A STUDY OF CAST SEARCH BY PORTRAIT

[1] *Shang-Fu Chen* (陳尚甫), [2] *Chiou-Shann Fuh* (傅楸善)

[1] Graduate Institute of Communication Engineering,
National Taiwan University, Taiwan

[2] Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan

E-mail: r07942144@ntu.edu.tw, fuh@csie.ntu.edu.tw

ABSTRACT

In this paper, we study an extension of person re-identification (re-ID) problem, cast search by portrait. While person re-ID problem has been widely researched, cast search by portrait remains its difficulty, since the distribution of images is varying in wide range. Traditional person re-ID problem only consider images or video sequences from street monitors, while cast search by portrait trying to retrieve cast from all kinds of movies, with different colors, illuminations. In addition, the size of human is variant. One can imagine person in romantic movie is usually large, while person in war movie usually much smaller. Moreover, the candidate images given by the IMDB dataset are detected automatically, which make the images quite noisy and some of them not even involves any person. In this paper, we evaluate different approaches and techniques to alleviate the problem mentioned above, and conduct ablation study for each component we use. By verifying the effectiveness of techniques and components, we conduct an early exploration of the challenging cast search by portrait task.

Keywords: person re-identification, cast search by portrait

1. INTRODUCTION

In computer vision researches, faces and person images are frequently researched subjects due to the importance of their potential applications. For instance, person re-identification (Re-ID) is a practical but challenging topic, which aims to search for a person in database with only one or few reference query images. Solving person Re-ID problem can bring benefit to applications, such as pedestrian retrieval from videos of street monitors, which can be helpful when identifying a certain criminal. In this paper, we consider another Re-ID extension, cast search by portrait. In movie industry, searching movie frames which involve certain cast is a desired solving problem, since it can bring abundant advantages for movie effects post processing. Thus, cast search by portrait is a task trying to retrieve frames from entire movie, given only one portrait for each cast. We consider IMDB movie data collected by [4]. They

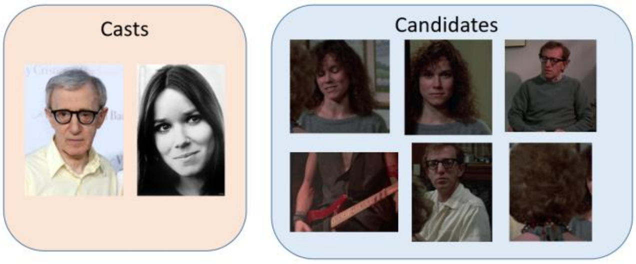

Fig. 1: An illustration of data. For a movie, there are several casts to be retrieved. The candidate images involve characters belong to one of the casts together with an additional class 'other', where no person can be identified.

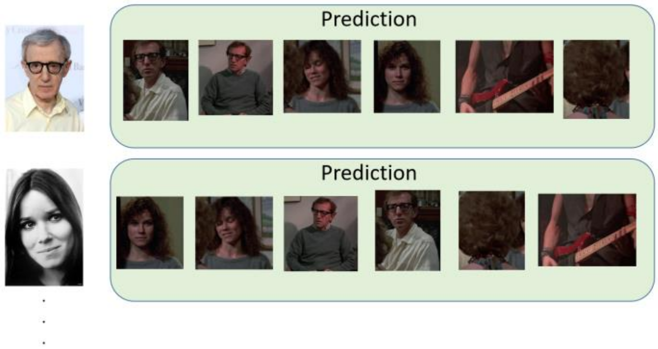

Fig. 2: An illustration of cast retrieval. For each cast, similarity score is computed for every candidate in the movie. Then, the candidates are ranked by the scores and mAP metric is applied for evaluation.

released a new movie dataset this year, which contains 250 movies. We further split the data into training, validation and testing set with 125, 25 and 100 movies, respectively.

2. RELATED WORKS

2.1. Person Re-Identification

Person re-identification can be categorized into two main approaches, single-shot approach and multipleshot approach, as mentioned in [5].

Single-shot approach focuses on associating pairs of images, each containing one in
stance from a class. Cast search by portrait studied in this paper is single-shot approach, since only one portrait of cast is used for retrieval, and the portrait is compared with each candidate in the movie. This process is associating a pair of images and evaluating if they are come from same class, which is consistent to the definition of single-shot approach in [5].

Multipleshot approaches consider multiple images of the same person as training data. When testing inference, multiple images of the same person is also accessible, making multipleshot a relatively easier task.

2.1.1. Single-shot re-ID approaches

As to single-shot approaches, in [6] the method tried to segment a pedestrian image into regions, and explore the relation between regions. This technique works well when the variety of viewpoint is small, since segmentation is hard to apply on a totally different viewpoint.

Viewpoint variance is considered and addressed in [7]. They utilized spatial and color information at the same time to make the extract feature more viewpoint invariant. Together with the technique of boosting and a help for local feature classifier, they alleviate the viewpoint problem.

Finally, [8] consider re-ID problem in another direction. Besides extracting the feature of the person to be retrieved, they also enrich the information of visual feature by considering background and other people. By considering more useful information, they achieved better performance on re-ID.

2.1.2. Multipleshot re-ID approaches

As to multiple-shot approaches, [9] applied SVM for classification, while local and global information is extracted and considered at the same time.

In [10], they firstly detected the bounding box of a pedestrian, and extracted feature considering HSL value and pose variations. By accumulating these feature across multiple images, they were able to predict person identity from input images.

[11] tried re-ID problem in a different setting: retrieve person for a short sequence of video. This was also categorized as multipleshot since consecutive frames could be viewed as multiple query images. They extracted spatial and temporal feature, together with a segmentation estimator for solving this problem.

Lastly, similar to the setting mentioned in [11], [12] conducted person re-ID from a short video sequence by matching SURF [13] interest points.

2.2. Face Detection

Face detection is a popular topic in recent years and many approaches are proposed. [16] proposed a fundamental model called detector cascade, which consists a simple-to-complex face classifier. Although this model is able to detect faces, but it fails when the images come from different viewpoint. To address multi-viewpoints problem, there are three categories of faces classifiers defined by [15]: cascade-based, DPM-based and neural network based.

2.2.1. Cascade-based face detection

[17] train a detector for each view of the face and accumulate their results, but the process is time-consuming. To speed up the detection process, [18] and [19] introduces multiclass boosting for multi-viewpoint object detection.

2.2.2. DPM-based face detection

These approaches are based on the technique proposed by [20], which define a face as multiple meaningful parts. These parts are further processed by classifier or latent SVM to extract relationships between them. [21] improves this approach by considering facial landmarks together with the pre-defined facial parts.

2.2.3. Neural-network-based face detection

Deep convolution neural networks are recently used for extracting visual features. [22] proposed R-CNN model for using segmentation technique, and the model can be used for image retrieval or object detection. [23] proposed spatial pyramid pooling technique to consider features in different scales, and keep the computational cost reasonable at the same time. Finally, a robust face detection model MTCNN is proposed by [1], which is the face detection model we use in this paper.

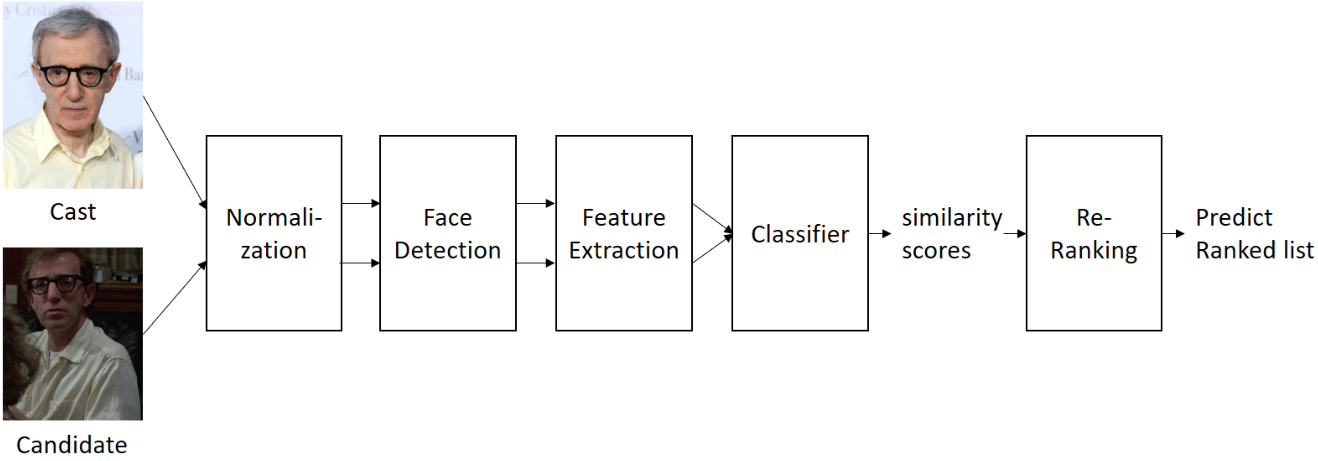

Fig. 3: A pipeline for cast search by portrait. Normalization & face detection is described in Section 3.3. The choice of feature extraction model is mentioned in Section 3.2. The classifier is simply a cosine similarity to evaluate the similarity between cast and candidate feature. The technique of re-ranking is described in Section 3.5.

3. CAST SEARCH BY PORTRAIT

3.1. Task Description

We define the cast search by portrait in detail in this section. For i-th movie, n_i is the number of casts and m_i is the number of candidates. For each cast in the movie, one has to predict a ranking list with size m_i, which represents the priority of retrieval given the cast as query. If one would like to retrieve top-k candidates for the query image, the first k elements in the ranking list is the prediction output for the model. Thus, for a movie, n_i ranking lists with size m_i should be generated and evaluated. The training and testing score is computed by averaging the scores over all movies in the set.

3.2. The Choice of Feature

Pre-trained convolution neural network (CNN) models [24, 25, 26, 27, 28, 29, 30] are widely used tools for feature extraction in computer vision tasks.

Imagenet challenge [2] has proved that the feature of 1000-class classification problem on a large scale data is general enough for many computer vision applications. However, the data distribution in cast search by portrait is biased and only involves images mostly related to human. We doubt that if CNN pre-trained on Imagenet still works well on this task. Thus, besides using CNN pretrained on Imagenet, we also experiment CNN model pretrained VGGFace2 [3], which is a face detection dataset and seems more consistent to the nature of our IMDB dataset.

3.3. Data Pre-Processing

Since we observe that, the distribution of the candidate images is quite noisy, since images are coming from different kinds of movies. In horror movie, the illumination of images is usually dark, while the illumination is quite light in sport movie. Moreover, the data even involves some gray scaled movies. In addition to the problem of the properties of images, candidate images are not always a human faces, since there is a classes named 'others', which contains eyes, legs or even non-human images. To alleviate the above problems, data pre-processing such as normalization and face detection are essential.

3.3.1. Image normalization

For image normalization, we simply convert every input images into RGB modes, that is, 3 channel representing red, green and blue pixel values. Then, normalization is conducted to ensure distribution of input is lies in certain mean and standard deviation.

3.3.1. Face detection

While many face detection techniques are useful [31, 32, 33, 34, 35, 36, 37], we applied MTCNN face detector [1] to find faces in the candidate images. For those images which no faces were found, we assumed they belonged to ‘others’ classes and rank them after those candidates with faces.

3.4. Classification

To generate the ranking list, classifiers for cast search by portrait role as functions with a pair of feature extracted from images as arguments, that is,

sim = C(cast, cand),

where *C* represents the classifier, *cast* and *cand* represents cast feature and candidate feature, respectively, and *sim* represents the similarity score of the cast and the candidate. By comparing a cast with every candidate in the same movie, the ranking list of the cast can be accessed by sorting the candidates by the generated similarity score. Candidates with higher similarities are ranked front while those with lower similarities are ranked behind.

For the choice of classifier, the simplest case is utilizing cosine similarity function as classifier. By calculating the cosine similarity between the feature vectors of the input cast and candidate, a similarity score between 1 and -1 is generated. Please note that for those candidates without faces which are filtered out by MTCNN face detector at previous step, we assign -1 to the similarity score between these candidates and any cast, since we assume these candidates do not refer to any cast.

Another choice of classifier is to train a linear multilayer perceptron (MLP) from the training data. However, we tried to train MLPs with different layers as classifier, but achieved mAP 2-3 percent lower than the simplest cosine similarity metric. We state that this is because the input feature is relatively good enough for classification. Since the parameters of the MLP classifier is trained from scratch, if important information is not distilled and reserved in those newly trained weight, the MLP classifier may have negative effect on the extracted face feature. On the other hand, since there is no parameter has to be learned for the cosine similarity metric, this problem would not occur.

We summarize that, cosine similarity is a good enough classifier and baseline for cast search by portrait task, and we would use this classifier in following experiments. If one wants to try deep classifiers on cast search by portrait task, naïve MLP brings disadvantages. Some better classifiers of re-ID tasks may be useful for this task, and we leave this as our future work.

3.5. Re-Ranking

The idea of re-ranking is taking the initial prediction of a ranking problem as pseudo label. By considering additional patterns or feature, we can refine the pseudo label and generate a more confident ranking list. [14] proposed a robust re-ranking algorithm for person re-identification problem, which can be applied on cast search by portrait task. Before the algorithm is applied, we can compute a similarity/distance matrix between each cast and each candidate, denoting as $M_{cast,\ cand}$. Instead of directly sorting $M_{cast,\ cand}$ for prediction, two additional matrixes $M_{cast,\ cast}$ and $M_{cand,\ cand}$, which represent similarity between casts and similarity between candidates, respectively. By computing k-reciprocal nearest neighbors, the information contains in $M_{cast,\ cast}$ and $M_{cand,\ cand}$ can be utilized and the prediction can be refined. To be more specific, if two candidates are found to have high similarity, they should have similar distance to a certain cast. For the cast, if the initial rankings of the two similar candidates differ a lot, the fact that these candidates are similar can be used for refinement.

4. EXPERIMENTS

4.1 Evaluation Metric: mAP

In image retrieval problem, mAP is a widely used matric, which evaluate scores in a ranking mannar. The metric can be computed as follow:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{m_q}\sum_{k=1}^{n_q}P_q(k)rel_q(k)$$

where Q is the number of query cast, m_q is the number of candidates with the same identity with the cast, n_q is the number of candidates in the movie, $P_q(k)$ is the precision at rank k for the q-th query, and $rel_q(k)$ is a indicator function that is 1 if k-th prediction is positive and 0 otherwise. Precision at rank k is $P_q(k)$ computed as n_P/k, where n_p denotes the number of positive candidates at rank k. For example, if 4 positive candidates are retrieved at rank 5, then $P_q(k) = 0.8$.

The function evaluates the retrieval problem in a ranking manner. It encourages all positive/correct candidates are list in the front of the prediction ranking list. Though the indicator function $rel_q(k)$ filters out negative candidates, the mAP score can be low if positive candidates are sorted behind negative candidates, which make the precision at rank k $P_q(k)$ decrease.

4.2 Feature Extraction Experiment

As mentioned in Section 3.2, we doubt if the Imagenet feature is suitable for cast search by portrait. We evaluate results with different pre-trained CNN feature. The first feature extractor is resnet152 pre-trained on Imagenet, which is a widely used extractor in computer vision tasks. The second feature extractor is resnet50 pre-trained on a face dataset, VGGFace2. After the feature is extracted, we simply feed the feature into the

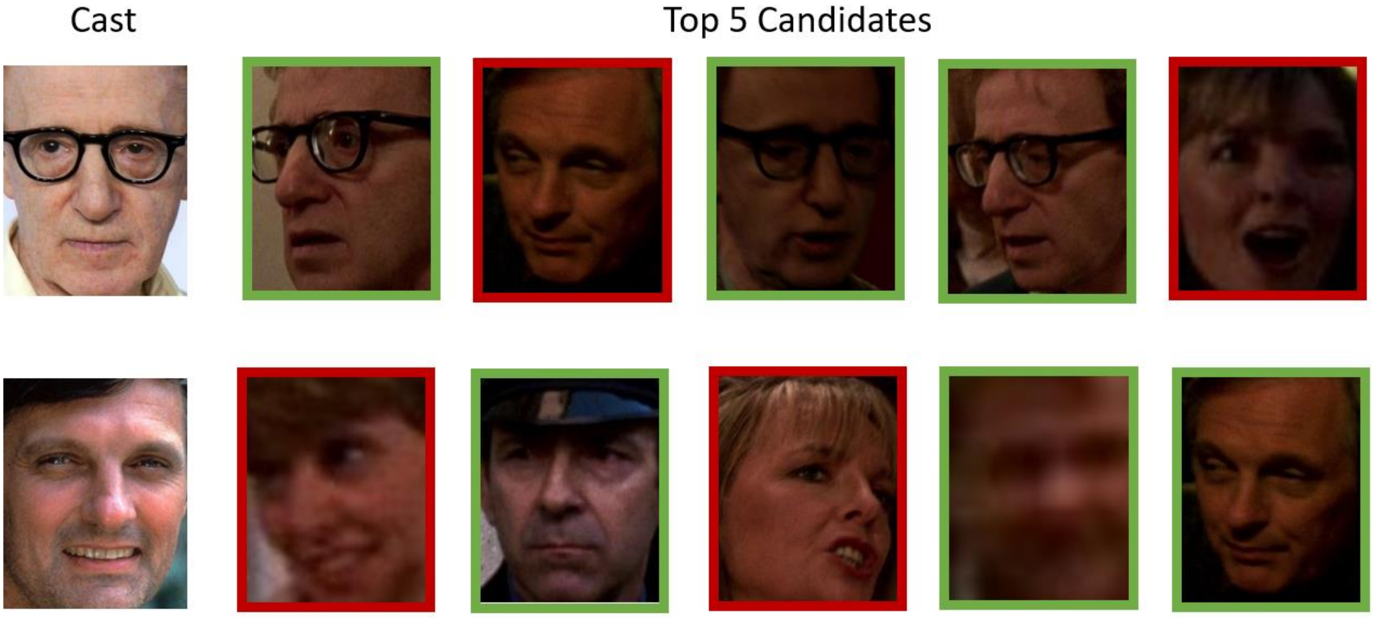

Fig. 4: Examples of prediction results. Since the length of complete ranking lists of candidates are too long, we only show top-5 candidates with highest similarity score. Correct candidates are annotated with green boxes, while wrong candidates are annotated with red boxes.

classifier, a cosine similarity metric and compute the mAP score on testing set. We show the results in Table 1. We observe that, even the first model resnet152 has more parameters and pre-trained on larger data, the performance is still inferior to the second model. This fact proves that more parameters and training data do not always lead to a better performance, and the nature and distribution of data should be also carefully considered. Thus, the second model, resnet50 pre-trained on VGGFace2, is used for following experiment, since its can extract discriminative face feature, which is more favorable for cast search by portrait task.

Table 1: The mAP score for CNN feature extractor pre-trained on different dataset

model	mAP
Resnet152	0.1403
Resnet50 on VGGFace2	0.1549

4.3 Face Detection Experiment

We conduct ablation study for applying face detection technique before the feature is extracted. In this experiment, we use same feature extractor, resnet50 pre-trained on VGGFace2, and the only control variable is whether MTCNN [1] face detector is applied. As described in Section 3.3.2, we crop the faces that MTCNN detect and extract the face feature. For those candidates which no faces are found, we assume that they belong to 'others' class, and rank them at the end of the ranking list. This idea can be viewed as a prior classification, since MTCNN actually filters out some noisy candidates.

As depicted in Table 2, we can observe that face detection technique brings great advantage to cast search by portrait task. While the improvement seems surprising, we think it is reasonable since the original candidate images are quite noisy. The face MTCNN can alleviate this problem in two aspects. First, for those images without any person or faces, MTCNN filters them out and thus make the following ranking problem easier.

Table 2: The ablation study for face detection

Resnet50 on VGGFace2	mAP
without face detection	0.1549
with face detection	0.3857

4.4 Re-Ranking Experiment

We applied the k-reciprocal nearest neighbors algorithm proposed in [14] for our cast search by portrait task. The backbone model is the best model found in Section 4.2, which use face-cropped feature extracted by resnet50 pre-pretrained on VGGFace2. As depicted in Table 3, re-ranking proves its power on person re-ID problem. Though there are parameters of re-ranking algorithm to be selected for the best performance, the default parameters described in [14] can reach mAP larger than 0.4, which proves the robustness of the algorithm.

Table 3: The ablation study of re-ranking on face-cropped feature

model	mAP
without re-ranking	0.3857
with re-ranking	0.4295

4.5 Experiment Results

Two examples of cast search by portraits are shown in Fig. 4. It is worth noting that, cast search by portrait is trying to retrieve correct cast from over 1000 candidates, with a lot of candidates are annotated as ‘others’, so only less than 1/10 candidates refer to the query cast. Thus, 3 correct predictions out of top 5 candidates are acceptable result.

From Fig. 4, it can be observed that the distributions of candidates are quite noisy. In some scenes with many people, face of a single person only occupy small area, and every cropped faces of candidates are resized to same size for the feature extractor, i.e. 224x224. This make some candidates blurry and look quite different from others images, such as 1st and 4th candidates in the second row of Fig.4. This is one of the major challenges of cast search by portrait, and prediction model for this task must be robust enough for all kind of input images.

From Fig. 4, we can also find that our model always gives some candidates high score in spite of which cast is queried, such as, 2nd candidate in the first row and 5th candidates in the second row. We think this problem is resulted from the feature extractor, because the features of some candidates can always lead to high confident scores. To address this problem, feature extractors with two input images can be applied. If the feature extractors can distill useful information while the similarity between two input images is also considered, better prediction mAP may be achieved.

5. CONCLUSIONS

In this paper, we study a new task proposed by WIDER [4], cast search by portrait. We examined techniques and approaches that are potentially beneficial to this task, and find out some important facts from our experiments.

In Section 4.2, we find that the most used feature extractors which are pre-trained on Imagenet are not quite suitable for cast search by portrait. Although these extractors proved they are general enough for predicting cast search by portrait, extractor pre-trained on VGGFace2 performances better with less model parameters.

In Section 4.3, we prove that face detection is an essential pre-process for cast search by portrait. Since the candidate images are noisy, filtering out candidates with faces is important, and comparing face features of these candidates with casts portrait make the classification problem easier.

In Section 4.4, we find re-ranking technique can further improve the performance of the initial prediction of models. By considering relationship between casts and candidates, the ranking results can be refined and a better mAP score can be achieved.

Finally, in Section 3.4, we mention that naïve neural network classifier such as MLP bring no advantage to cast search by portrait problem. A simplest similarity metric, cosine similarity, is a good enough baseline classifier for this task. In re-ID problem, many strong and robust classifiers have been proposed, and they may be utilized for cast search by portrait. We leave searching for better classifier as future work to be explored.

REFERENCES

[1] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499-1503.

[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[3] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018, May). Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 67-74). IEEE.

[4] Loy, C. C., Lin, D., Ouyang, W., Xiong, Y., Yang, S., Huang, Q., ... & Yan, J. (2019). WIDER Face and Pedestrian Challenge 2018: Methods and Results. arXiv preprint arXiv:1902.06854.

[5] Farenzena, M., Bazzani, L., Perina, A., Murino, V., & Cristani, M. (2010, June). Person re-identification by symmetry-driven accumulation of local features. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 2360-2367). IEEE.

[6] Wang, X., Doretto, G., Sebastian, T., Rittscher, J., & Tu, P. (2007, October). Shape and appearance context modeling. In 2007 ieee 11th international conference on computer vision (pp. 1-8). Ieee.

[7] Gray, D., & Tao, H. (2008, October). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In European conference on computer vision (pp. 262-275). Springer, Berlin, Heidelberg.

[8] Prosser, B. J., Zheng, W. S., Gong, S., Xiang, T., & Mary, Q. (2010, August). Person re-identification by support vector ranking. In BMVC (Vol. 2, No. 5, p. 6).

[9] Nakajima, C., Pontil, M., Heisele, B., & Poggio, T. (2003). Full-body person recognition system. Pattern recognition, 36(9), 1997-2006.

[10] Bird, N. D., Masoud, O., Papanikolopoulos, N. P., & Isaacs, A. (2005). Detection of loitering individuals in public transportation areas. IEEE Transactions on intelligent transportation systems, 6(2), 167-177.

[11] Gheissari, N., Sebastian, T. B., & Hartley, R. (2006). Person reidentification using spatiotemporal appearance. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 1528-1535). IEEE.

[12] Hamdoun, O., Moutarde, F., Stanciulescu, B., & Steux, B. (2008, September). Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In 2008 Second ACM/IEEE International Conference on Distributed Smart Cameras (pp. 1-6). IEEE.

[13] Bay, H., Tuytelaars, T., & Van Gool, L. (2006, May). Surf: Speeded up robust features. In European conference on computer vision (pp. 404-417). Springer, Berlin, Heidelberg.

[14] Zhong, Z., Zheng, L., Cao, D., & Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1318-1327).

[15] Farfade, S. S., Saberian, M. J., & Li, L. J. (2015, June). Multi-view face detection using deep convolutional neural networks. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (pp. 643-650). ACM.

[16] Viola, P., & Jones, M. J. (2004). Robust real-time face detection. International journal of computer vision, 57(2), 137-154.

[17] Wu, B., Ai, H., Huang, C., & Lao, S. (2004, May). Fast rotation invariant multi-view face detection based on real adaboost. In Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings. (pp. 79-84). IEEE.

[18] Torralba, A., Murphy, K. P., & Freeman, W. T. (2007). Sharing visual features for multiclass and multiview object detection. IEEE Transactions on Pattern Analysis & Machine Intelligence, (5), 854-869.

[19] Saberian, M., & Vasconcelos, N. (2014). Multi-resolution cascades for multiclass object detection. In Advances in Neural Information Processing Systems (pp. 2186-2194).

[20] Felzenszwalb, P. F., McAllester, D. A., & Ramanan, D. (2008, June). A discriminatively trained, multiscale, deformable part model. In Cvpr (Vol. 2, No. 6, p. 7).

[21] Ramanan, D., & Zhu, X. (2012, June). Face detection, pose estimation, and landmark localization in the wild. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2879-2886).

[22] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[23] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9), 1904-1916.

[24] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence.

[25] Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016, October). Xnor-net: Imagenet classification using binary convolutional neural networks. In European

Conference on Computer Vision (pp. 525-542). Springer, Cham.

[26] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[27] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence.

[28] He, K., Zhang, X., Ren, S., & Sun, J. (2016, October). Identity mappings in deep residual networks. In European conference on computer vision (pp. 630-645). Springer, Cham.

[29] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[30] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[31] Viola, P., & Jones, M. J. (2004). Robust real-time face detection. International journal of computer vision, 57(2), 137-154.

[32] Rowley, H. A. (1999). Neural network-based face detection (No. CMU-CS-99-117). CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

[33] Hsu, R. L., Abdel-Mottaleb, M., & Jain, A. K. (2002). Face detection in color images. IEEE transactions on pattern analysis and machine intelligence, 24(5), 696-706.

[34] Osuna, E., Freund, R., & Girosi, F. (1997, June). Training support vector machines: an application to face detection. In cvpr (Vol. 97, No. 130-136, p. 99).

[35] Sung, K. K., & Poggio, T. (1994). Example Based Learning for View-Based Human Face Detection (No. AI-M-1521). MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB.

[36] Jesorsky, O., Kirchberg, K. J., & Frischholz, R. W. (2001, June). Robust face detection using the hausdorff distance. In International conference on audio-and video-based biometric person authentication (pp. 90-95). Springer, Berlin, Heidelberg.

[37] Ramanan, D., & Zhu, X. (2012, June). Face detection, pose estimation, and landmark localization in the wild. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2879-2886).