Journal of Computers for Society

Journal homepage: https://ejournal.upi.edu/index.php/JCS

Exemplar Based Convolutional Neural Network for Face Search on CCTV Video Recording

Winda Mauli Kristy[*] *, Yaya Wihardi , Erlangga*

Department of Computer Science Education, Universitas Pendidikan Indonesia, Indonesia
*Correspondence: E-mail: windamkristy@gmail.com

ABSTRACT

Many techniques can perform effective face searches, but generally, these methods require numerous samples, particularly when using deep learning approaches. However, there are scenarios where face searches must be conducted with limited samples, such as those obtained from CCTV video recordings, making prior training infeasible. In these situations, a method based on exemplars must be implemented. This investigation utilizes a convolutional neural network (CNN) approach coupled with two unique matching techniques: cross-correlation matching (CCM) and normalized cross-correlation matching (NCC). The study makes use of the Chokepoint Face Dataset, training the data through the optimization of triplet loss. The goal of the study is to evaluate the performance of these combined methods. Two different architectures are created and tested within each method to determine the accuracy of each architecture. The CNN-NCC method has been found to yield accuracy rates that surpass those of the CNN-CCM method by 2 to 17.9%. Nevertheless, it is important to note that the accuracy of the results is greatly influenced by the variations observed in the CCTV video recordings.

ARTICLE INFO

Article History:
Submitted/Received 20 Feb 2023
First Revised 18 Apr 2023
Accepted 12 Jun 2023
First Available Online 13 Jun 2023
Publication Date 15 Sep 2023

Keyword:
CCTV,
Convolution Neural Network,
Cross-correlation Matching,
Normalized Cross-correlation Matching,
Exemplar-Based,
Face Search,
Triplet Loss Optimization.

1. INTRODUCTION

Face search is a technique that involves searching for individuals by utilizing their facial features as a distinguishing factor, focusing on the specific area of interest for each person. The application of face search has found extensive use in tackling diverse issues, including searching for faces within images of human crowds (Dunn, 2018), CCTV surveillance cameras (Mileva & Burton, 2019), facial images shared on websites or social media (Wang *et al,* 2017), and many others.

Face search is not easy to do because you have to accommodate face variations with various variants of rotation, lighting, illumination, expression, and so on. To solve these face-search-related problems, many studies have been conducted in various different domains. One of them is in image-based face search research on a large scale using tens of millions to billions of data (Zou *et al*, 2019). The study used a combination of methods and Hash-based Similarity Search. The results of experiments in the study show that the proposed method is very effective for large search scales in both aspects namely accuracy and real-time properties.

For face searches on CCTV video recordings have also been carried out recently, with the aim of finding perpetrators of crimes at train stations (Mileva & Burton, 2019). Face searches on CCTV video footage are based on images gathered from a variety of sources, including passports, driver's licenses, images of detainees or people searches, and images found on social media (Irshad *et al*, 2021). By using visual search, the results obtained are very dependent on the number of facial images used as sources, the more images, the better the accuracy obtained.

Nevertheless, there are instances where the face being searched is relatively unfamiliar, making it necessary to conduct a face search. So that data about the person being searched and the photo does not exist. For example, if there is a loss of an item in a place and do not know the person who took the item. All the victim remembers is the face of the perpetrator without knowing his identity. The only source that can be used as a source to conduct a face search is CCTV video footage contained in the building. By taking facial images from the CCTV video footage used as a target sample, making facial search difficult to do (Mileva & Burton, 2019). Nevertheless, there are instances where the face being searched is relatively unfamiliar, making it necessary to conduct a face search.

Within this investigation, an exemplar-based approach using the CCN method is proposed, where this method has gained recognition in many image processing tasks, and many approaches are used in the use of CNN in addition to the exemplar-based approach. For example, in research that uses CNN to classify images with high resolution, using an object-based approach. The proposed method is built by combining feature learning strategies with object-based classification. The results obtained from this study get a fairly high level of accuracy (Zhao *et al*, 2017).

For face searches on CCTV video recordings have also been carried out recently, with the aim of finding perpetrators of crimes at train stations (Mileva & Burton, 2019). Face searches on CCTV video footage are based on images gathered from a variety of sources, including passports, driver's licenses, images of detainees or people searches, and images found on social media. By using visual search, the results obtained are very dependent on the number of facial images used as sources, the more images, the better the accuracy obtained.

Nevertheless, there are instances when conducting a facial recognition search involves an unfamiliar face. So that data about the person being searched and the photo does not exist. For example, if there is a loss of an item in a place and do not know the person who took the item. All the victim remembers is the face of the perpetrator without knowing his identity. The only source that can be used as a source to conduct a face search is CCTV video footage contained in the building (Davis & Valentine, 2015). By taking facial images from the CCTV video footage used as a target sample, making facial search difficult to do. The reason for this is that the initial stage of the process does not allow for the training to be conducted on the target face directly. Hence, an exemplar-based approach must be adopted in order to proceed (Zhao *et al*, 2017).

Within this investigation, an exemplar-based approach using the CCN method is proposed, where this method has gained recognition in many image processing tasks, and many approaches are used in the use of CNN in addition to the exemplar-based approach (Tudavekar *et al*, 2021). For example, in research that uses CNN to classify images with high resolution, using an object-based approach. The proposed method is built by combining feature learning strategies with object-based classification. The results obtained from this study get a fairly high level of accuracy

2. METHODS

he methodology employed in this study is illustrated in **Figure 1**, which depicts the experimental design used for training purposes.

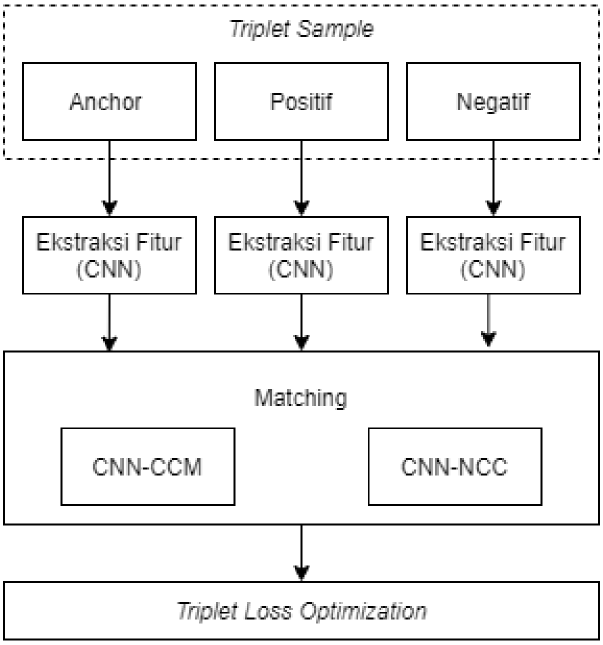

Figure 1. Design Method for Training

The approach utilizes sets of three images, known as triplet samples, as its input. These samples consist of an anchor image, a positive image, and a negative image. To conduct this study, the Chokepoint Face Dataset is utilized. Prior to using the data as a dataset, it is organized into triplets, with each triplet containing an anchor, positive, and negative image. The anchor and positive images depict the same individual, while the negative images portray different individuals from the anchor and positive samples. Notably, the anchor images possess a higher resolution compared to the positive and negative samples, as the latter are extracted from CCTV video frames. Once the triplet samples are established, Haar Cascade Face Detection is employed to detect faces. Subsequently, the facial region of interest (ROI) that is detected is resized from 800x600 pixels to 120x96 pixels.

Once the data has been preprocessed, the feature extraction process is applied to each data pair. Subsequently, the matching process is carried out, and finally, the data proceeds to the optimization phase of triplet loss.

2.1. Feature Extraction

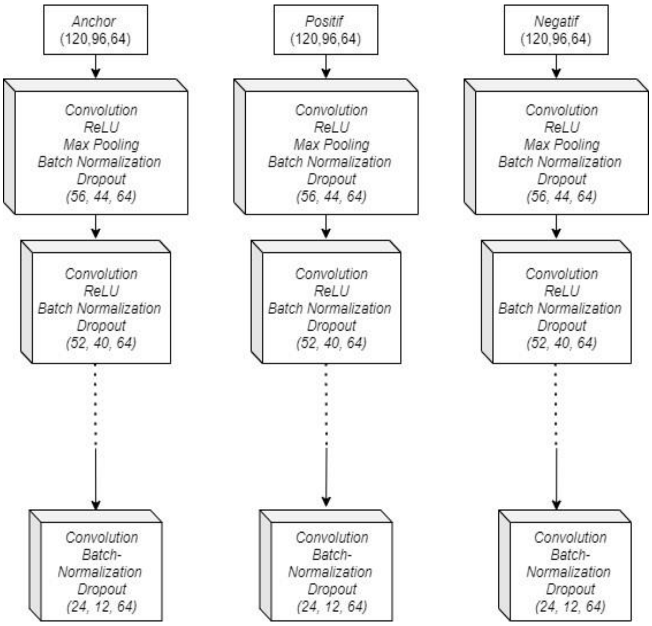
Anchor (120,96,64)
Positif (120,96,64)
Negatif (120,96,64)
Convolution ReLU Max Pooling Batch Normalization Dropout (56, 44, 64)
Convolution ReLU Batch Normalization Dropout (52, 40, 64)
Convolution Batch-Normalization Dropout (24, 12, 64)

Figure 2. Stages of Feature Extraction

In Figure 2, the Convolutional Neural Network (CNN) approach is utilized for feature extraction. Features are extracted for each image pair simultaneously by evenly distributing the weights across each subnetwork. Each subnetwork consists of nine convolutional layers, with batch normalization, dropout, and ReLU activation functions following each layer. However, the ninth convolutional layer does not have the ReLU activation function to prevent the loss of informative data from facial features.

Each convolutional layer uses 64 filters of size 5x5, without padding. The input images for feature extraction are 120x96 pixels in size, resulting in a 24x12x64 filter output. To reduce the facial image size, a MaxPooling layer is added to the first convolutional layer with a pool size and strides of 2x2.

2.2. Matching

The evaluation of the matching process involves the utilization of two separate techniques to assess their effectiveness: cross-correlation matching and normalized cross-correlation matching (Cui *et al*, 2020). Furthermore, both methods are executed using two distinct architectures, leading to a total of four architectures encompassing the two techniques (CNN-CCM and CNN-NCC).

2.2.1. Cross-correlation Mathing (CCM)

The comparison process involves the utilization of two separate techniques to assess their effectiveness: cross-correlation matching and normalized cross-correlation matching (Zui *et* *al*, 2020). Additionally, both techniques are integrated into two unique structures, leading to a total of four structures encompassing the two methods (CNN-CCM and CNN-NCC).

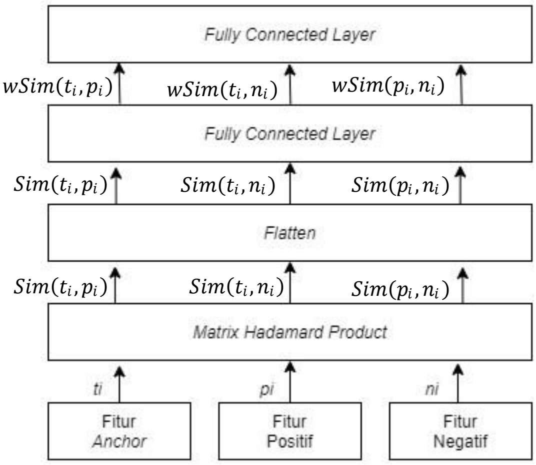

Figure 3. Matching Stages with CCM Method

Figure 3 illustrates a process that comprises of four layers: the Hadamard Product Matrix, a Flatten layer, and two Fully Connected Neural Network layers. This approach involves merging two vectors derived from the extracted features of two image inputs. The merging is accomplished by multiplying the two matrices using the Hadamard Product. The three feature-extracted inputs, namely anchor, positive, and negative, are combined. More specifically, the anchor is combined with the positive, the anchor with the negative, and the negative with the positive.

$$Sim(ti, pi) = (ti \odot pi) \quad (1)$$

$$wSim(ti, pi) = wm.RELU(wn.\ Sim(ti, pi) + bn) + bm \quad (2)$$

The CCM technique is based on a formula developed from the research conducted by Pachami and colleagues (Parchhami *et al*, 20). This approach involves obtaining weights and biases (wm, wn, bn, bm) through multiplication using the Hadamard Product Matrix, which includes anchor, positive, and negative samples (ti, pi, ni). The multiplication process yields a similarity value, which is then inputted into a fully connected neural network. Prior to reaching the fully connected layer, the matrix undergoes a flatten layer transformation to convert it into a vector format, as fully connected layers necessitate vector inputs. The initial fully connected layer incorporates the ReLU activation function. In the CNN-CCM design, the ultimate fully connected layer integrates a Softmax activation function in architecture A, and a Sigmoid activation function in architecture B.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

2.2.2. Normalized Cross-correlation Matching (NCC)

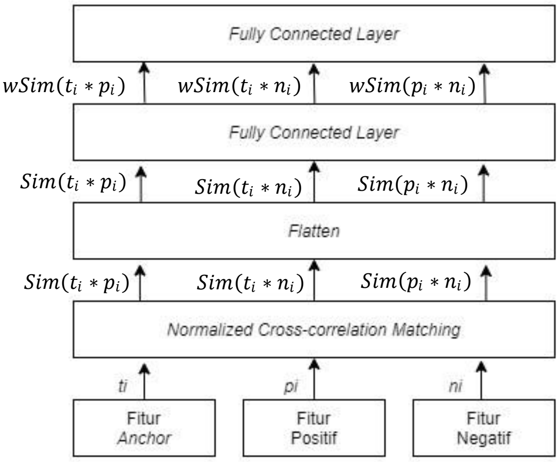

Figure 4. Matching Stages with NNC Method

In **Figure 4** above, it is not much different from the CNNCCM method. The only difference is in the first layer and the last layer. The first layer uses the formula from NCC, the formula used is based on research conducted by Subramnian and Chatterjee (Subramaniam & Chatterjee, 2016). In the formula, the input received will be normalized before being filtered.

$$Sim(ti \,\bar{*}\, pi) \;=\; \frac{\sum_{i=1} N \;(ti - \mu t)(pi - \mu p)}{(N-1).\sigma t.\sigma p} \, (3)$$

Where $\bar{*}$ is a symbol that represents normalized cross-correlation matching. While N represents the total pixels t and p. μt and μp are the average pixel values. σt and σp are the standard deviations of the pixel values t and p.

$$wSim(ti \,\bar{*}\, pi) = wm.RELU(wn.Sim(ti \,\bar{*}\, pi) + bn) + bm \quad (4)$$

wm, wn, bn, bm is the weight and bias obtained using normalized cross-correlation matching. The result of normalized cross-correlation matching, will then enter into a flatten and fully connected layer.

In the first fully connected layer, a ReLU activation function is applied, but in the final fully connected layer, no activation function is used. Despite this, two distinct architectures were developed using the CNN-NCC method (Adem, 2022). Architecture A in the CNN-NCC method features a fully connected layer with 2 units, whereas Architecture B has a fully connected layer with a single unit.

2.2.3 Triplet Loss Optimization

The triplet loss function in this study uses the equation,

$$\text{Triplet Loss} = \frac{1}{L} \sum_{R_i \in B} \sqrt{\left(1 - S_{tipi}\right)^2 + S^2{}_{tini} + S^2{}_{nipi}} \quad (5)$$

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

Where S_{tipi}, S_{tini}, and S_{nipi} is the similarity values obtained from cross-correlation matching (CCM) and normalized cross-correlation matching (NCC) are used for loss optimization based on the research by Parchami et al. The training process involves 50 epochs for each method. After training, the models are tested to evaluate their performance. Testing is conducted with data different from the training set, following an exemplar-based approach, where test data is not used during training.

During training, a triplet network is used, while testing employs a siamese network with two inputs to determine similarity. The first input is a target facial image, and the second input comprises frames from CCTV footage. Testing uses the embedding model from the trained triplet network. After predictions are made using this model, the output vectors are used to calculate similarity values.

For architecture A of both the CNN-CCM and CNN-NCC methods, cosine similarity is used to obtain similarity values. For architecture B of these methods, the similarity value is directly obtained from the prediction results using the embedding model.

3. RESULTS AND DISCUSSION

Testing was conducted using three different scenarios. In the first scenario, testing is done using testing data. Testing with the second scenario was carried out using case studies that had been provided by the dataset used. While in the third scenario, CCTV video footage is used on a crowd of people, where in the video recording every frame there are at least two subjects caught by CCTV cameras. Here are the test results from all three scenarios:

3.1. Test Results with Test Scenarios Using Data Testing

Testing using pre-made testing data, in which there are 1000 pairs of data. This test data is a facial image of a different identity from the training data. This is because the research conducted is research using exemplar based. Testing will be carried out with several thresholds to find the highest accuracy value. The following are the test results with testing data in **Table 1**.

Table 1. Test Results with Data Testing

Method	Architecture	Threshold	Accuracy
CNN-CCM	Arsitektur A (Softmax)	0.9999998	0.686
		0.9999999	0.629
		1.0	0.667
		1.0000001	0.749
		1.0000002	0.757
	Arsitektur B (Sigmoid)	0.627	0.763
		0.628	0.767
		0.629	0.726
		0.630	0.691
		0.631	0.582
CNN-NCC	Arsitektur A (2 *units*)	0.9999994	0.646
		0.9999998	0.772
		0.9999999	0.700
		1.0	0.677
		1.0000001	0.775
	Arsitektur B (1 *unit*)	0.1	0.709
		0.2	0.594

Method	Architecture	Threshold	Accuracy
		0.3	0.599
		0.4	0.773
		0.5	0.799
		0.6	0.787
		0.7	0.780

In this test, the greatest accuracy results were found in tests using CNN-NCC architecture B with an accuracy of 0.799. The accuracy is obtained using a threshold of 0.5.

3.2. Test Results with Test Scenarios Using Data Testing

Testing using pre-made testing data, in which there are 1000 pairs of data. This test data is a facial image of a different identity from the training data. This is because the research conducted is research using exemplar based. Testing will be carried out with several thresholds to find the highest accuracy value. The following are the test results with testing data in **Table 2**.

Table 2. Recorded Video Data

Folder Name	Type Scene	Total Frame	Number of Frames to Find
P2E_S4_C2	Outdoor	803	100
P2E_S3_C1	Outdoor	873	106
P2L_S4_C2	Indoor	716	111
P2L_S3_C3	Indoor	853	98

Where E, signifies the case with the subject entering the room (entering) and L, signifies the case with the subject leaving the room (leaving). S indicates the sequence of the video and C indicates the camera recording. Testing is carried out using the threshold with the highest accuracy of testing using testing data. For test results with case studies, here are the results obtained in **Table 3**.

Table 3. Test Results with Case Studies

Folder Name	CNN-CCMA	CNNCCMB	CNNNCCA	CNNCCB
P2E_S4_C2	0.761	0.842	0.831	0.940
P2E_S3_C1	0.792	0.751	0.809	0.852
P2L_S4_C2	0.668	0.724	0.651	0.740
P2L_S3_C3	0.722	0.668	0.708	0.813

From the tests carried out, the highest accuracy was found using the CNN-NCC method architecture B. Where CNN-CCM is superior to about 2-17.9% of other methods.

3.3. Test Results with with Test Scenarios On Crowds

If testing using case studies is carried out using CCTV video recordings where there is only one subject in each frame, then this test is carried out to determine the performance of each method on CCTV video recordings where in each frame there are at least two subjects. Tests are carried out using new thresholds, this is because the B architecture of each method outputs similarity values with different ranges from the previous two tests. The video footage used in this test consisted of 775 frames, with approximately 2113 facial images detected. In the video footage, there are 159 frames with the target's face that must be found. Here are the test results with scenarios in crowds in **Table 4**.

Table 4. Test Results on Crowds

Method	Architecture	Threshold	Accuracy
CNN-CCM	Arsitektur A (Softmax)	0.9999994	0.756
		0.9999998	0.798
		0.9999999	0.927
		1.0	0.756
		1.0000001	0.926
	Arsitektur B (Sigmoid)	0.632	0.922
		0.633	0.770
		0.634	0.601
		0.635	0.825
CNN-NCC	Arsitektur A (2 *units*)	0.9999994	0.849
		0.9999998	0.951
		0.9999998	0.945
		1.0	0.653
		1.0000001	0.787
	Arsitektur B (1 *unit*)	0.1	0.819
		0.2	0.602
		0.3	0.865
		0.4	0.967

Just like the other two test scenarios, the highest accuracy was found with the CNN-NCC method of architecture B. Where the accuracy obtained was 0.967. The accuracy is obtained using a threshold of 0.4.

From the results of accuracy, the CNN-NCC method of architecture B outperforms other methods. Although the accuracy found is not too much different, which is only 2-17.9% higher than other methods. However, in conducting face search, the CNN-NCC method with architecture B has good performance. In the face search performed, the method performs a face search according to the example of the target face sought. This is different from the other three architectures, where there is still an error in finding faces. Where the face found is different from the image of the face used as an example.

Although it has done a good face search, this method still has a drawback where, the method cannot perform a face search if there are variations in poses, expressions, and scales that are different from the target face that is the example. So that the accuracy obtained depends very much on the variations that exist in the CCTV video recording.

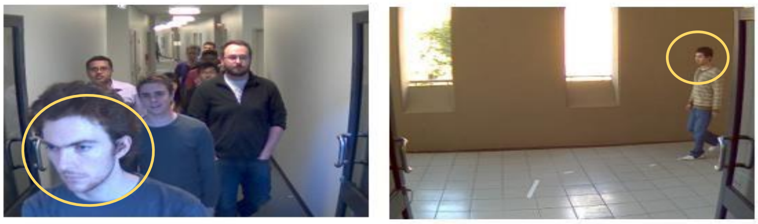

Figure 5. Examples of Pose Variations That Can't Be Found

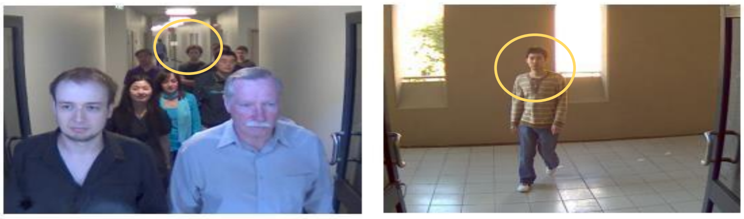

Figure 6. Examples of Pose Scale That Can't Be Found

4. Conclusion and Advice

4.1. Conclusion

Research was conducted on facial search in CCTV video footage utilizing Convolutional Neural Network (CNN) techniques. These were complemented by two distinct matching approaches: Cross-correlation Matching (CCM) and Normalized Cross-correlation Matching (NCC). Training involved optimizing triplet loss, necessitating three samples as input.

Before entering the training process, face detection was carried out and changed the image size from 800x600 pixels to 120x96 pixels. Training is conducted with two architectures (A and B) from two methods, namely CNNCCM and CNN-NCC. After the training, the resulting model will be tested with three different scenarios.

In tests conducted by the CNN-NCC method architecture B, showed better performance than other methods. Where the CNN-NCC method has 2-17.9% accuracy which is higher than the other three methods. Although it has been able to outperform the other three methods, it still cannot handle variations in expressions, poses, and scales. So that the accuracy obtained is very dependent on the variations that exist in the CCTV video recording.

4.2. Advice

Although the proposed method is already capable of conducting face searches, and it can already outperform the CNN-CCM method in terms of accuracy. But the method still does not handle variations that are commonly found in search, recognition, and face detection. Variations that still cannot be addressed, especially variations in poses, variations in expression, and variations in scale.

Therefore, in order for future research to get better results, the authors suggest replacing the matching method that can handle variations in face search. This is because, CCTV video recordings in general have many variations that are a challenge in conducting face searches. So it is hoped that by overcoming the problem of variation, the accuracy obtained can be further increased.

5. REFERENCES

Adem, K. (2022). Impact of activation functions and number of layers on detection of exudates using circular Hough transform and convolutional neural networks. *Expert Systems with Applications*, *203*, 117583.

Bhople, A. R., and Prakash, S. (2021). Learning similarity and dissimilarity in 3D faces with triplet network. *Multimedia Tools and Applications*, *80*(28), 35973-35991.

Cui, Z., Qi, W., and Liu, Y. (2020, December). A fast image template matching algorithm based on normalized cross correlation. *In Journal of Physics: Conference Series*, *1693*(1), 012163.

Davis, J. P., and Valentine, T. (2015). Human verification of identity from photographic images. *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and CCTV*, 209-238.

Dunn, J. (2018). Robust search templates for face-in-crowd search (*Doctoral dissertation, UNSW Sydney*).

EnsarE, T., and Günay, M. (2017). Comparison of face recognition algorithms. I*n 2017 25th Signal Processing and Communications Applications Conference (SIU),* 2017, 1-4.

Irshad, M., Zhou, X., Noman, S. M., Murthy, A., Hu, B., Haider, S. A., and Olawale, O. A. (2021). City vision: CCTV images based public surveillance model. In 2021 *International Conference on Electronic Information Technology and Smart Agriculture (ICEITSA)*, *2021, 416*-420.

Mileva, M., and Burton, A. M. (2019). Face search in CCTV surveillance. Cognitive Research: Principles and Implications, *4*, 1-21.

Parchami, M., Bashbaghi, S., and Granger, E. (2017). Cnns with cross-correlation matching for face recognition in video surveillance using a single training sample per person. *In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2017,* 1-6.

Subramaniam, A., Chatterjee, M., and Mittal, A. (2016). Deep neural networks with inexact matching for person re-identification. *Advances in Neural Information Processing Systems*, *29*.

Sumbul, G., Ravanbakhsh, M., and Demir, B. (2021). Informative and representative triplet selection for multilabel remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-11.

Tudavekar, G., Patil, S. R., and Saraf, S. S. (2021). Dual-tree complex wavelet transform and deep CNN-based super-resolution for video inpainting with application to object removal and error concealment. *In Computational Intelligence Methods for Super-Resolution in Image Processing Applications*, *2021*, 231-248.

Wang, D., Otto, C., and Jain, A. K. (2016). Face search at scale. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1122-1136.

Zhao, W., Du, S., and Emery, W. J. (2017). Object-based convolutional neural network for high-resolution imagery classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(7), 3386-3396.

Zou, F., Yang, F., Chen, W., Li, K., Song, J., Chen, J., and Ling, H. (2020). Fast large scale deep face search. *Pattern Recognition Letters, 130*, 83-90.