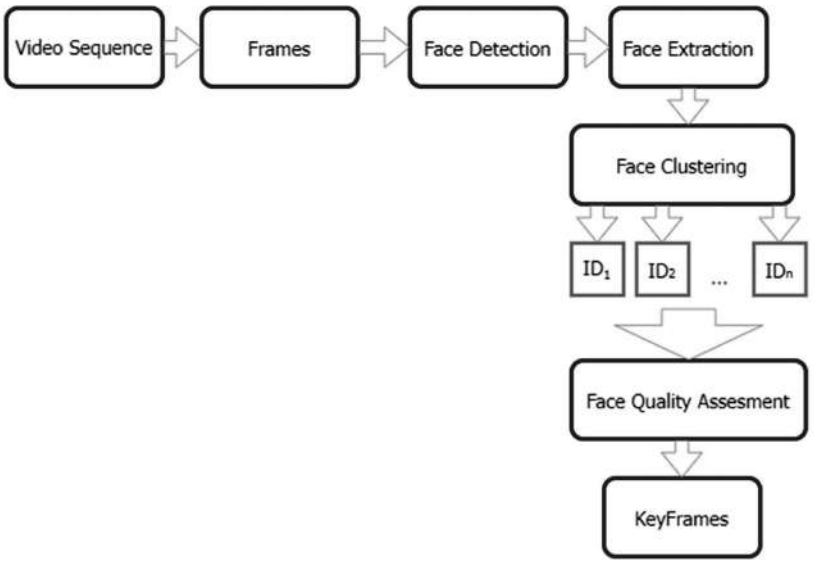

Рис. 2. Загальна схема алгоритму

який описує ключові характеристики зображення. Для цього використано попередньо натреновану модель FaceNet, яка призначена для створення embedding фото та переводить значення пікселів у 128-вимірний вектор ознак. Далі для кластеризації застосовуємо ієрархічну кластеризацію. Кількість кластерів, як необхідний параметр алгоритму, визначаємо шляхом побудови дендрограми. В результаті отримуємо вхідний набір зображень із наданими мітками класів.

5. Для зображень кожного класу проводимо оцінку якості облич. У результаті отримуємо відсортований набір для кожного класу.
6. Обираємо *K* ключових кадрів кожного класу. Значення параметра *K* задає користувач.

Функціонування алгоритму в такому вигляді є достатньо оптимальним із погляду швидкодії та отриманої якості оброблення.

Оцінка якості кадрів

Оцінку якості обличчя (FQA) проводять для кожного кадру-кандидата. У цій роботі використано п'ять метрик для оцінки якості обличчя на зображенні: поза голови, різкість, яскравість, роздільна здатність і точність детектора. Найкращі за якістю кадри для кожного обличчя буде відібрано як ключові. Вибрано саме ці показники через їхню важливість для оцінки якості обличчя. По-перше, саме обличчя в анфас є найбільш адекватним для представлення особистості. По-друге, через переміщення людей перед камерою на зображенні можуть бути шуми, це призводить до низької якості зображення обличчя. По-третє, незбалансована яскравість зображення може приховати деталі обличчя в будь-якому положенні (засвітити деякі важливі ділянки). По-четверте, у зображенні з низькою роздільною здатністю неможливо візуалізувати окремі компоненти обличчя. По-п'яте, також потрібно врахувати похибку, з якою спрацьовує детектор обличчя. Модель MTCNN надає показник confidence, який відображає впевненість алгоритму, що результатом детекції було обличчя.

Розглянемо докладніше кожен із цих показників і спосіб їх обчислення.

1. Оцінка впевненості – один із параметрів, що повертається після роботи детектора MTCNN.
2. Оцінка пози. Для підрахунку потрібно визначити, якою мірою ключова точка обличчя зміщена від центра зображення. Для визначення ключової точки обличчя використано точку, яка відповідає за ніс і визначається детектором у процесі оброблення. Для визначення центра зображення використаємо формулу:

$$x_c = \frac{x_2 - x_1}{2}\ y_c = \frac{y_2 - y_1}{2},$$

де (x_1, y_1) – координати верхнього лівого кута, (x_2, y_2) – координати нижнього правого кута. Для оцінки вимірюємо евклідову відстань між цими точками (див. рис. 3):

$$Pose\ Score = \frac{1}{1+\sqrt{(x_c - x_n)^2 + (y_c - y_n)^2}},$$

де (x_n, y_n) – координати точки «ніс» відповідно до результатів роботи детектора.

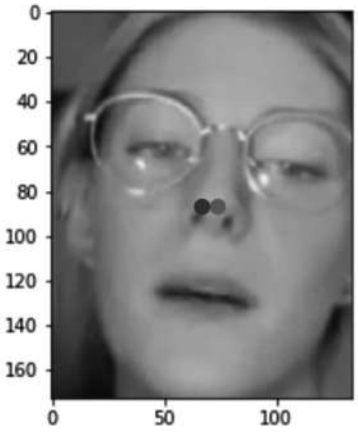

Рис. 3. Оцінка пози. Синя крапка (ліворуч) – центр зображення, червона (праворуч) – ніс

3. Оцінка яскравості. Деякі ділянки обличчя можуть бути погано помітні на темних зображеннях. Тому нам потрібно знайти зображення обличчя з найкращим розподілом освітлення. Показник яскравості дорівнює сумі значень усіх пікселів, поділених на площу зображення:

$$Brightness\ Score = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} B_{ij}}{H \cdot W},$$

де H, W – висота й ширина зображення, а $B_{i,j}$ – значення пікселя в координатах i, j. Діапазон значень кожного пікселя $B \in [0, 255]$.

4. Оцінка різкості. Оскільки набір зображень відображатиме рух людей на відео, то деякі з кадрів матимуть ефект розмиття. Ранжування за різкістю має допомогти знайти найбільш нейтральні зображенні й відсіяти розмиті. Щоб обчислити показник різкості, потрібно виконати розмиття зображення обличчя за допомогою оператора Гауса (рис. 4). Оцінку визначають як абсолютну різницю між вхідним і розмитим зображенням, поділену на розмір зображення:

$$Sharpness\ Score = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} |I - G(I)|}{H \cdot W},$$

де I – вихідне зображення, $G(I)$ – зображення після розмиття за Гаусом. Ядро $G(I)$ має розмір (5,5) зі стандартним відхиленням 1,0 у напрямках X і Y. W, H – ширина та висота зображення відповідно.

5. Оцінка роздільної здатності відображає, наскільки придатними до подальшого оброблення є зображення. Розраховують як добуток висоти на ширину зображення:

$$Resolution\ Score = H \cdot W.$$

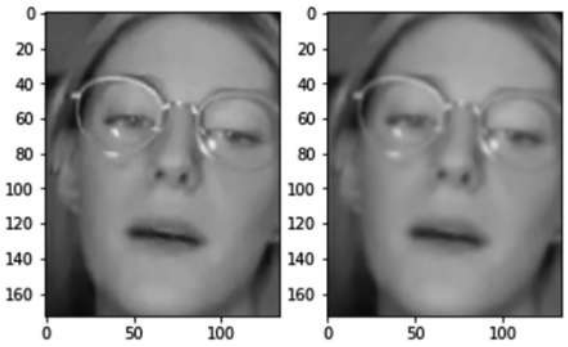

Рис. 4. Розмиття за допомогою оператора Гауса

Після обчислення п'яти показників для кожного зображення обличчя вилучені ознаки об'єднують у єдине значення якості. Спочатку кожну оцінку нормалізують відповідно до максимального значення цього показника для цієї послідовності. Потім усі оцінки характеристик об'єднують у загальне значення номінальної якості зображення:

$$Face\ Quality = \sum_{i=1}^{N} \frac{S_i}{S_{i_{max}}},$$

де S_i – значення оцінки i, $S_{i_{max}}$ – максимальне значення оцінки i для заданої послідовності зображень, $N=5$ – кількість оцінок.

Отже, для кожного набору зображень, які буде отримано після кластеризації (розділення на окремих людей), застосовуємо FQA. В результаті отримуємо набір відсортованих за спаданням якості зображень для кожної людини, що зображена на відео та яку вдалося розпізнати детектором облич. Залежно від параметра K, який задає користувач і який вказує на бажану кількість ключових кадрів для кожного обличчя, формуємо результат із відсортованого списку.

Аналіз результатів

Для реалізації цього проєкту було вибрано мову програмування Python 3.7. Використано бібліотеки: decord, OpenCV, mtcnn, tensorflow(keras), scikit-learn. Для демонстрації роботи алгоритму було створено користувацький інтерфейс на платформі **Streamlit.io**. Вона має пряму інтеграцію з Python і дає змогу досить швидко зробити обгортку для задач машинного навчання, містить інструменти для швидкого деплою вебдодатків.

Реалізований інтерфейс представлено на рис. 5.

Аналізуючи роботу системи на різних вхідних даних, вдалося виокремити два основні фактори, які впливають на якість оброблення.

1. Час оброблення. Оскільки використано готові рішення (методи детекції, дискретизації,

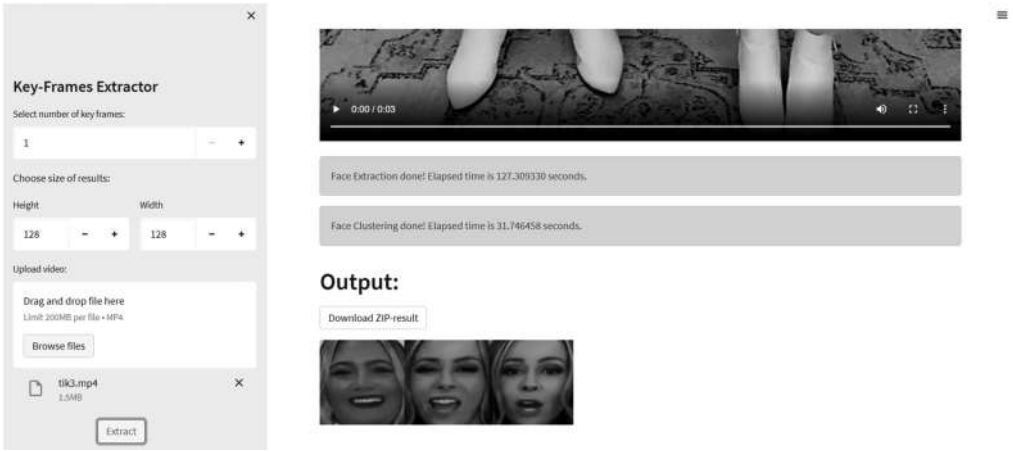

Рис. 5. Графічний інтерфейс програми

кластеризації) для цього алгоритму, достатньо оптимізовані на швидкодійній мові C++, основним чинником є потужність процесора комп'ютера, на якому ведуться обчислення. Для відпрацювання використовували стандартний процесор CPU. Тому для дуже довгих відео або відео з великою кількістю людей час оброблення був досить великим. Цю проблему можна вирішити шляхом запуску алгоритму на процесорах GPU, використовуючи принципи паралелізму.

2. Параметр кількості кластерів. Оскільки в запропонованому алгоритмі для визначення цього параметра використовують метод побудови дендрограми, автоматичне обрання кластерів може бути не дуже точним. Однак ручний метод оцінки не є зручним і не може виконуватись автоматично, тому в процесі кластеризації можуть виникати неточності. Цю проблему можна спробувати вирішити або іншим методом підбору кількості кластерів, або винесенням цієї задачі на розсуд користувача як додатковий параметр алгоритму. Відповідно у вікні редагування параметрів користувач вказуватиме, скільки облич наявні в цьому відеофайлі.

Висновки

У статті досліджено функціонування систем відеопошуку на основі вмісту. Такі системи є потужними інструментами для задач комп'ютерного зору й мають широкий спектр використання. Для реалізації алгоритму вилучення ключових кадрів обличчя з відеофайлів було визначено основні кроки та описано необхідні методи оброблення облич. Сформовано алгоритм повного оброблення вхідного відеозапису, запропоновано метрики оцінки якості зображень обличчя.

Отримані результати роботи алгоритму можна аналізувати з двох сторін: час оброблення й точність. На час оброблення дуже впливає потужність комп'ютера, на якому виконується програма. На точність впливає оцінка кількості людей на відео. Відкритим питанням залишаються задача кластеризації зображень і метод оцінки якості. Для розділення облич на окремі класи можна спробувати реалізувати інші методи кластеризації. Для оцінки якості своєю чергою можуть бути досліджені показники, що базуються на даних пікселів або використовують для оцінки інші методи комп'ютерного зору. Також можлива адаптація алгоритму до систем, які працюють у реальному часі.

Запропонований алгоритм можна використовувати як самостійне рішення для оброблення відео при вирішенні задач комп'ютерного зору та машинного навчання або як повноцінний модуль системи відеопошуку на основі контенту.

Список літератури

1. Ciaparrone G. A comparison of deep learning models for end-to-end face-based video retrieval in unconstrained videos [Electronic resource] / G. Ciaparrone, L. Chiariglione, R. Tagliaferri // Neural Computing and Applications – 2021. – No. 34. – Pp. 7489–7506 pp. – Mode of access: https://link.springer.com/content/pdf/10.1007/s00521-021-06875-x.pdf.
2. Nouyed M. I. Evaluation and Understandability of Face Image Quality Assessment [Electronic resource] / M. I. Nouyed // West Virginia University Libraries. – 2019. – Mode of access: https://researchrepository.wvu.edu/cgi/viewcontent.cgi?article=8473&context=etd.
3. Saoudi E. M. A distributed Content-Based Video Retrieval system for large datasets [Electronic resource] / E. M. Saoudi,

S. Jai-Andaloussi // Journal of Big Data. – 2021. – No. 8. – Mode of access: https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-021-00479-x.pdf.
4. Schroff F. FaceNet: A Unified Embedding for Face Recognition and Clustering [Electronic resource] / F. Schroff, D. Kalenichenko, J. Philbin // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2015. – Pp. 815–823. – Mode of access: https://arxiv.org/pdf/1503.03832.pdf.
5. Shi Y. Face Clustering: Representation and Pairwise Constraints [Electronic resource] / Y. Shi, C. Otto, A. K. Jain // IEEE Transactions on Information Forensics and Security. – 2018. – No. 7. – Pp. 1626–1640. – Mode of access: https://arxiv.org/pdf/1706.05067.pdf.
6. Yang Y. Content-Based Video Retrieval (CBVR) System for CCTV Surveillance Videos [Electronic resource] / Y. Yang, B. C. Lovell, F. Dadgostar // 2009 Digital Image Computing: Techniques and Applications. – 2009. – Pp. 183–187. – Mode of access: https://ieeexplore.ieee.org/document/5384989.

References

Ciaparrone, G., Chiariglione, L., & Tagliaferri, R. (2021). A comparison of deep learning models for end-to-end face-based video retrieval in unconstrained videos. *Neural Computing and Applications*, *34*, 7489–7506. https://link.springer.com/content/pdf/10.1007/s00521-021-06875-x.pdf.

Nouyed, M. I. (2019). *Evaluation and Understandability of Face Image Quality Assessment*. West Virginia University Libraries https://researchrepository.wvu.edu/cgi/viewcontent.cgi?article=8473&context=etd.

Saoudi, E. M., & Jai-Andaloussi, S. (2021). A distributed Content-Based Video Retrieval system for large datasets. *Journal of Big Data*, *8*. https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-021-00479-x.pdf.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823. https://arxiv.org/pdf/1503.03832.pdf.

Shi, Y., Otto, C., & Jain, A. K. (2018). Face Clustering: Representation and Pairwise Constraints. *IEEE Transactions on Information Forensics and Security*, *7*, 1626–1640. https://arxiv.org/pdf/1706.05067.pdf.

Yang, Y., Lovell, B. C., & Dadgostar, F. (2009). Content-Based Video Retrieval (CBVR) System for CCTV Surveillance Videos. *2009 Digital Image Computing: Techniques and Applications*, 183–187. https://ieeexplore.ieee.org/document/5384989.

A. Afonin, I. Oksiuta

ALGORITHM FOR EXTRACTION OF KEYFRAMES OF IMAGES IN VIDEO RETRIEVAL SYSTEMS

As a part of this work, there was a study of image processing algorithms used in video search systems.

With the development of search engines and an increase in the types of queries possible for searching, the need for indexing an increasing amount of diverse information is growing. New data in the form of images and videos require new processing techniques to extract key content descriptions. In video search engines, according to this description, users can find the video files most relevant to the search query. The search query, in turn, can be of various types: text, search by image, search by video file to find a similar one, etc. Therefore, it is necessary to accurately describe the objects in the video in order to assign appropriate labels to the video file in the search engine database.

In this article, we focused on the algorithm for extracting key frames of faces from a video sequence, since one of the important objects in the video are people themselves. This algorithm allows you to perform the initial processing of the file and save the identified frames with faces in order to later process this data with the help of the face recognition algorithm and assign the appropriate labels. An alternative application for this algorithm is the current processing of video files to form datasets of faces for the development and training of new computer vision models. The main criteria for such an algorithm were: the accuracy of face detection, the ability to distinguish keyframes of all people from each other, comprehensive evaluation of candidate frames and sorting by the relevance of the entire set for each face.

After an analysis of existing solutions for specific stages of the algorithm, the article proposes a sequence of steps for the algorithm for extracting key frames of faces from a video file. An important step is to assess the quality of all candidates and sort them by quality. For this, the work defines various metrics for assessing the quality of the frame, which affect the overall assessment and, accordingly, the sorting order. The article also describes the basic version of the interface for using the proposed algorithm.

Keywords: content-based video retrieval systems, video processing, face detection, computer vision, machine learning.

Матеріал надійшов 17.08.2022