UDC 004.93

DEVELOPING A COMPUTER VISION RE-IDENTIFICATION SYSTEM

[1]Maksym S. Ostapenko, [1]Olena S. Shtogrina, [1]Larysa S. Globa,
[1]Andrii A. Astrakhantsev, [2]Siemens Eduard

[1] Institute of Telecommunication Systems
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine
[2]Anhalt University of Applied Sciences, Germany

Background. The rapid growth of computational power of machines and amount of data caused exploding usage of computer vision in a large variety of tasks and in particular for people recognition.

Objective. The aim of the paper is to propose a computer vision re-identification system based on research. Also improvements for detection and recognition models of the system are made.

Methods. We used classical computer vision and deep learning techniques to create the system.

Results. The main contribution of the research is a description of the optimal system structure with a trade-of between speed and quality. Furthermore, requirements for an environment are proposed, which allows to set up the system in the real world with guaranteed quality.

Conclusions. Real-time computer vision re-identification system was developed and can be used in a production environment which satisfy requirements.

Keywords: machine learning**;** video surveillance**;** re-identification**;** computer vision**;** deep learning.

I. Introduction

Computer vision-based people re-identification systems are technologies that can identify or verify a person from an image or from a video stream. There are several main classes of re-identification systems, in general, they are all based on a comparison of selected features or generated vectors of face descriptors with images or descriptors that are already in the database. Also, re-identification systems are described as a biometric application based on artificial intelligence, they can uniquely identify a person by analyzing features based on personal textures and face sizes.

Computer vision face re-identification system is often used in access control systems and can be compared to other biometric protocols such as fingerprint recognition and retinal scanning. The recognition accuracy of the computer vision system is less due to the more accessible possibility of deception of the system and also the system is sensitive to environmental conditions.

The paper has the next structure:

- In Section 2 problems are set up.
- In Section 3 are described requirements to the system and the system will be presented.
- In Section 4 there is a discussion about key parts of the system and their selection.
- In Section 5 are described improvements to detection and recognition models.
- Finally, in the end, there is a conclusion about the results and further work.

II. Problem Definition

The problems that need to be solved with a proposed computer vision re-identification system:

1. Requirements to an environment where the system would be used.
2. Structure of the system.
3. Choosing optimal system parts in order to make a real-time system.
4. Increasing robustness of detection and recognition models.

III. People Re-Identification system Based on Computer Vision

Any computer vision-based system critically depends on environmental conditions, hence for the first we are going to concentrate on them then the system will be described.

A. Requirements for the system

Computer vision systems depend on environmental factors, among which the main are camera position, background, and light.

The quality of face recognition depends on camera position and how static camera is. The camera should be placed in a way that it will be possible to highlight the main characters of the object on an output image. On top of that, image resolution and bounding box of the object to the whole image is important as well. Thus, it will be possible to highlight the main face features. It is critical for recognition, as the main goal of embedding models is providing difference for different classes objects and dimensional similarity for objects within the same class. For preventing small deviation of the object and camera, face stabilization on the key points could be used.

The background of the environment can affect both positively and negatively, so it must be taken into account and if possible, to modify for maximizing efficiency. The negative effect of background on the quality of detection is in

increasing the number of false detections due to the high heterogeneity of the background and a large number of different textures. Reflecting surfaces should also be avoided as they cause the target objects to be detected several times and simultaneously. The positive impact of background on detection and recognition occurs when the background is homogeneous and has a high level of contrast to the possible targets.

When considering the room lighting setting, the following parameters must be considered:

1. The dimensional position of the light source. The light source has to be installed to avoid lighting and shadows on the target.
2. The intensity of light. It should be sufficient to show all the details on the possible targets.
3. Scattering of light. Has to be evenly scattered throughout the room.

Failure to take these factors into account will reduce the amount of visible detail of the object and, as a result, generated embedding would have the poor representational capacity, which in turn reduces the quality of recognition.

With slight deviations from the desired illumination values, digital image processing techniques are used to improve the image quality and highlight the target. Often used in the alignment of the histogram image to increase its contrast, variation of Gaussian filters to suppress noise and highlight the contours of objects. One of the most effective is the bilateral filter.

The main requirement for the camera is to provide sufficient resolution of the image, its contrast and frame rate per second.

Computing resources should be selected depending on what frame rate you can provide.

B. The system

The main idea behind a computer vision re-identification system is to describe a unique person with a unique embedding and make an identification decision by comparing the embedding. In general, the system consists of the following parts: detector, tracking model, object descriptor retrieval model, data-based search model, and authorization decision model. Fig. 1 shows the structure of the system.

Separate images from camera stream are fed to a detection model. It detects faces in the input images and output consists of the coordinates of the bounding boxes around the face.

Then, for each detected face, the trajectory of motion is initialized by the tracking algorithm. Each image of each trajectory is saved. If there are more than 5 frames (hyper parameter) for a single face, then all trajectory images are then fed to the model for obtaining unique face embedding.

The embedding generation model generates a unique 512 descriptor vector (hyper parameter) for each face image input. Each of the embedding received is queried to find five close

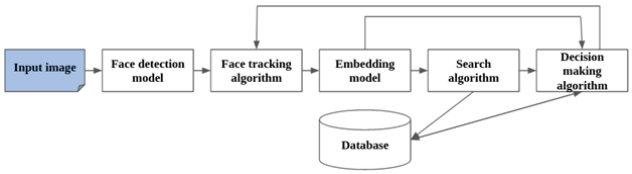

Fig.1. The structure of the proposed computer vision-based re-identification system.

neighbors (hyper parameter) from the database. After that, each input embedding is compared with the corresponding one from the database as a result of these comparisons, for each embedding the class to which it belongs. The initial prediction of a class is obtained by selecting the class which is most often found among the predictions for each input embedding if all classes are different, then the trajectory is assigned to the class closest to each of the input embedding, but if the distance is less than the threshold then they are initialized, new class.

The system then calls the corresponding trajectory the name of the resulting class for it, and continues to track the object between frames, but no longer submits its image to the face recognition model.

IV. Choosing Algorithms for the System Parts

A. Comparison of detection algorithms

YOLO [3] and SSD [4] detectors are similar in structure of prediction blocks and differ only in the base network and the number of prediction blocks. YOLO detector has a single block for object detection, which provides speed but loses quality when compared to SSD. SSD has several extra layers to refine the detection of objects with different sizes, followed by an aggregation block of all the predictions produces the final predictions.

As a result of the proposed improvements to the SSD detection algorithm (described below), we have obtained a detector that is fast enough and has a sufficiently high quality of detection that satisfies the optimality criteria, so it will be used in the proposed system.

B. Comparison of tracking algorithms.

Comparison of tracking algorithms with MOTA (Multiple object tracking accuracy) quality metrics is given in Table 2.

It is advisable to use a tracker based on correlation filters in an environment with and in the absence of an object detector. Also, based on correlation filters, the tracker is more valuable relative to the computing resource than the IOU tracker [5]. If the frame rate is low, the IOU between the detection of one object in the adjacent frames becomes low and the quality of the tracking is significantly reduced. Because the object background is unchanged and the detector selected is fast enough and has high detection quality, the system will use an IOU tracker.

C. Comparison of face recognition algorithms

After the comparison, the ArcFace [8] model was chosen as the architecture for the system. The rapid convergence of the model optimizing the error function was taken into consideration. The one does not require a complicated

procedure of generating batches within one epoch of training while using the ResNet50 [1] model achieves the highest results on standardized datasets (Table 3). Improvements will be made to increase the representativeness of the model.

D. Comparison of search algorithms

The naive k-nearest neighbor method is the most accurate, its results are standard, but as the database grows it becomes impossible to use because of the large amount of time spent for computing the distances between the input embedding with all embedding from a database. Hence, it was decided to take an approximate k-nearest neighbor method. HNSW [9] was chosen. The main advantages of the HNSW algorithm are speed, a high approximation to the results of the k-nearest neighbors method, and effective implementation in the nmslib library.

E. Decision-making algorithm

Distance threshold comparisons between embedding will be used to make authorization decisions in the system. As a distance metric, the cosine distance between vectors will be used, because the ArcFace error function operates with cosine distance during training.

TABLE I. COMPARISON OF DETECTION ALGORITHMS

Name	*FPS (frame per second)*	*mAP*	*Dataset*
YOLO	40	48.1	COCO test-dev
SSD	8	50.4	COCO test-dev
Improved SSD	15	54.1	COCO test-dev

TABLE II. COMPARISON OF TRACKING ALGORITHMS

Name	*FPS (frame per second)*	*MOTA*	*Dataset*
IoU Tracker	100 000	76.5	DETRAC
Background aware CF [6]	156	77.8	DETRAC

TABLE III. COMPARISON OF RECOGNITION ALGORITHMS

Name	*Accuracy (identification)*	*Accuracy (verification)*	*Dataset*
Center [7]	65.49	80.14	MEGA FACE
ArcFace	81.72	96.98	MEGA FACE
Improved ArcFace	84.16	95.34	MEGA FACE

V. DESCRIPTION OF IMPROVEMENTS TO THE DETECTION AND RECOGNITION MODELS

The key parts that most affect the quality of the re-identification system are the detection model and recognition models. The detection model depends on whether each individual target is detected in the video stream for further processing. The recognition model depends on the uniqueness of the representation of individual objects in the multidimensional space of embedding, and further the ability to distinguish between objects of separate classes. Therefore, it was decided to improve and optimize existing detection models and to obtain descriptors.

A. Description of improvements to SSD detection model

The YOLO and SSD detection models described here have some important ideas for improving the quality of detection and its speed.

The YOLO and SSD detectors are similar in structure to the prediction blocks but differ in the base network and the number of prediction blocks. The YOLO detector has one extra layer to predict the position of objects in the image, thus providing speed but losing quality compared to the SSD. The SSD has several extra layers to improve detection of objects of different sizes, followed by an extra layer followed by a block that aggregates the prediction of all extra layers and produces the final predictions. This complexity results in a lower detection rate than YOLO.

The purpose of improving the SSD detection algorithm is to improve the quality of its operation without much loss in detection speed.

SSD and MobileNet:

Replacing the VGG16 network with the MobileNet [2] network will reduce the number of network parameters and increase the speed. Improved SSD uses the MobileNet network as the base neural network. Predictive outputs that relate to the base network join the eleventh layer and the thirteenth are also added eight extra layers to the network after the thirteenth layer. Prediction is made from each extra layer.

CBAM (Constitution block attention module) module for increasing sensitivity to details:

To improve the quality of detection and reduce the number of negative detections in each extra layer was added an attention block - CBAM [10]. The CBAM module is used to increase the representative power of convolution neural networks. It does not require a lot of calculations, hence it can be effectively used without much loss in terms of speed. The channel attention module uses the inter-channel interaction of features and tries to highlight more important ones with greater weight. The channel attention module can be considered a detector of important attributes; it concentrates on what is important.

The spatial attention module takes into account the spatial interaction of features. The spatial attention module focuses on the position of the features.

The composition of the two modules of attention allows distinguishing channel and spatial information from the input data.

Training details:

Initially, an SSD detector with the MobileNet core network was taken. It was pretrained on the WIDER Face training

dataset [11]. After adding CBAM attention blocks, all layers up to and including ten were frozen and did not participate in training to optimize the loss function to avoid the problem of over fitting. For training and testing, WIDER Face datasets were used.

The training was performed over 35,000 iterations with the batch size of 16. The initial training speed was set to 0.01 and changed each iteration. The Adam [12] optimization algorithm was taken for the model optimization, with the following values of momentum and weight decay 0.9 and 0.0001, respectively. Augmentation of the training dataset was used to increase the number of training samples and reduce the likelihood of over fitting. The following augmentation techniques were used: random reflection, random scaling with a scaling factor of 0.2 to 2, random rotation between -30 and 30 degrees and Gaussian noise.

Thus, by changing the base model from VGG16 to MobileNet, we get an increase in speed and accuracy due to the greater representation power of the MobileNet network. By adding the CBAM special attention module, we have reduced the number of false detections. The results of the training are shown in Table 1.

B. Description of the recognition model and its improvements

After review and analysis of face recognition models, it was decided to use ArcFace based models as a face recognition architecture. The disadvantage of training models with the ArcFace loss function is the sensitivity to outliers in the training data, so the training dataset must be validated.

The goal of the ArcFace error-based model modifications is to increase the model's training speed, increase the model's representative ability, and reduce sensitivity to data anomalies.

Description of the backbone network:

The SE-ResNet network will be used as the base neural network to improve the representation power of network and convergence speed of model training. Recent studies in the field of CNN have shown that their representational capacity can be enhanced by integrating specialized learning mechanisms that help to pay attention to the features and parameters of the network itself.

The authors of the article Squeeze-and-Excitation Networks [13] propose a new unit to CNN - the Squeeze-and-Excitation (SE) block, it performs the task of nonlinear interaction between channels of one layer. The unit adaptively calibrates the inter-channel interaction and their responses to the input data, this unit models the interdependence of the channels. As a result, the SE unit learns to focus on important features while still compressing less important ones.

The block is common, and it performs different roles depending on the layer of the neural network on which it is located. In the early layers, the block draws attention to the features of ignoring image classes, enhancing the low-level representativity of the network. In the last SE layers, the blocks become more dependent on the input classes: each class has its own response. By installing SE blocks on all layers, weighted features can be accumulated across the entire network.

The advantage of SE block over similar blocks is the simple ability to integrate into any network, reduce the tendency to over fitting, increase the representative capacity of the network and use a small number of computing resources compared to the whole network.

Training details:

CASIA datasets [14], VGGFace2 [15], MegaFace [16] and LFW [17] were used to train the network.

The SE-ResNet-50 network is used to obtain the descriptors. The output vector size of the descriptors is 512. The feature vector is obtained by passing the output data from the last convolutional layer to the batch normalization layer and then to the fully connected neuron layer and outputting the final vector of the input descriptor output.

The scaling vector of the descriptors is used by the scaling factor s, it is equal to 64, and the angular indentation parameter of the ArcFace m error function is 0.5, as was shown in the original article. The batch size was set to 64.

The initial training speed, as for the detector, was set to 0.01 and changed every iteration. The training process continued for 150,000 iterations. For optimization, Adam was taken with the following values of momentum and weight decay 0.9 and 0.0005, respectively. Augmentation of the training dataset was used to increase the training sample and reduce the likelihood of over fitting. The following augmentation techniques were used: random reflection, random scaling with a scaling factor of 0.2 to 2, random rotation between -30 and 30 degrees and Gaussian noise.

Thus, setting the base model SE-ResNet-50, we get a slight drop in speed, but a significant improvement in quality due to the greater representation power due to SE blocks. The results of the training are shown in Table 3.

Conclusion

In the paper was proposed a new computer vision-based re-identification system. In the research the next contributions were made:

1. A thorough analysis and comparison of the components of a computer vision system of people re-identification: detection, tracking, recognition, search, decision-making. As a result of the analysis, the existing shortcomings of the algorithms were identified and the possibilities of their improvement to increase their quality of work.

2. Improvements to detection and facial recognition models have been done. For the detection model was added MobileNet as base network and prediction blocks ware combined with CBAM attention block.
3. Improved speed and quality of people re-identification by developing a computer vision-based re-identification system that allows execution in real-time and with guaranteed quality. As a result, the quality of detection was increased to 54.1 mAP and identification accuracy of the recognition model ware increased to 84.16%.
4. The system could be used for re-identification task in environments where it is possible to satisfy the environmental requirements, for example in offices or at border.

Further research will include system testing with new state of the art parts and in a real-world environment with challenging conditions.

REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun "Deep 1Residual Learning for Image Recognition". arXiv:1512.03385, 2015.

[2] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". arXiv:1704.04861, 2017.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi "You Only Look Once: Unified, Real-Time Object Detection". arXiv:1506.02640, 2015.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg "SSD: Single Shot MultiBox Detector". arXiv:1512.02325, 2016.

[5] E. Bochinski, V. Eiselein, T. Sikora. High-Speed Tracking-by-Detection Without Using Image Information. In International Workshop on Traffic and Street Surveillance for Safety and Security at IEEE AVSS 2017, 2017.

[6] Hamed Kiani Galoogahi, Ashton Fagg, Simon Lucey "Learning Background-Aware Correlation Filters for Visual Tracking". arXiv:1703.04590, 2017.

[7] Yandong Wen, Kaipeng Zhang, Zhifeng Li and Yu Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition. The Chinese University of Hong Kong, Sha Tin, Hong Kong, 2016.

[8] Jiankang Deng, Jia Guo, Niannan Xue, Stefanos Zafeiriou "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" arXiv:1801.07698, 2018.

[9] Yu. A. Malkov, D. A. Yashunin "Efficient and robust approximate nearest neighbour search using Hierarchical Navigable Small World graphs". arXiv:1603.09320, 2016.

[10] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon "CBAM: Convolutional Block Attention Module". arXiv:1807.06521, 2018.

[11] Yang, Shuo and Luo, Ping and Loy, Chen Change and Tang, Xiaoou. WIDER FACE: A Face Detection Benchmark. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[12] Diederik P. Kingma, Jimmy Ba "Adam: A Method for Stochastic Optimization". arXiv:1412.6980, 2014.

[13] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, Enhua Wu "Squeeze-and-Excitation Networks" arXiv:1709.01507, 2017.

[14] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch.arXiv:1411.7923, 2014.

[15] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. InFG, 2018.

[16] Ira Kemelmacher-Shlizerman, Steve Seitz, Daniel Miller, Evan Brossard "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale". arXiv:1512.00596, 2015.

[17] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labelled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report,2007.

Остапенко М.С., Штогріна О.С., Глоба Л.С., Астраханцев А.А., Сіменс Едуард

Розробка системи комп'ютерного зору повторної ідентифікації

Проблематика. Швидкий розвиток обчислювальних потужностей обчислювальних машин і збільшення кількості даних стали причинами набуття популярності комп'ютерного зору і великої кількості завдань. Одне з них - розпізнавання людей.

Мета досліджень. Метою статті, на основі досліджень, є запропонувати систему комп'ютерного зору. Також покращити існуючі моделі детекції та розпізнавання, які входять в систему.

Методика реалізації. Алгоритми класичного комп'ютерного зору та глибинного навчання були використані для створення системи.

Результати досліджень. Головним внеском дослідження є опис оптимальної структури системи повторної ідентифікації, в термах швидкості та якості. Більше того, описані вимоги до середовища, яке дозволить використовувати системи в реальних умовах із гарантованою якістю.

Висновки. Система комп'ютерного зору повторної ідентифікації, яка є системою реального часу, була розроблена. Вона може використовуватись в реальних умовах.

Ключові слова: машинне навчання; відеоспостереження; повторна ідентифікація; комп'ютерний зір; глибинне навчання.

Остапенко М.С., Штогрина А.С., Глоба Л.С., Астраханцев А.А., Сименс Эдуард

Разработка системы компьютерного зрения повторной идентификации

Проблематика. Быстрое развитие вычислительных мощностей вычислительных машин и увеличения количества данных стали причинами приобретения популярности компьютерного зрения и большого количества задач. Одно из них - распознавание людей.

Цель исследований. Целью статьи, на основе исследований, является предложить систему компьютерного зрения. Также улучшить существующие модели детекции и распознавания, которые входят в систему.

Методика реализации. Алгоритмы классического компьютерного зрения и глубинного обучения были использованы для создания системы.

Результаты исследований. Главным вкладом исследования является описание оптимальной структуры системы повторной идентификации, в термах скорости и качества. Более того, описаны требования к среде, которое позволит использовать системы в реальных условиях с гарантированным качеством.

Выводы. Система компьютерного зрения повторной идентификации, которая является системой реального времени, была разработана. Она может использоваться в реальных условиях.

Ключевые слова: машинное обучение; видеонаблюдения; повторная идентификация; компьютерное зрение; глубинное обучения.