FRI-ONLINE-1-CCT1-06

DEEP LEARNING BASED FACE RECOGNITION SYSTEM WITH SMART GLASSES[6]

Eng. Nikolay Gospodinov, PhD Student
Department of Computer Systems and Technologies,
"Angel Kanchev" University of Ruse
Phone: +359 88 834 5830
E-mail: ngospodinov@uni-ruse.bg

Prof. Georgi Krastev
Department of Computer Systems and Technologies,
"Angel Kanchev" University of Ruse
Phone: +359 82 888 672
E-mail: geork@uni-ruse.bg

Abstract: *A key element of a system that supports people with visual impairments using augmented reality glasses is the creation of a system for recognizing faces and objects. The system is created with a single purpose, and it is to help people with visual problems more easily perceive their environment and reduce the social distance between them and ordinary people. To be able to create such a system, it is necessary to create a model through which faces, and objects can be detected and recognized and through sound reproduction of their labels (names) they are transmitted to the user. The creation of such a model requires the use of artificial intelligence and in particular neural networks.*

Keywords: *Augmented reality glasses, Smart glasses, Visual Impairments, Artificial Intelligence, Neural networks.*

ВЪВЕДЕНИЕ

Deep learning (дълбоко машинно самообучение) е област, която се развива изключително бързо, успоредно с увеличаващите се изчислителни способности на модерните компютърни системи. Алгоритмите за Deep learning се справят успешно в ситуации, при които се изисква откриване на различни модели на поведение, взаимовръзки и зависимости между различни елементи в дадено множество от данни. С тяхна помощ се генерира код, който позволява разпознаването на тези модели в нови данни.

Като *добавена реалност* (Augmented reality, AR) може да се определи всеки дигитален обект или информация, който зрителя директно или индиректно наблюдава като добавен слой върху реалната физическа среда, която вижда. Разпространени са следните разновидности на *добавена реалност*:

- през телефон. За момента, е най-разпространения вариант. Това става с използването на маркер, който стартира AR изживяването/съдържанието. Може да стане и без маркер, като тогава сензорите на телефона разпознават околната среда и поставят дигиталния образ/обект спрямо желанието на потребителя. Друга възможност за активиране на съдържанието е чрез геолокации.
- с AR очила (Smart glasses), през който човек вижда насложена върху реалния свят дигитална информация. Известни производители на AR очила са: Google, ODG, Toshiba, Epson и др.
- Прожектирана *добавена реалност* са всички прожектирани върху реални физически обекти дигитални образи- този процес е по известен като projection mapping.

Напоследък популярност набира една нова концепция намираща се в етап на разработка, наречена „Augmented ID" – Самоличност чрез добавена реалност. Тази идея е в основата на предлаганото изследване. Чрез AR очила да се визуализира цифрова идентичност на хората, които срещаме в реалния живот. Разработката ни дава възможност да открием предварително

[6] Докладът е представен на заседание на секция 3.2 на 29 октомври 2021 с оригинално заглавие DEEP LEARNING BASED FACE RECOGNITION SYSTEM WITH SMART GLASSES

подбрана и подготвена информация за хората около нас. Потребителите могат да контролират и подбират съдържанието и информацията за себе си, връзките към профилите си в социалните мрежи (Facebook, Twitter), които искат да споделят с останалия свят и друга лична информация.

Настоящата работа представя възможно решение на проблема за класификация на изображения с очила за добавена реалност. Целта е да се определи към кой клас принадлежи дадено входно изображение. Начинът за постигане на това е създаването и тренирането на изкуствена невронна мрежа върху изображения, при което невронната мрежа трябва да определи към кой клас принадлежи даденото изображение (Saad Albawi, 2017).

ИЗЛОЖЕНИЕ

Интелигентната система за обектно и лицево разпознаване използваща добавена реалност се реализира чрез третото поколение очила за добавена реалност Moverio BT-300 (фиг. 1.). Бинокулярните очила Moverio BT-300 работят под управлението на операционна система Android 5.1, включват сензор за осветеност, 9-осен сензор за движение (акселерометър, жироскоп, магнитометър), поддържат WAV, MP3, AAC аудио формати и др. В таблица 1 са представени част от характеристиките на очилата (Moverio, 2017).

Таблица 1. Технически характеристики на Moverio BT-300

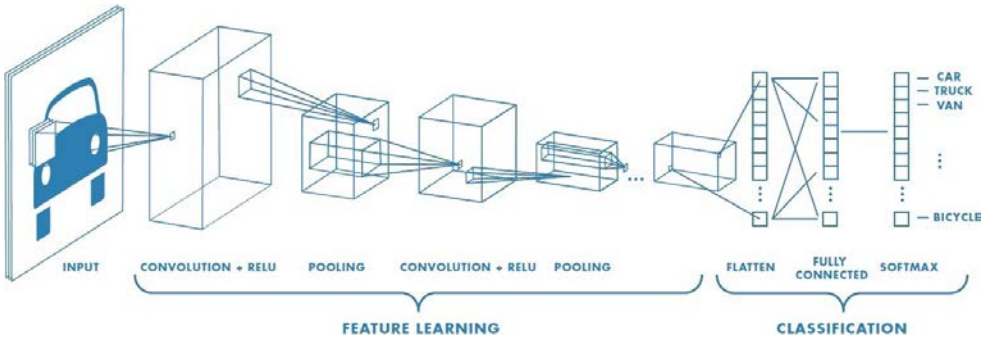
Model Number		BT-300
Material		Si-OLED
Panel size		0.43" wide panel (16:9)
Pixel count		Horizontal 1280 × Vertical 720 × RGB
Angle of view		Approximately 23 degrees (diagonally)
Virtual screen size		80" support (virtual viewing distance 5m)
Color reproduction		24 bit color (approximately 16,770,000 colors)
Platform		Android ™ 5.1
Supported movie formats		MP4(MPEG4/H.264+AAC), MPEG2(H.264+AAC), VP8
Supported still image formats		JPEG, PNG, BMP, GIF
Supported audio formats		WAV, MP3, AAC
Supported 3D		Side-by-side format
Internal memory	Main memory	2GB
	User memory	16GB
External memory		microSD (2 GB maximum),microSDHC (32 GB maximum)
Connector ports		micro-USB, controller, 4 pin mini jack (earphones with microphone complying with the CTIA standard)*1,microSD card slot

Operating temperature		5 to 35℃, humidity 20 to 80% (no condensation)
Storage temperature		-10 to 60℃, humidity 10 to 90% (no condensation)
Power voltage	Controller	5VDC (when USB connected)
Battery		Lithium polymer battery 2950 mAh
Frequency band		IEEE802.11a/b/g/n/ac、Miracast® (Source/Sink) Bluetooth® Ver.4.1 (Bluetooth® Smart Ready)
Frequency range		2.4 GHz band 1 to 13 ch, 5 GHz band 36 to 144 ch
Modulation		OFDM,DS-SS
Battery life		Approx. 6 hours (When playing movies)
Exterior dimensions (W×D×H)	Headset	191mm × 178mm × 25mm (without the shade)
	Controller	56mm × 116mm × 23mm (without protruding section)
Mass	Headset	69g (without cables)
	Controller	129g (without batteries, 1 battery is approx. 50g)
Camera		5 million pixels
Earphones with mic	Reception	100dB/mW
	Playback frequency band	20 to 20kHz
	Impedance	32ohm
Sensors	GPS/geomagnetic sensor/accelerometer/gyroscopic sensor/illumination sensor	

Очилата разполагат с видеокамера и микрофон, които позволяват потребителя да предава в реално време информацията около него със звук на други хора (Michalski,1985).

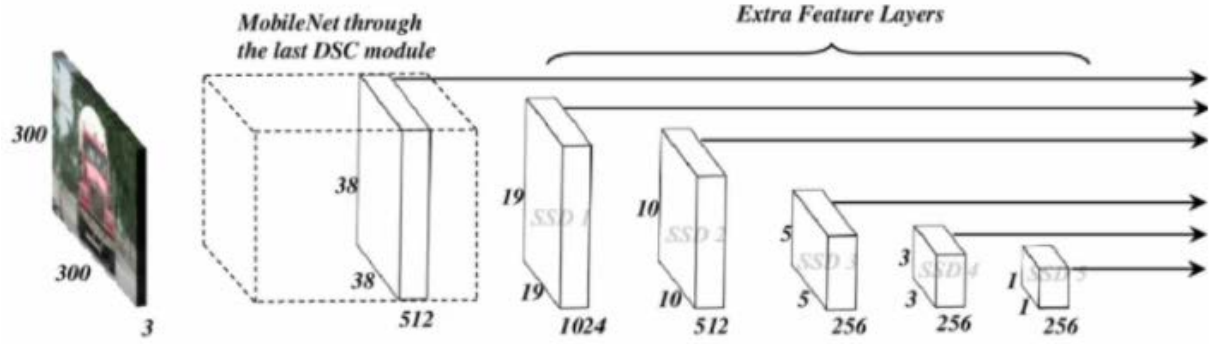
Съществуват многобройни алгоритми за лицево и обектно разпознаване. Един от тях е R-CNN, който е разработен върху дълбока невронна мрежа и произлизащите от него SPP-NET, Fast-RCN, Faster RCNN, and R-FCN. Всички те постигат високи резултати, но изискват много изчисления и много променливи, които не са подходящи за мобилни устройства (Ayesha Younis, 2020). За създаването на интелигентна система за обектно и лицево разпознаване е необходимо създаването на подходящ модел, чрез който да се разпознават необходимите обекти и лица. Моделът е изграден на принципите заложени при дълбокото самообучение, а именно използване на разновидност на конволюционна невронна мрежа (фиг. 2). Конволюционните невронни мрежи са част от „Feedforward" невронните мрежи. Те са вдъхновени от начина на действие на невроните от визуалната кора на мозъка. Всеки неврон от нея реагира само на определен регион от зрителното поле, наричано поле на възприятие. Полетата на възприятие на невроните се припокриват частично и така покриват цялостно зрителното поле.

Фиг. 1. Очила с добавена реалност Epson Moverio BT-300



Фиг. 2. Конволюционна невронна мрежа

За разлика от напълно свързаната невронна мрежа, при конволюционната невронна мрежа се добавя нов слой – Конволюционнен слой. Както е показано на фиг. 2., конволюционната невронна мрежа е разделена на две части – базова част (изучаване на черти на изображение) и класификационна част (откриване и класифициране на изображение). Първата част е съставена от три основни слоя, които се повтарят през цялата мрежа: конволюционнен, активационен и пулингов слой. Втората част е съставена от напълно свързана невронна мрежа, която завършва с Softmax функция, която връща предполагаема стойност на разпознатото изображение. За първата част на конволюционната невронна мрежа е използвана MobileNet архитектура (фиг. 3).

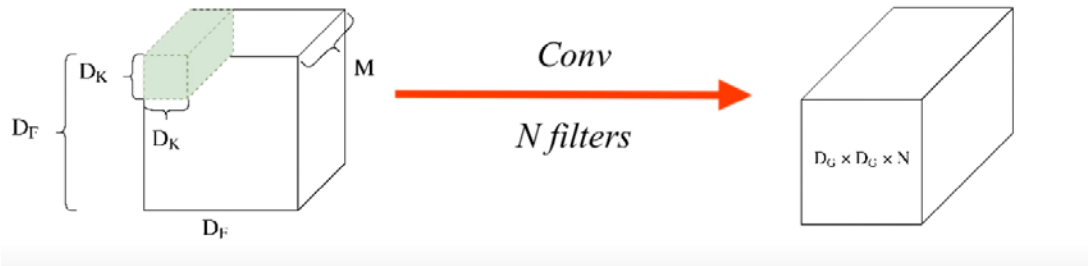


Фиг. 3. MobileNet архитектура

MobileNet архитектурата е специално създадена за приложения, които разпознават обекти в реално време и работят върху устройства с ограничен хардуер (мобилни телефони, видео камери, очила с добавена реалност). Тази архитектура не е толкова прецизна в разпознаването и засичането на обекти в сравнени с други архитектури, които постигат това използвайки повече време за обработка (Simonyan & Zisserman, 2014). Използвана е с цел

времето за разпознаване да бъде оптимизирано (Howard et al., 2017). MobileNet е лека архитектура на дълбока невронна мрежа, която е предназначена за мобилни телефони и приложения за вградено зрение. Тя е изградена върху дълбочинни разделителни конволюции.

При стандартната конволюция (СТ) филтър К с размери (D_k, D_k, M) се прилага върху цялата дълбочина на входен елемент F с размери (D_F, D_F, M). След прилагане на стандартна конволюция се получава изходен елемент G с размери (D_G , D_G , N) (фиг. 4). За извършването на една конволюция са необходими $D_K^2 \; X \; M$ брой умножения. Тъй като конволюцията се прилага по цялата височина и широчина на входния елемент, са необходими $D_G^2 \; X \; D_k^2 \; X \; M$ брой умножения за един филтър. При прилагане на N на брой филтри са необходими - $N \; X \; D_G^2 \; X \; D_k^2 \; X \; M$ брой умножения.

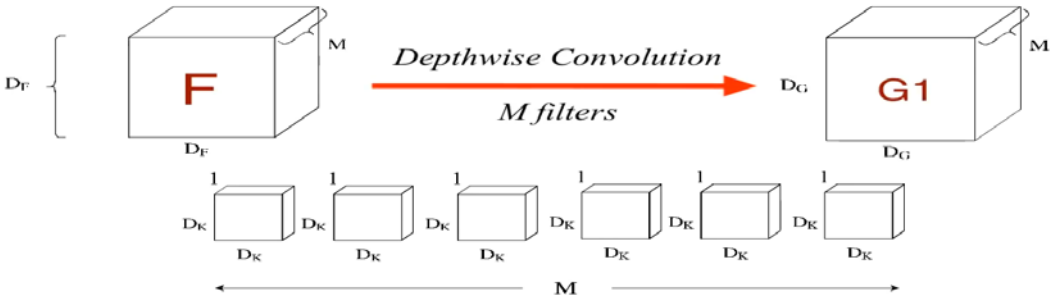


Фиг. 4. Прилагане на стандартна конволюция

Дълбочинно разделителната конволюция (ДРК) е изградена от два етапа. На първия етап се прилага конволюция в дълбочина (Depthwise Convolution), а на втория етап се прилага Точкова конволюция (Pointwise Convolution).

Конволюция в дълбочина

При конволюция в дълбочина, филтър К с размери (D_k , D_k , M) се прилага върху всеки слой по отделно на входен елемент F с размери (D_F , D_F, M). След прилагане на конволюция в дълбочина се получава изходен елемент G с размери (D_G , D_G , M) (фиг. 5). За извършването на една конволюция са необходими D_K^2 брой умножения. Тъй като конволюцията се прилага по цялата височина и широчина на входния елемент, са необходими $D_G^2 \; X \; D_k^2$ брой умножения за един филтър. При прилагане на M на брой филтри са необходими - $D_G^2 \; X \; D_k^2 \; X \; M$ брой умножения.



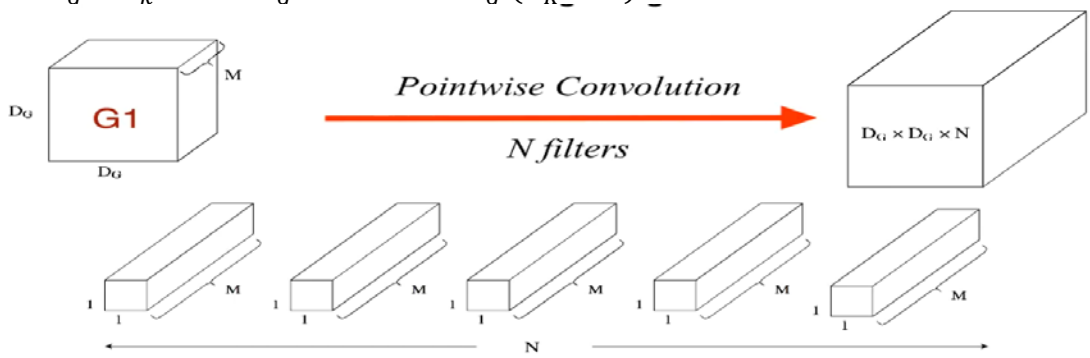
Фиг. 5. Конволюция в дълбочина

Точкова конволюция

При точкова конволюция, филтър с размер (1, 1, M) се прилага върху входен елемент G с размери (D_G , D_G , M) и се получава изходен елемент G с размери (D_G , D_G , N) (фиг. 6).

За извършването на една конволюция са необходими M брой умножения. Тъй като конволюцията се прилага по цялата височина и широчина на входния елемент, са необходими $D_G^2 \; X$ M брой умножения за един филтър. При прилагане на N на брой филтри са необходими - $N \; X \; D_G^2 \; X \; M$ брой умножения (Saeed Arabli, 2019). Общия брой умножения от двата етапа е:

$$M\ X\ D_G^2\ X\ D_k^2\ + N\ X\ D_G^2\ X\ M\ = M\ X\ D_G^2\ (D_K^2 + N)$$



Фиг. 6. Точкова конволюция

След сравняване на стандартната конволюция с дълбочинно разделителната конволюция се установява, че при стандартната конволюция се използват 9 пъти повече обучителни елементи и 9 пъти повече брой умножения. Това от своя страна е признак, че дълбочинно разделителната конволюция е по-бърза:

$$\frac{\text{Брой умножения при ДРК}}{\text{Брой умножения при СК}} = \frac{\mathrm{M}\ X\ D_G^2 (D_K^2\ + N)}{N\ X\ D_G^2\ X\ D_K^2\ X\ M}$$

$$\frac{\text{Брой умножения при ДРК}}{\text{Брой умножения при СК}} = \frac{D_K^2\ + N}{(D_K^2\ X\ N)} = \frac{1}{N} + \frac{1}{D_K^2}$$

Пример: Нека имаме изходен елемент с размер N = 1024 и филтър $D_K = 3$

$$\frac{\text{Брой умножения при ДРК}}{\text{Брой умножения при СК}} = \frac{D_K^2\ + N}{(D_K^2\ X\ N)} = \frac{1}{1024} + \frac{1}{3^2} = 0.122$$

Където:

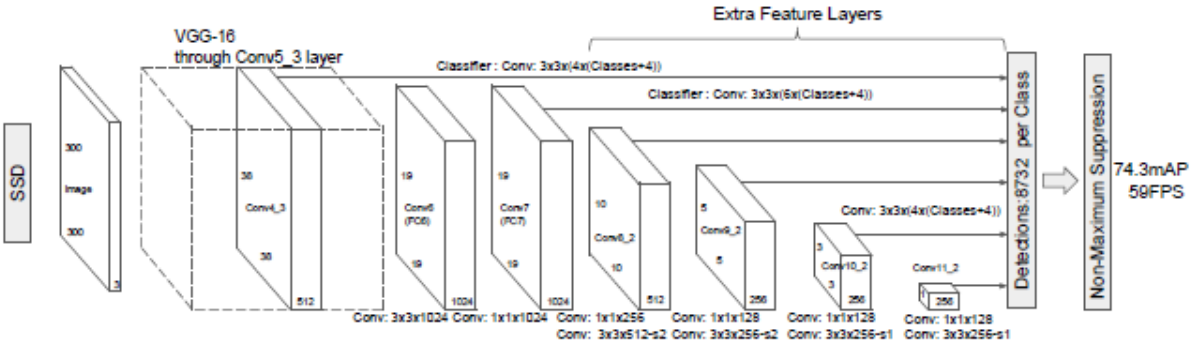
- ДРК - Дълбочинно разделителна конволюция.
- СК – Стандартна конволюция.

В таблица 2 е представена сравнителна характеристика на MobileNet архитектура използваща стандартна конволюция и MobileNet архитектура, използваща дълбочинно разделителна конволюция.

Таблица 2. Сравнителна характеристика на Con MobileNet и MobileNet

Модел	Точност при обработка на изображенията	Брой умножения (милиони)
Conv MobileNet	71.7 %	4866
MobileNet	70.6 %	569

За втората част на конволюционната невронна мрежа е използвана SSD (Single Shot Detection) архитектура (фиг. 7).

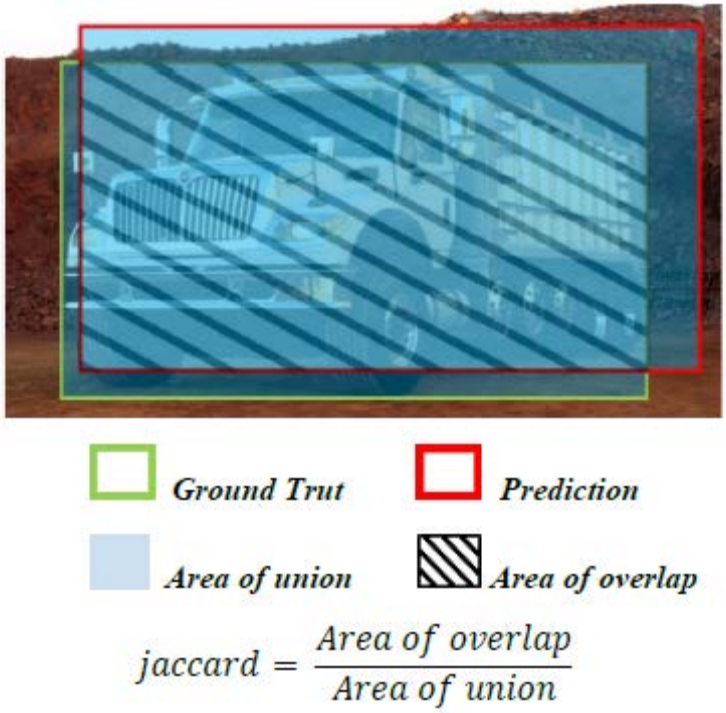


Фиг. 7. SSD Архитектура

Съществуват два типа детектори на обекти – детектор на обекти, които работят с една стъпка и детектори на обекти, които работят на две стъпки. При детекторите, които работят на две стъпки, процеса на локализация и класификация са разделени, което води до по-добра точност, но изискват много изчисления, което не е подходящо за мобилни устройства. (Girshick, 2015; Girshick et al., 2013; Ren et al., 2015). За разлика от тях, детекторите, които работят с една стъпка , съчетават процеса на локализация и класификация в едно.

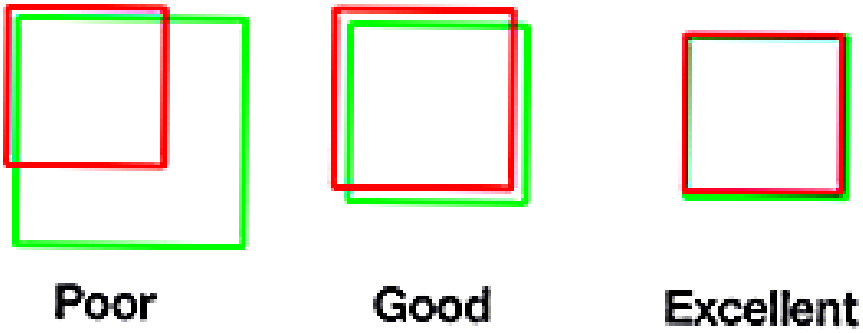
SSD използва VGG-16 (Visual Geometry Group) като базова мрежа. Използваме го заради силното му представяне при класифицирането на изображения с високо качество. При него, за разлика от VGG, където имаме напълно свързани слоеве, тук имаме добавени помощни конволюционни слоеве (Започват от Conv6). Това позволява извличане на функции в множество мащаби и постепенно намаляване на размера на входа за всеки следващ слой. При SSD всяка клетка на карта с характеристики е свързана с набор от рамки за ограничение по подразбиране с различни размери и съотношения. Това позволява на SSD да обобщава за всякакъв вид входни данни, без да изисква фаза на предварително обучение за предишно поколение.

За изчисляване на процента на препокриване между предполагаемата позиция и реалната позиция на обект се използва коефициент на подобие на Джакар (фиг. 8).



Фиг. 8. Коефициент на препокриване на Джакар

В оптималния случай предполагаемата зона напълно ще се препокрива с реалното положение на обекта в пространството (фиг. 9).



Фиг. 9. Препокриване на предполагаемата позиция и реалната позиция на обекта

В обобщение този модел е използван поради факта, че въпреки, че модела има високо бързодействие, той е достатъчно ефективен. MobileNet архитектурата позволява използването на метода в устройства с ограничена хардуерна възможност, а именно очила с добавена реалност.

Създаването на такъв модел се разделя на няколко етапа. За реализирането е използван програмния език Python и средата за разработване Jupyter Notebook. В добавка към тях са използвани и допълнителни библиотеки, като основната библиотека ползване при реализирането му е TensorFlow. Етапите на разработка са 4:

1. Подготовка на набор от изображения

На този етап трябва да се подготви набора от данни. В случая той представлява набор от изображения, за всеки обект и лице, което модела ще разпознава. За тази цел се използва подходящ за тази работа скрипт на Python, с помощта, на който, чрез камерата на компютър се събират необходимите изображения. След събиране на необходимия брой изображения се преминава към етикиране на изображенията. Етикирането е процес, при който се създава придружаващ xml файл (анотация) към всяко изображение, съдържащ информация за името на етикета, размера на изображението и др.

2. Трениране на модела MobileNet-SSD

След разделяне на изображенията и придружаващите ги анотации в две директории (тестова и обучителна) се преминава към процеса на трениране на модела. За създаването на модел, отговарящ на изискванията на системата за лицево и обектно разпознаване, се използва MobileNet-SSD модела.

Чрез Python скрипт, в няколко стъпки се извършва обучението на модела:

- Изтегляне на MobileNetSSD модела, който е предварително трениран с помощта на COCO набор от данни (изображения).
- Създаване на карта на етикетите (име и уникален номер) в обособена за това директория
- Създаване на TF записи – двоичен файлов формат за съхранение на данни. Използва се за намаляване на времето при трениране на модела в обособена за това директория
- Модифициране на модела да използва нашите данни
- Трениране (може да го трениране използвайки процесора (CPU) или графичната кара (GPU).

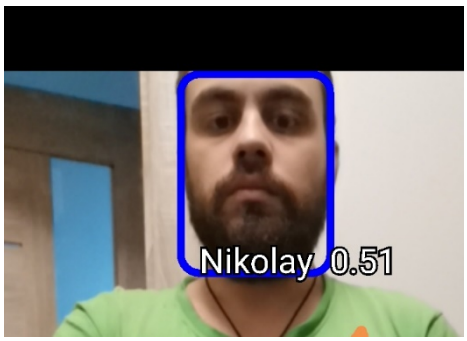
За да е възможно използването на GPU при графична карта NVIDIA е необходимо инсталиране на CUDA и cuDNN.

3. Конвертиране на тренирания модел в подходящ формат

След като модела премине успешно етапа на обучение, се преминава в етап на конвертиране във формат, който позволява имплементирането му специализирания модул за разпознаване на лица и обекти на интелигентната система, предназначена за хора със зрителни проблеми, която използва очила с добавена реалност. Този формат е Tensorflow Lite. Той може да се използва при разработване на мобилни приложения, използващи машинно самообучение и задълбочено самообучение и са използват операционни системи Android и iOS.

4. Имплементиране на форматирания модел в софтуера, за очила с добавена реалност

След конвертирането на модела той преминава към имплементиране в специализирания модул за разпознаване на лица и обекти на интелигентната система, използваща очила с добавена реалност. За да е възможно това трябва да се използват и допълнителни библиотеки (OpenCV, Tensorflow и др.). На фиг. 10, използвайки MobileNet-SSD, е осъществено засичане и разпознаване на лице на човек.



Фиг. 10. Разпознато лице с MobileNet-SSD

ЗАКЛЮЧЕНИЕ

Използването на метода на дълбокото самообучение позволява създаване на система за лицево разпознаване, която удволетворява изискванията на потребителя, а именно точност и бързодействие при засичане и разпознаване на лицата на хората около него. Използваната архитектура, не само покрива изискванията за бързина на работа, но и хардуерните изисквания на системата (очила с добавена реалност). Разработката е първа стъпка при създаването на интелигентна система за откриване на предварително подбрана и подготвена информация за хората около нас. Тя е в основата на бъдеще в което хората със зрителни проблеми ще могат безпрепятствено да комуникират със всички около тях и да водят пълноценен начин на живот.

БЛАГОДАРНОСТ

Този доклад е реализиран с подкрепата на Националната научна програма „Информационни и комуникационни технологии за единен цифров пазар в науката, образованието и сигурността (ИКТ в НОС)“, финансирана от Министерството на образованието и науката на България.

REFERENCES

Saad Albawi, Tareq Abed Mohammed, Saad Al-Zawi (2017), Understanding of a convolutional neural network, 2017 International Conference on Engineering and Technology (ICET), DOI: 10.1109/ICEngTechnol.2017.8308186.

Ayesha Younis, Li Shixin, Shelembi JN, Zhang Hai, (2020), Real-Time Object Detection Using Pre-Trained Deep, Tianjin University of Technology and Education

Simonyan, K., & Zisserman, A., 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & Andreetto, M. ,2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

Saeed Arabi, Arya Ketabchi Haghighat, A deep learning based solution for construction equipment detection: from development to deployment, 2019, Department of Civil, Construction and Environmental Engineering Iowa State University, Ames, IA

Girshick, R., 2015. Fast R-CNN.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. ,2013. Rich feature hierarchies for accurate object detection and semantic segmentation Tech report (v5).

Epson Moverio - https://moverio.epson.com/2017/13/05

Michalski, R.S.; Carbonell, J.G.; Mitchell, T.M. Machine learning: An artificial intelligence approach. Artif. Intell. 1985, 25, 236–238.

Kress, B.; Starner, T. A Review of Head-Mounted Displays (HMD) Technologies and Applications for Consumer Electronics. In Photonic Applications for Aerospace, Commercial, and Harsh Environments IV; International Society for Optics and Photonics: Bellingham, WA, USA, 2013; Volume 8720, p. 87200A

Seifabadi, R.; Li, M.; Long, D.; Xu, S.; Wood, B.J. Accuracy Study of Smartglasses/Smartphone AR Systems for Percutaneous Needle Interventions. In Proceedings of the SPIE Medical Imaging, Houston, TX, USA, 16 March 2020; Volume 11315.