© 2022 г. А.В. БОБКОВ, канд. техн. наук
(Alexander.Bobkov@bmstu.ru),
Х. АУНГ (happyland27057@gmail.com)
(Московский государственный технический университет
им. Н.Э. Баумана, Москва)

ИДЕНТИФИКАЦИЯ ЧЕЛОВЕКА ПО ВИДЕОИЗОБРАЖЕНИЮ В РЕАЛЬНОМ ВРЕМЕНИ НА ОСНОВЕ СЕТЕЙ YOLOv2 И VGG 16

Данная работа посвящена задаче распознавания лиц по видео. На сегодняшний день методы распознавания лиц сделали большой шаг вперед, однако распознавание видео с его низким качеством, сложными условиями освещенности и требованиями работы в реальном времени попрежнему остается сложной и до конца нерешенной задачей.

В работе используется аппарат сверточных сетей для различных этапов обработки: для захвата и обнаружения лица, для построения вектора признаков, и наконец, для распознавания. Все алгоритмы реализованы и исследованы в среде Matlab для упрощения их дальнейшего экспорта во встраиваемые приложения.

Ключевые слова: сверточная нейронная сеть VGG16, распознавание лиц, алгоритм обнаружения объектов YOLOv2, глубокое обучение, база данных лиц.

DOI: 10.31857/S0005231022100099, **EDN:** AKSNXL

1. Введение

Задача распознавания лиц — одна из наиболее интересных тем в области компьютерного зрения. Это способность распознавать или идентифицировать личность человека, анализируя различные черты лиц. Система распознавания лиц обеспечивает огромные преимущества по сравнению с другими решениями биометрической безопасности, такими как распознавание радужной оболочки глаза и отпечатков пальцев. Система фиксирует биометрические измерения человека с определенного расстояния, не взаимодействуя с ним. В приложениях для сдерживания преступности эта система может помочь многим организациям идентифицировать человека, у которого есть какое-либо уголовное прошлое или другие юридические проблемы.

В последние годы задача распознавания лиц оставалась крайне популярной и актуальной темой исследований, и в этом направлении было разработано значительное количество методов, которые обеспечивают очень высокую точность распознавания и которые можно использовать на практике [1, 2]. Из классических методов машинного обучения наиболее широкое распространение получили фильтр Виолы–Джонса и его модификация, использующая гистограмму направленности градиента. Основным направлением развития исследований сегодня остается поиск эффективных подходов с использованием сверточных сетей и глубокого обучения.

Современный подход к распознаванию лиц связан с построением сетей со сложной топологией и большим числом сверточных слоев, что обеспечивает возможность высокоточной классификации при наличии смены ракурса, мимики, маскирующих признаков и т.д. Это такие сети, как VGG-Face [11], Google FaceNet [12], Facebook DeepFace [13], FaceID и другие. Однако попытки применять их распознаванию лиц на видеоизображении дают низкие результаты из-за низкой скорости работы и низкого качества изображения. Попытки же использовать более быстрые и менее точные сети, такие как OpenFace на основе MobieNet [14], сразу ведут к резкому снижению точности распознавания.

При распознавании лиц на видео, помимо высокой точности, также требуется обеспечить высокую скорость обработки больших массивов информации, причем, как правило, качество изображения достаточно низкое из-за воздействия шума и ракурса съемки. Тем не менее видео несет в себе гораздо больше информации о лице, нежели одиночная фотография. Все это заставляет искать новые методы для распознавания лиц по видео. Особенности задачи распознавания по видео не позволяют решить задачу каким-либо одним инструментом, например, обучив нейронную сеть, как во многих других задачах: такое решение будет либо слишком громоздко и неспособно работать в режиме реального времени, либо, наоборот, будет быстрым, но неспособным решать задачу с требуемой точностью. Решением здесь будет являться использование совокупности сетей, каждая со своими свойствами, для наиболее качественного решения задач каждого из этапов распознавания.

В данной работе рассматривается подход, связанный с обнаружением и отслеживанием области интереса при помощи быстродействующей поисковой сети. Наличие области интереса позволяет распознавать не все изображение во все моменты времени, а лишь отдельные фрагменты наиболее удачных кадров. Потенциально такой подход позволяет повышать качество распознавания, собирая более качественное изображение по последовательности кадров, с коррекцией ракурса съемки и восстановлением трехмерной формы, но в данной работе данная задача не ставилась.

Методы детектирования лиц традиционно делятся на методы жестких шаблонов и методы деформируемых моделей.

Методы деформируемых моделей строят модель лица на основе деформируемых частей [15–17] для моделирования потенциальной деформации между элементами лица. Методы также могут сочетать обнаружение всего лица и локализацию его элементов [18].

Методы с использованием жестких шаблонов, в свою очередь, делятся на следующие группы:

— каскадные фильтры и вариации бустинга; к основным представителям этого семейства алгоритмов относятся алгоритм распознавания лиц Виолы–Джонса и его варианты [1, 19];

— методы на основе обобщенного преобразования Хафа и его вариации [20, 21];

— алгоритмы, основанные на сверточных нейронных сетях и сетях глубокого обучения [22, 23–25].

Сети глубокого обучения в последнее время показывают очень высокую точность, поэтому в настоящее время именно они рассматриваются как наиболее перспективный подход для поиска лиц.

В данной статье используется аппарат сверточных сетей с отдельным детектором общих признаков, детектором лиц и классификатором лиц.

В роли детектора признаков выступает относительно небольшая и производительная сеть ResNet-18 с редуцированными верхними слоями, предварительно обученная на большом количестве классов объектов.

В качестве быстродействующего детектора лиц выбрана модифицированная сеть YOLOv2 [5] (You Only Look Once — посмотри на изображение только один раз). Сеть YOLO обладает очень высокой производительностью и широко применяется, например, в задачах распознавания дорожных сцен, однако ее применение для идентификации лиц затруднено низкой точностью классификации.

Для решения задачи окончательного высокоточного распознавания лица в найденном положении использовалась предварительно обученная сверточная сеть VGG16 [10] без последних слоев многослойного классификатора. Эта сеть обеспечивает построение вектора признаков, устойчивого к изменениям освещенности и к небольшим изменениям ракурса, что является важным для рассматриваемой задачи.

Наконец, для поиска наилучшего совпадения с параметрами лиц из базы использовалась косинус-метрика, величину которой удобно рассматривать как вероятность совпадения лиц.

Для упрощения экспорта полученных алгоритмов на целевую вычислительную платформу решено было использовать среду MATLAB. Данная среда, с одной стороны, содержит библиотеки с открытым исходным кодом, доступным для изучения и воспроизведения, а с другой — позволяет непосредственно переносить написанный код на другие языки программирования.

Использование среды MATLAB потребовало создания интерактивной среды для исследования алгоритмов и методов распознавания. Для этого использовался дизайнер приложений MATLAB, который представляет собой интерактивную среду разработки для проектирования макета приложения и программирования его поведения.

Исследование, проведенное в статье, направлено на изучение работоспособности, тестирование производительности и сравнение точности результатов обнаружения и распознавания лица комбинации методов YOLOv2 и VGG16 с другими известными методами. Эксперименты показали как высокую производительность подхода, позволяющего работать в режиме реального времени, так и высокую точность, позволяющую использовать подход в реальных практических задачах.

2. Обнаружение лиц с использованием метода YOLOv2 на базе ResNet-18

Сеть ResNet — это одна из самых мощных глубоких нейронных сетей, которая достигла прорывных результатов в классификации ILSVRC 2015 [7].

Таблица 1. Архитектура ResNet-18 для извлечения признаков

Имя слоя	Размер фильтра	Размер выхода
Входной слой изображения	224×224×3	224×224×3
Conv_1	7×7×64 Maxpool 3×3	112×112
Conv_2	[3×3×64]×2	56×56
Conv_3	[3×3×128]×2	28×28
Conv_4	[3×3×256]×2	14×14
Входной слой признаков	14×14	14×14

ResNet добилась отличных результатов обобщения по другим задачам распознавания и заняла первое место по обнаружению ImageNet, локализации ImageNet, обнаружению COCO и сегментации COCO в конкурсах ILSVRC и COCO 2015. Существует много вариантов архитектуры ResNet: это одна и та же структура, но с разным количеством слоев. Различные версии ResNet — это ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-110 и т.д.

В данной работе использовалась предварительно обученная сеть ResNet-18 только для извлечения признаков изображений. Размер входного изображения составляет 224×224 с RGB. После того, как изображение прошло через сеть, на выходе получается отклик размером 14×14, т.е. сеть осуществляет 16-кратную понижающую дискретизацию. Это достаточно для извлечения признаков лица (табл. 1).

Выход алгоритма обнаружения объектов YOLO v2 — отклик размера $S \times S$, где S — количество ячеек сетки. Каждая ячейка содержит пять параметров (x, y, w, h) и $Pr(obj)$, где x, y — координаты центра ограничивающей рамки, w, h — ее ширина и высота, $Pr(obj)$ — вероятность нахождения объекта внутри рамки. Показатель достоверности отражает вероятность включения в модель целевого объекта и точность блока обнаружения предсказания. Показатель $C(obj)$ достоверности определяется так:

$$C(obj) = Pr(obj)^* IoU(Pred, Gtruth).$$

Если искомый объект отсутствует в ячейке, то $Pr(obj)$ будет равен нулю, а доверительный балл должен быть равен нулю: $C(obj) = 0$.

IoU — это величина перекрытия найденной ограничивающей рамки и рамки из обучающей выборки, т.е. отношение их пересечения и объединения:

$IoU(Pred, Gtruth)$ = (перекрывающаяся область предсказанной рамки и рамки обучающей выборки)/(вся область предсказанной рамки и рамки обучающей выборки).

После получения достоверности каждой рамки, рамки с низкой достоверностью удаляется путем сравнения с пороговым значением, а затем выполняется удаление оставшихся рамок, отклик которых не является локальным максимумом.

Метод YOLO v2 использует суммарную квадратическую ошибку в качестве функции потерь. Метод пытается оптимизировать следующие многосо-

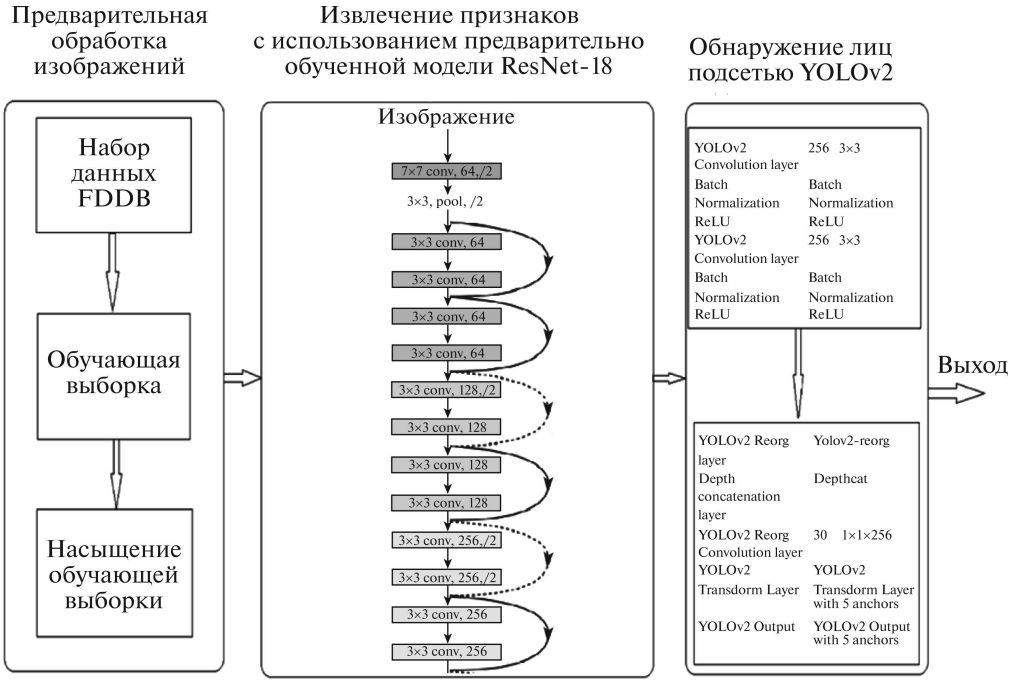

Рис. 1. Схема процесса обнаружения лиц предлагаемой модели.

ставные потери: потери локализации (ошибка определения положения объекта), потери доверия (ошибка определения вероятности обнаружения) и потери классификации (ошибка определения класса объекта).

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{obj}\left[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2\right]+$$

$$+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{obj}\left[\left(\sqrt{w_i}-\sqrt{\hat{w}_i}\right)^2+\left(\sqrt{h_i}-\sqrt{\hat{h}_i}\right)^2\right]+$$

$$+\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{obj}\left(C_i-\hat{C}_i\right)^2+\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{noobj}\left(C_i-\hat{C}_i\right)^2+$$

$$+\sum_{i=0}^{S^2}\mathbb{1}_{i}^{obj}\sum_{C\in classes}\left(P_i(c)-\hat{P}_i(c)\right)^2,$$

где $\mathbb{1}_{ij}^{obj}=1$ если j-й граничный прямоугольник в i-й ячейке сетки отвечает за обнаружение объекта, то в противном случае 0;

$\lambda_{coord}=$ увеличение веса для потери в координатах граничного поля, по умолчанию 5,

$\lambda_{noobj}=0{,}5$;

$(x, y, w, h) =$ потери локализации между обучающей выборкой и предсказанной рамкой;

C — потери достоверности обнаружения и $P(c)$ — вероятности принадлежности к классу. На рис. 1 представлена схема процесса обнаружения лиц предлагаемой модели.

3. Распознавание лиц с использованием предварительно обученной модели VGG16

Для распознавания лиц использовалась предварительно обученная модель сети VGG16. Модель VGG16 имеет большое количество гиперпараметров. Размер входного изображения первого слоя составляет 224×224 с кодированием RGB. Изображение пропускается через последовательность сверточных слоев, в которых использовался сверточный фильтр размером 3×3 с шагом 1, и всегда используется один и тот же слой субдискретизации maxpool 2×2 с шагом 2. Расположение слоев в этой архитектуре выглядит следующим образом: сверточные слои, слои ReLU и слои субдискретизации. В конце модели есть два полносвязных слоя, за которыми следует слой классификатора softmax для вывода данных. Эта сеть VGG16 является довольно большой сетью и имеет около 138 млн обучаемых параметров (рис. 2). При предъявлении сети изображения лица на выходе сети появляется его описание в виде вектора признаков. При этом одинаковые лица будут иметь схожие признаки, а разные соответственно несхожие, даже при наличии мешающих факторов: изменения ракурса, освещенности и т.д. Это позволяет распознавать лица, заранее неизвестные сети, путем сравнения вектора признака, сгенерированного сетью, с ранее заданным образцом.

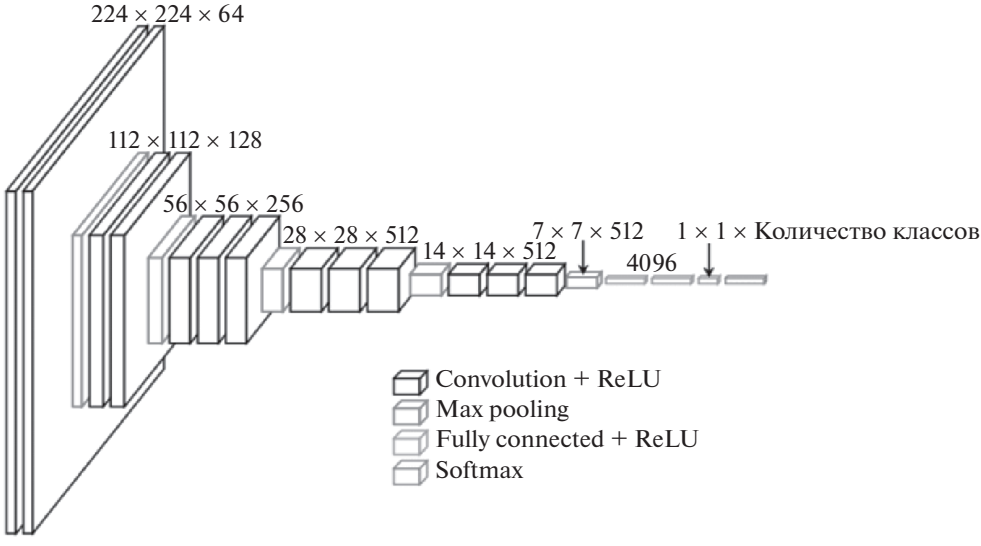

Рис. 2. Предварительно обученная сетевая модель VGG16.

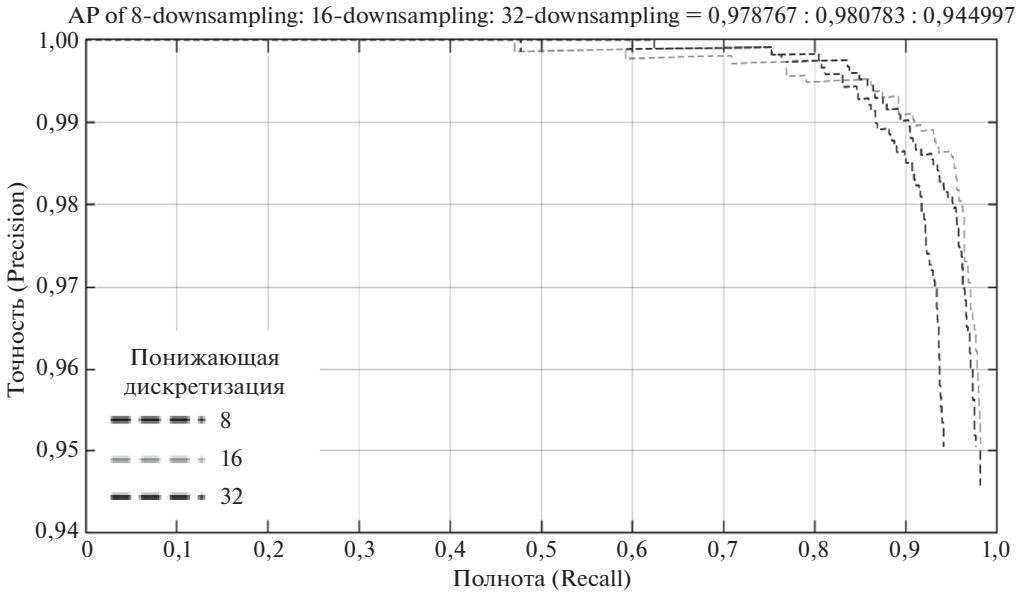

Рис. 3. Сравнение результатов по различным критериям понижающей дискретизацией.

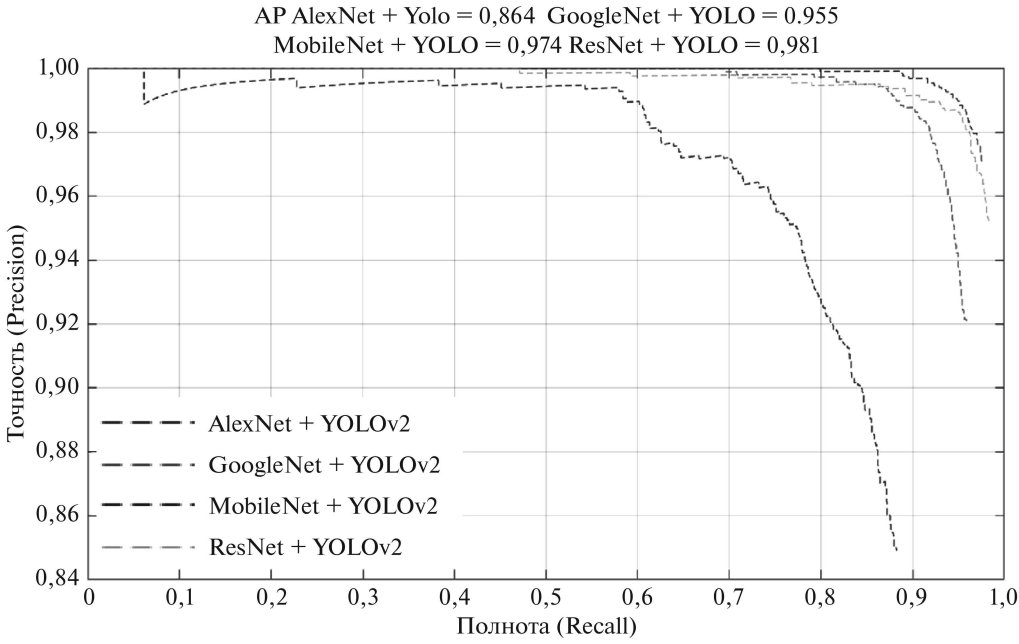

Рис. 4. Сравнение результатов на основе других предварительно обученных сетевых моделей.

Таблица 2. Сравнение результатов по различным критериям понижающей дискретизации

Понижающая дискретизация	Средняя точность
8-кратная	0,978%
16-кратная	0,980%
32-кратная	0,945%

Таблица 3. Сравнение результатов на основе других предварительно обученных сетевых моделей

Архитектура	Средняя точность набора тестовых изображений
Alexnet+YOLOv2	0,86%
Googlenet+YOLOv2	0,95%
Mobilenet+YOLOv2	0,97%
Resnet18+YOLOv2 (предлагаемая модель)	0,98%

Таблица 4. Точность распознавания лиц с использованием сети VGG 16

Сеть	База данных FEI	База данных Face94
VGG16	98%	98,5%

4. Результаты экспериментов

В табл. 2 и на рис. 3 представлены результаты экспериментов по обнаружению лиц сетями с различной кратностью понижающей дискретизации. Из таблицы видно, что 16-кратная понижающая дискретизация более эффективна для извлечения признаков лица: на тестовых данных была получена средняя точность 98%. В табл. 3 и на рис. 4 представлено сравнение результатов на основе других предварительно обученных сетевых моделей. На рис. 5 представлены примеры работы предлагаемой модели системы обнаружения лиц.

В табл. 4 представлены результаты исследования точности распознавания лица с использованием VGG16. Здесь использовались две различные базы данных (FEI и Face94) и получили точность распознавания более 98%. На рис. 6 представлены примеры работы полученного алгоритма поиска и распознавания лиц, реализованного в виде приложения MATLAB.

Для реализации системы, способной работать в режиме реального времени, был использован тулбокс App designer среды MATLAB. App designer позволяет инженерам и исследователям легко создавать профессиональные приложения, не требуя специализированных навыков программирования. App designer объединяет две основные задачи создания приложений: создание визуальных компонентов графического пользовательского интерфейса (GUI) и программирование поведения приложения. App designer также предоставляет

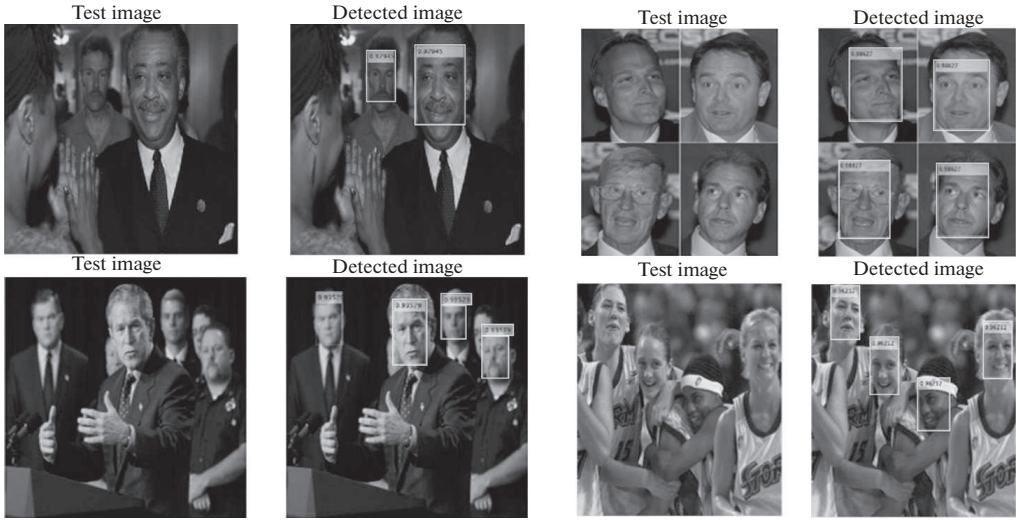

Рис. 5. Пример работы предлагаемой модели обнаружения лиц.

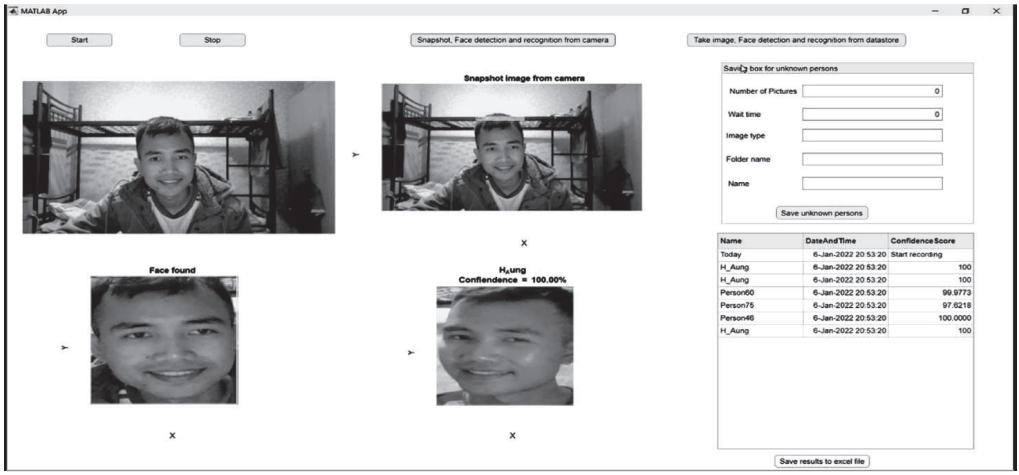

Рис. 6. Общий вид интерфейса системы распознавания лиц по видеоизображению.

различные библиотеки компонентов, такие как кнопки, флажки, оси пользовательского интерфейса, раскрывающиеся списки и т.д. Используя его, была создана система распознавания лиц, общий вид интерфейса которой показан на рис. 6.

5. Заключение

Эксперименты показали, что система обнаружения лиц на базе YOLOv2 обладает не только высокой точностью, но и позволяет находить лицо на видео в режиме реального времени, с высокой скоростью обнаружения. Для этапа распознавания лиц использовалась предварительно обученная сетевая

модель VGG16, перенесенная в MATLAB. Для 101 разной персоны и общего количества изображений 5050 получена точность более 98%. Это довольно хороший результат, достаточный для работы многих приложений захвата и распознавания изображения с видеокамеры в реальном времени.

СПИСОК ЛИТЕРАТУРЫ

1. *Viola P., Jones M.* Rapid object detection using a boosted cascade of simple features // Institute of Electrical and Electronics Engineers (IEEE), 15 April 2003, ISSN: 1063-6919, 9 pages, https://doi.org/10.1109/CVPR.2001.990517.
2. *Guennouni S., Ahaitouf A., Mansouri A.* Face Detection: Comparing Haar-like combined with Cascade Classifiers and Edge Orientation Matching // Institute of Electrical and Electronics Engineers (IEEE), 29 May 2017, ISBN:978-1-5090-6681-0, 4 pages, https://doi.org/10.1109/WITS.2017.7934604.
3. *Хачумов М.В., Нгуен Т.З.* Распознавание лиц по фотографиям на основе инвариантных моментов // Современные проблемы науки и образования. 2015. № 2-2, url: http://science-education.ru/ru/article/view?id=23235 (дата обращения: 23.05.2021).
4. *Рудинская Е.А., Парингер Р.А.* Разработка алгоритма детектирования лиц с использованием комбинаций каскадов Хаара // Сб. тр. ИТНТ-2019. Новая техника. 2019. С. 6–12.
5. *Redmon J., Farhadi A.* YOLO9000: Better, Faster, Stronger // Institute of Electrical and Electronics Engineers (IEEE), 09 November 2017, ISSN: 1063-6919, 9 pages, https://doi.org/10.1109/CVPR.2017.690.
6. *Simonyan K., Zisserman A.* Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556v6 [cs.CV] // 10 Apr 2015, 14 pages.
7. *Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun.* Deep Residual Learning for Image Recognition, arXiv:1512.03385v1 [cs.CV] // 10 Dec 2015.
8. *Коломиец В.* Анализ существующих подходов к распознаванию лиц [Электронный ресурс]. URL: http://habrahabr.ru/company/synesis/blog/ 238129/ (дата обращения: 15.12.2021).
9. *Redmon J., Santosh D., Girshick R., Farhadi A.* You only look once: Unified, real-time object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788. Las Vegas, NV: CVPR, 2016.
10. *Russakovsky O., Deng J., Su H., et al.* ImageNet Large Scale Visual Recognition Challenge // International Journal of Computer Vision (IJCV). 2015, Vol. 115, Issue 3, p. 211–252.
11. *Qawaqneh Z., Mallouh A.A., Barkana B.D.* Deep convolutional neural network for age estimation based on VGG-face model // arXiv preprint arXiv:1709.01664. – 2017.
12. *Schroff F., Kalenichenko D., Philbin J.* FaceNet: A unified embedding for face recognition and clustering // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2015.
13. *Taigman Y., Yang M., Ranzato M., Wolf L.* DeepFace: Closing the gap to human-level performance in face verification // Proc IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2014, p. 1701–1708.
14. *Amos B., Ludwiczuk B., Satyanarayanan M.* Openface: A general-purpose face recognition library with mobile applications // CMU School of Computer Science. 2016, Vol. 6. No. 2, p. 20.

15. *Felzenszwalb P.F., Girshick R.B., McAllester D., Ramanan D.* Object detection with discriminatively trained part-based models // IEEE Trans. Pattern Anal. Mach. Intell. 2010, Vol. 32. Issue 9, p. 1627–1645.
16. *Felzenszwalb P.F., Huttenlocher D.P.* Pictorial structures for object recognition // Int. J. Comput. Vision. 2005, Vol. 61. Issue 1, p. 55–79.
17. *Fischler M.A., Elschlager R.A.* The representation and matching of pictorial structures // IEEETrans. Comput. 1973, Vol. 22, Issue 1, p. 67–92.
18. *Zhu X., Ramanan D.* Face, detection pose estimation, and landmark localization in the wild // 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, p. 2879–2886.
19. *Viola P., Jones M.J.* Robust real-time face detection // Int. J. Comput. Vis. 2004, Vol. 57. Issue 2, p. 137–154.
20. *Li H., Lin Z., Brandt J., Shen X., Hua G.* Efficient boosted exemplar-based face detection // 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
21. *Shen X., Lin Z., Brandt J., Wu Y.* Detecting and aligning faces by image retrieval // 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, p. 3460–3467.
22. *Krizhevsky A., Sutskever I., Hinton G.E.* Imagenet classification with deep convolutional neural networks // Advances in Neural Information Processing Systems, 2012, p. 1097–1105.
23. *LeCun Y., Bottou L., Bengio Y., Haffner P.* Gradient-based learning applied to document recognition // Proc. IEEE. 1998, Vol. 86. Issue 11, p. 2278–2324.
24. *Girshick R., Donahue J., Darrell T., Malik J.* Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. https://doi.org/10.48550/arXiv.1311.2524.
25. *Zhang C., Zhang Z.* Improving multiview face detection with multi-task deep convolutional neural networks // 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2014, p. 1036–1041.

Статья представлена к публикации членом редколлегии А.А. Лазаревым.

Поступила в редакцию 17.02.2022

После доработки 22.04.2022

Принята к публикации 29.06.2022