Image and video face retrieval with query image using convolutional neural network features

Imane Hachchane[1], Abdelmajid Badri[1], Aïcha Sahel[1], Ilham Elmourabit[1], Yassine Ruichek[2]
[1]Laboratoire d'Electronique, Energie, Automatique and Traitement de l'Information (EEA and TI), Faculté des Sciences et Techniques Mohammedia, Université Hassan II Casablanca, Mohammedia, Morocco
[2]IRTES-Laboratoire SET, Université de Technologie de Belfort Montbéliard, Belfor, France

Article Info

Article history:

Received May 7, 2021
Revised Dec 22, 2021
Accepted Dec 29, 2021

Keywords:

Classification
Convolutional neural network
Faster R-CNN
Image and video retrieval
Image processing
Image to video instance retrieval
Object recognition

ABSTRACT

This paper addresses the issue of image and video face retrieval. The aim of this work is to be able to retrieve images and/or videos of specific person from a dataset of images and videos if we have a query image of that person. The methods proposed so far either focus on images or videos and use hand crafted features. In this work we built an end-to-end pipeline for both image and video face retrieval where we use convolutional neural network (CNN) features from an off-line feature extractor. And we exploit the object proposals learned by a region proposal network (RPN) in the online filtering and re-ranking steps. Moreover, we study the impact of finetuning the networks, the impact of sum-pooling and max-pooling, and the impact of different similarity metrics. The results that we were able to achieve are very promising.

Corresponding Author:

Imane Hachchane
Laboratoire d'Electronique, Energie, Automatique and Traitement de l'Information (EEA and TI), Faculté des Sciences et Techniques Mohammedia, Université Hassan II Casablanca
Mohammedia, Morocco
Email: hachchaneimane@gmail.com

1. INTRODUCTION

The massive advances in internet technologies and the proliferation of smartphones, digital cameras and storage devices led to an increase in the popularity of visual search applications such as image retrieval, video retrieval or precisely instance search. By comparing a query against a database, instance search is used to extract images or videos of a particular object from large databases. It has been commonly used in product recognition, property identification, and other applications [1]–[3].

We should note that in one hand, image-to-image retrieval is a well-known field where large-scale face image retrieval has recently attracted attention, and a wide variety of methods have been proposed for face recognition and retrieval [4]–[7]. Following proper adaptation, well-known techniques for image retrieval were used for face recognition/retrieval, such as bag-of-visual words (BoVW). Other recent studies used convolutional neural network (CNN) for the feature extraction task [6].

On the other hand, image-to-video retrieval [8]–[10] is an asymmetric problem where the lack of temporal information in images stops us from using standard techniques for extracting video descriptors [11]–[14]. Traditionally, image-to-video retrieval techniques are based on a classic extraction methodes of hand-crafted features scale invariant feature transform (SIFT) [15], and binary robust independent elementary features (BRIEF) [16]. Smaller effort has been made to adapt deep learning techniques. We can apply standard features for image retrieval [17]–[20] by processing each frame as an independent image. More recent works showed that is possible to use CNN for feature extraction when working on videos [21], [22].

But not much work has been done in combining both, meaning having one pipeline for both image retrieval and video retrieval using one query image. Hence, in this paper, we investigate this issue. We are trying to retrieve the top N most relevant images and/or videos of an instance from a single image query instance. More specifically, we are working on face retrieval. In other words, giving an instance of a face in a query image, we are trying to retrieve the top N most relevant image instances and/or video instances from our database of videos and images of that specific face.

The main contribution of this paper is to build an end-to-end pipeline, for both image and video face retrieval using one query image. The pipeline takes advantage of off-the-shelf and fine-tuned features from an object detection CNN. We tested the impact of multiple similarity metrics, different network architectures, max-pooling and sum-pooling as well as the impact of most common reranking strategies.

2. RELATED WORK

Visual search and retrieval are in general an indexing and querying problem for visual data, which can be further divided into categories depending on the query type and database used. The most studied field in visual retrieval is image-to-image retrieval, where we use a query image to find the most relevant images from an image dataset [23], [24]. Generally speaking, visual search and retrieval remains an issue of indexing and querying visual data. This issue can be categorized depending on the type of queries and databases used. The most studied area in visual retrieval is image-to-image retrieval, were a we use a query image to retrieve the most relevant images from an image dataset [23], [24]. Another area of visual retrieval is video-to-video retrieval where a query video is used to retrieve relevant videos from a video dataset [25]. A further variant is video-to-image retrieval in which we use a query video to search a dataset of images [26], it is usually used in augmented reality. And of course we have the image-to-video retrieval where we search a database of videos using a query image [21]. In this paper, we merge two of those areas: Image-to-image retrieval and image-to-video retrieval. We focus on both image and video retrieval using one query image. More precisely, we are targeting face retrieval. Meaning, giving a query face image we are trying to retrieve the most relevant images and/or videos of that specific face.

Face retrieval is a difficult task because it is hard to adapt traditional image retrieval methodes (like bag of words) are difficult to apply to the field of face research [27]. Because the traditional descriptor based on the detection of key points (like SIFT) often fails due to the smooth surface of the face. Previous work, using a previously trained image classification convolutional neural network as a feature extractor, showed that it is more appropriate to use a fully connected layer for image retrieval [17]. Razavian *et al.* [28] Improved results by combining fully connected layers extracted from different image submatches. Later, the new work found that the convolutional layer is significantly better than the fully connected layer in image retrieval tasks [3], [28].

When working on image-to-image retrieval, a variety of CNN-based object detection pipelines have been proposed. In this paper, we are interested in Faster R-CNN [29], a CNN network created by Ren *et al.* They used a region proposal network (RPN) [30] in Faster R-CNN to remove the dependence of object propositions that exists in older CNN object detection systems. And, even though Faster R-CNN is designed to detect genral objects, Jiang and Learned-Miller [31] were able to highlight its impressive face detection performance, especially when retrained on a suitable face detection training set [6]. The current pipeline, that we are working on, uses off-the-shelf and finely tuned features of Faster R-CNN's end-to-end object detection architecture to extract global and local convolutional features in one pass and test their utility for image and video face retrieval using one query face image. We also test the impact of different similarity metrics, network architectures, max-pooling and sum-pooling, as well as reranking strategies.

3. METHODOLOGY

3.1. CNN-based representations

In our new pipeline, Figure 1, we examine the importance of using local and global CNN features extracted from pre-trained Faster R-CNN models [29] for image and video face retrieval. We use bounding boxs above our query images to define the instances that we are looking for. Faster R-CNN had two major parts that share a convolutional layer. The first one is RPN; it is a small neural network that glides over the last feature map of the convolution layers to predict whether an object is present or not, as well as the bounding box of those objects called windows. The second one is the classifier that learns to label each of those objects as one of the classes in the learning dataset [3].

As with earlier works [3], [32], and [33] our objective is to derive a compact image representation from Faster R-CNN activations. We construct the global descriptor by ignoring all of Faster R-CNN's layers that work with object propositions, and we derive features from the last convolutional layer. Taking the extracted activations of the convolution layer for an image or a frame into consideration, we group the activations of each filter to form an image descriptor with the same dimension as the number of filters in the convolution layer.

When working on constracting the local descriptor, the region pooling layer attached to the last convolutional layer is used to extract the convolutional activations for each of the object propositions gathered by the RPN for the local descriptor. This provides the capability of creating a local descriptor for every window proposal by aggregating the activations of that window in the RoI pooling layer. Sum-pooled features are l2-normalized in a manner similar to those described by several other authors [18], [32], followed by whitening and a second round of l2-normalization, while max-pooled features are only l2-normalized once without any whitening.

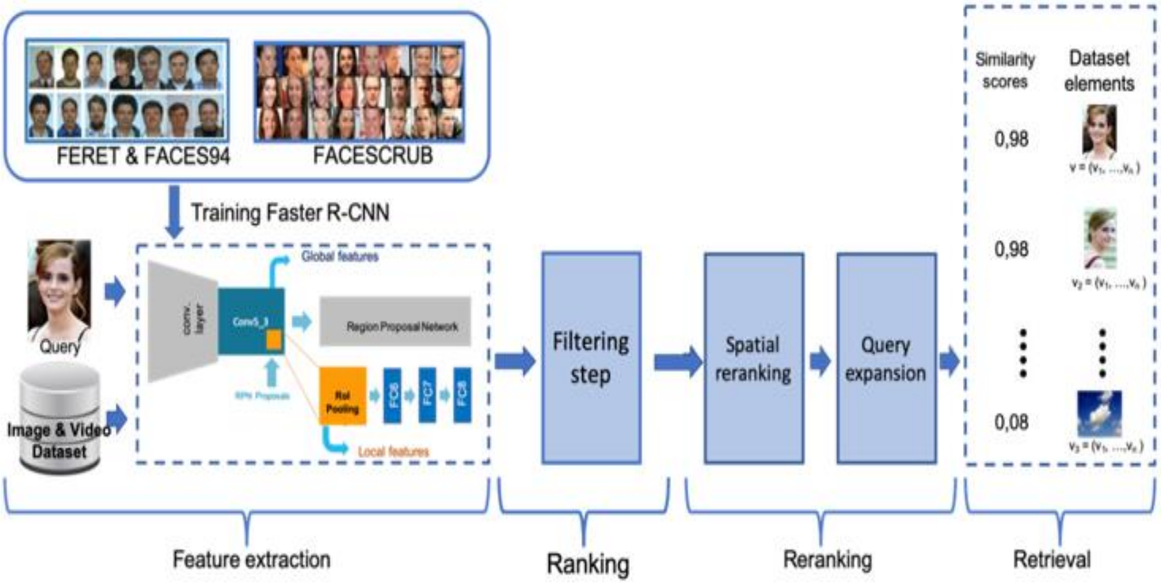

Figure 1. Proposed pipeline’s architecture

3.2. Video and image retrieval

The feature extracting is done offline where we create the descriptors for the images, the video frames and the query images. At testing time (the online portion of the pipeline) we follow the raking strategies described in this section. We start with a filtering step, where the query features are compared to all the dataset items and then ranked using a similarity measure. At this step, we are still considering the entire frame as a query. After the filtering step, we locally analyze and re-rank the N upper elements. It is the spatial re-ranking. Last, we use query expansion (QE), in which we combine the descriptors of the M higher elements of the first ranking with the query descriptor to conduct a new search (M=5).

4. EXPERIMENTS

4.1. Utilized datasets

To test our methods, we need to use a dataset of images and videos. We could not find one, so we decided to merge two existing ones. These are the datasets we used:

- YouTube faces database [34]: The dataset contains 3,425 videos of 1,595 people, all of which were downloaded from YouTube. The database contains an average of 2.15 videos for each subject, with 48 frames being the shortest clip and 6,070 frames being the longest.
- FaceScrub [35]: 22,507 unconstrained face images amassed from the Internet. We added a framing box to the query images to surrounde the target faces.

The datasets we used to fine-tune the network:

- FERET [36]: This dataset has 3,528 images. We provide a framing box to the query images in order to surrounding the target faces.
- FACES94 [37]: This dataset has 2,809 images.

We also used the 55,127 unconstrained face images of the original FaceScrub dataset to fine-tune the network. When testing, we used 111 query images.

4.2. Experimental setup

According to previous works [3], [6], [21] deeper networks achieved better performance in extracting global and local features. Therefore, we decided to use the VGG16 architectures of Faster R-CNN and compare it with the ZF architecture to test the validity of the theory when working on both image and video retrieval. When working with the VGG16 architecture, the global descriptors are extracted from the last convolution layer “conv5_3” and are of dimension 512. And when working with the ZF architecture, the global descriptors are extracted from the last convolution layer “conv5” and are of dimension 256. For the local features, we group them from the Faster R-CNN RoI clustering layer. The global descriptors for the VGG16 architecture are extracted from the last convolution layer “conv5_3” and are of dimension 512, while the global descriptors for the ZF architecture are extracted from the last convolution layer “conv5” and are of dimension 256. We group local features using the Faster R-CNN region of interest (RoI) clustering layer.

We also experimented with widely used similarity metrics to see which one is more suitable for our pipeline. We tested the following similarity metrics: Cosine similarity metric, Euclidien similarity metric, Manhatan similarity metric, Chebychev similarity metric, Minkowski similarity metric, Canberra similarity metric, and Corrolation similarity metric. The following specifications were used for the experiments: Processor: Intel(R) Core (TM) i7-7700K CPU 4.20 GHz, RAM: 16 GB, OS: Ubuntu 16.04, Graphics card: NVIDIA GeForce GTX 1070.

We should note the extraction time for the VGG16 required an average of 16h 11min 30 seconds compared to an average of 7h 34min and 22 seconds when using ZF. This time difference can be explained by the sizes of the networks. The ranking took on average 2 seconds per query image; the re-ranking took an average of 16 seconds per query image, and when using the QE, the re-ranking took an average of 17 seconds per query image.

4.3. Off-the-shelf CNN features

In this section we evaluate using Faster R-CNN features for face image and/or video retrieval. We have tested different similarity metrics, as detailed above. The results, displayed in Table 1, were similar and close, but the best results were obtained using the cosine and the euclidien similarity metrics combined with our re-ranking strategies with a precision of 55.4%. But with the other similarity metrics, the query expansion and the spatial reranking did not improve the results.

Moreover, a comparative study of the sum and max pooling strategies of image-wise and region-wise descriptors was also conducted, with the results summarized in Table 1. Sum-pooling is better than max-pooling, according to our tests. It also confirms that Faster R-CNN with a VGG16 architecture trained on pascal VOC datasets performed best, which is consistent with previous research that had demonstrated that deep networks could deliver better results when extracting global and local features.

4.4. Fine-tuning the CNN

More importantly, we investigated the effects of fine-tuning a pre-trained network on recovery performance with the query objects to retrieve. We used the model VGG16 of Faster R-CNN pre-trained with the pascal VOC objects. We refined it using two datasets:

- We refined the first network with FERET and Faces94 datasets and we called it VGG16 (Feret and Faces94). Because of their small size, the Feret and the Faces94 datasets were combined, and the network’s output layer was modified to return 422 class probabilities and their corresponding bounding box coordinates [6] (the 422 counts for the 269 classes in the FERET dataset and the 152 classes in the Faces94 dataset, plus one additional class for the background).
- We refined the second network with using the FaceScrub dataset. We called it VGG16 (Facescrub). For this network the output layer was modified to return 530 class probabilities and their corresponding bounding box coordinates (530 classes, plus one additional class for the background).

The initial parameters of Faster R-CNN as described in [19] did not change, but due to a reduced number of training samples, the number of iterations was reduced from 80,000 to 20,000. We use the refined networks of the tuning strategy (VGG16 (Feret and Faces94) and VGG16 (Facescrub)) on our image and video dataset to extract the descriptors and perform image and video face retrieval. Those results are presented in Table 2. This time the Manhattan similarity metric, also called city block, produced the best results. We should also note that the query expansion and spatial reranking slightly improved the results. When comparing the sum-pooling strategie to the max-pooling strategie of the image-wise and region-wise descriptors, sum-pooling gave better results than max-pooling with most similarity metrics. But max-pooling gave the best results when used with the Manhattan similarity metric with an accuracy of 76.2%.

We also compared different Faster R-CNN architectures trained on different datasets. We determined that deeper networks gave better results, which is consistent with the literature. We also noticed the datasets, on which the network was previously trained, had the most impact on the results. As we can see, when working with off-the-shelf networks, the networks trained on pascal VOC gave average results. But the best results were obtained when working with the networks trained for face classification, meaning trained on Fasecrub and Feret and Faces94 in our case. On that account, the VGG16 trained on Facescrub gave the best results because the nature of the photos in this dataset is more similar to the dataset that we are working on. Feret and Faces94 images were taken in a controlled environment, but Fasecrub images were amassed from the web and showcase the subject in different positions with different lighting setups and facial expressions which is closest to what videos can be. That is why the VGG16 trained on Facescrub gave the best results when used for retrieving face images and videos from a dataset of images and videos using one query image with a precision of 76.2%. So, we were able to improve the results with 13.7%.

Table 1. Mean average precision (mAP) of pre-trained Faster R-CNN models trained with microsoft COCO or pascal VOC

Metrics	Models	Pooling	Ranking	Re-ranking	QE
Cosine similarity metric	VGG16 (Pascal VOC)	sum	0.551	0.551	0.554
		max	0.538	0.545	0.544
	VGG16 (Microsoft COCO)	sum	0.545	0.521	0.516
		max	0.524	0.525	0.522
	ZF (Pascal VOC)	sum	0.550	0.539	0.538
		max	0.534	0.544	0.540
Euclidien similarity metric	VGG16 (Pascal VOC)	sum	0.551	0.551	0.554
		max	0.538	0.545	0.544
	VGG16 (Microsoft COCO)	sum	0.545	0.521	0.516
		max	0.524	0.525	0.522
	ZF (Pascal VOC)	sum	0.550	0.539	0.538
		max	0.534	0.544	0.540
Manhatan similarity metric	VGG16 (Pascal VOC)	sum	0.550	0.550	0.545
		max	0.540	0.543	0.538
	VGG16 (Microsoft COCO)	sum	0.543	0.513	0.507
		max	0.527	0.529	0.526
	ZF (Pascal VOC)	sum	0.547	0.535	0.530
		max	0.538	0.549	0.546
Chebychev similarity metric	VGG16 (Pascal VOC)	sum	0.497	0.482	0.493
		max	0.470	0.451	0.469
	VGG16 (Microsoft COCO)	sum	0.513	0.465	0.487
		max	0.488	0.437	0.453
	ZF (Pascal VOC)	sum	0.518	0.515	0.517
		max	0.499	0.459	0.490
Minkowski similarity metric	VGG16 (Pascal VOC)	sum	0.551	0.551	0.544
		max	0.538	0.545	0.544
	VGG16 (Microsoft COCO)	sum	0.545	0.521	0.516
		max	0.524	0.525	0.522
	ZF (Pascal VOC)	sum	0.550	0.544	0.536
		max	0.534	0.544	0.540
Canberra similarity metric	VGG16 (Pascal VOC)	sum	0.547	0.544	0.539
		max	0.528	0.516	0.518
	VGG16 (Microsoft COCO)	sum	0.538	0.516	0.512
		max	0.526	0.524	0.524
	ZF (Pascal VOC)	sum	0.540	0.538	0.537
		max	0.524	0.530	0.521
Corrolation similarity metric	VGG16 (Pascal VOC)	sum	0.551	0.551	0.544
		max	0.539	0.549	0.548
	VGG16 (Microsoft COCO)	sum	0.545	0.520	0.524
		max	0.524	0.522	0.517
	ZF (Pascal VOC)	sum	0.549	0.544	0.545
		max	0.537	0.542	0.537

4.5. Comparison

In this section we present a comparative study between our results and other results obtained using fisher vector (FV) and bag of visual word (BOVW). When working on video retrieval and image and video retrieval, our pipeline, which utilizes raw faster R-CNN features, outperformed all other techniques. The results are displayed in Table 3.

Table 2. Mean average precision (mAP) of the fine-tuned Faster R-CNN models with VGG16 architectures fine-tuned with Facescrub or Feret and Faces9 respectively

Metrics	Models	Pooling	Ranking	Re-ranking	QE
Cosine similarity metric	VGG16(Facescrub).	sum	0.757	0.737	0.706
		max	0.738	0.731	0.756
	VGG16(Feret and Faces94)	sum	0.577	0.570	0.563
		max	0.554	0.564	0.572
Euclidien similarity metric	VGG16(Facescrub).	sum	0.757	0.737	0.706
		max	0.738	0.731	0.756
	VGG16(Feret and Faces94)	sum	0.577	0.570	0.563
		max	0.554	0.564	0.572
Manhatan similarity metric	VGG16(Facescrub).	sum	0.738	0.695	0.734
		max	0.750	0.746	0.762
	VGG16(Feret and Faces94)	sum	0.565	0.561	0.553
		max	0.562	0.573	0.580
Chebychev similarity metric	VGG16(Facescrub).	sum	0.545	0.555	0.562
		max	0.564	0.579	0.605
	VGG16(Feret and Faces94)	sum	0.504	0.513	0.514
		max	0.495	0.501	0.500
Minkowski similarity metric	VGG16(Facescrub).	sum	0.757	0.727	0.747
		max	0.738	0.731	0.756
	VGG16(Feret and Faces94)	sum	0.577	0.570	0.560
		max	0.554	0.564	0.572
Canberra similarity metric	VGG16(Facescrub).	sum	0.742	0.742	0.760
		max	0.723	0.731	0.737
	VGG16(Feret and Faces94)	sum	0.567	0.569	0.568
		max	0.556	0.558	0.552
Corrolation similarity metric	VGG16(Facescrub).	sum	0.757	0.728	0.749
		max	0.741	0.731	0.748
	VGG16(Feret and Faces94)	sum	0.577	0.570	0.563
		max	0.557	0.568	0.573

Table 3. Comparative study with other techniques. Results provided as mAP

Method	YouTube Faces Database+Facescrub (an image and video dataset)	YouTube Faces Database (a video dataset)	FERET (an image dataset)
Our pipeline	0.762	0.903	0.8913
Faster R-CNN features+FV [21]	0.006	0.006	-
Faster R-CNN features+BOVW [21]	-	0.001	-
Log ICA II+KNN [38]	-	-	0.3553
Log ICA I+KNN [38]	-	-	0.3608
LGHP descriptor [7]	-	-	0.5460

5. CONCLUSION

In this paper, we demonstrate how to use CNN features from an object detection network for image and video face retrieval using one query image. We used Faster R-CNN features as our global and local descriptors in our end-to-end pipeline. We demonstrated that the best similarity metric to use with the off-the-shelf feature is the cosine similarity metric, and that the best one to use with refined networks is the Manhattan similarity metric. We also found that sum-pooling generally performs better, but when using the fine-tuned networks with the Manhattan similarity metrics, max-pooling gave the best results. We established that reranking strategies can improve the results. Most importantly, we proved that finetuned networks give the best results. So, when working on image and video face retrieval using one query image, we found the best results were obtained using a fine-tuned network combined with max-pooling, all our reranking strategies and using the Manhattan similarity metric. We determined that Finetuned CNN feature can give great results (76,2%) in real time (17 seconds per query image) when working on image and video face retrieval using a query image.

ACKNOWLEDGEMENTS

This work falls within the scope of Big Data and Connected Object (BDCO). We would like to thank the Hassan II University of Casablanca for financing this project.

REFERENCES

[1] D. Feng, M.-G. Liang, F. Gao, Y.-C. Huang, X.-F. Zhang, and L.-Y. Duan, “Towards large-scale object instance search: A multi-

block N-ary Trie," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 372–386, Jan. 2021, doi: 10.1109/TCSVT.2020.2966541.

[2] S. S. Tsai *et al.*, "Mobile product recognition," in *Proceedings of the international conference on Multimedia-MM '10*, 2010, Art. no. 1587, doi: 10.1145/1873951.1874293.

[3] A. Salvador, X. Giro-i-Nieto, F. Marques, and S. Satoh, "Faster R-CNN features for instance search," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2016, pp. 394–401, doi: 10.1109/CVPRW.2016.56.

[4] C.-W. Lin and S. Hong, "High-order histogram-based local clustering patterns in polar coordinate for facial recognition and retrieval," *The Visual Computer*, Mar. 2021, doi: 10.1007/s00371-021-02102-9.

[5] F.-C. Lin, H.-H. Ngo, and C.-R. Dow, "A cloud-based face video retrieval system with deep learning," *The Journal of Supercomputing*, vol. 76, no. 11, pp. 8473–8493, Nov. 2020, doi: 10.1007/s11227-019-03123-x.

[6] I. Hachchane, A. Badri, A. Sahel, and Y. Ruichek, "New faster R-CNN neuronal approach for face retrieval," in *Lecture Notes in Networks and Systems*, Springer International Publishing, 2019, pp. 113–120.

[7] S. R. Dubey, "Local directional relation pattern for unconstrained and robust face retrieval," *Multimedia Tools and Applications*, vol. 78, no. 19, pp. 28063–28088, Oct. 2019, doi: 10.1007/s11042-019-07908-3.

[8] L. Liu, J. Li, L. Niu, R. Xu, and L. Zhang, "Activity image-to-video retrieval by disentangling appearance and motion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 2145–2153.

[9] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 128–140, Jan. 2017, doi: 10.1109/TPAMI.2016.2537320.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[11] A. F. De Araujo, "Large-scale video retrieval using image queries a dissertation submitted to the department of electrical engineering and the committee on graduate studies of stanford university in partial fulfillment of the requirements for the degree of doctor of philos," 2016.

[12] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 4694–4702, doi: 10.1109/CVPR.2015.7299101.

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.2199.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 4489–4497, doi: 10.1109/ICCV.2015.510.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

[16] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2010, pp. 778–792.

[17] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," Springer International Publishing, 2014, pp. 584–599.

[18] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Lecture Notes in Computer Science*, Springer International Publishing, 2016, pp. 685–701.

[19] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 251–258, 2016, doi: 10.3169/mta.4.251.

[20] L. Wu, Y. Wang, Z. Ge, Q. Hu, and X. Li, "Structured deep hashing with convolutional neural networks for fast person re-identification," *Computer Vision and Image Understanding*, vol. 167, pp. 63–73, Feb. 2018, doi: 10.1016/j.cviu.2017.11.009.

[21] I. Hachchane, A. Badri, A. Sahel, and Y. Ruichek, "Large-scale image-to-video face retrieval with convolutional neural network features," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 1, pp. 40–45, Mar. 2020, doi: 10.11591/ijai.v9.i1.pp40-45.

[22] C. Zhang, B. Hu, Y. Suo, Z. Zou, and Y. Ji, "Large-scale video retrieval via deep local convolutional features," *Advances in Multimedia*, vol. 2020, pp. 1–8, Jun. 2020, doi: 10.1155/2020/7862894.

[23] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1224–1244, May 2018, doi: 10.1109/TPAMI.2017.2709749.

[24] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific rank fusion for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 803–815, Apr. 2015, doi: 10.1109/TPAMI.2014.2346201.

[25] S. Poullot, S. Tsukatani, A. Phuong Nguyen, H. Jégou, and S. Satoh, "Temporal matching kernel with explicit feature maps," in *Proceedings of the 23rd ACM international conference on Multimedia*, Oct. 2015, pp. 381–390, doi: 10.1145/2733373.2806228.

[26] D. M. Chen and B. Girod, "A hybrid mobile visual search system with compact global signatures," *IEEE Transactions on Multimedia*, vol. 17, no. 7, pp. 1019–1030, Jul. 2015, doi: 10.1109/TMM.2015.2427744.

[27] C. Herrmann and J. Beyerer, "Fast face recognition by using an inverted index," in *Procedings SPIE 9405, Image Processing: Machine Vision Applicationd VIII, 940507*, Feb. 2015, Art. no. 940507, doi: 10.1117/12.2078988.

[28] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2014, pp. 512–519, doi: 10.1109/CVPRW.2014.131.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[30] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[31] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, May 2017, pp. 650–657, doi: 10.1109/FG.2017.82.

[32] A. B. Yandex and V. Lempitsky, "Aggregating local deep features for image retrieval," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1269–1277, doi: 10.1109/ICCV.2015.150.

[33] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," *Computer Vision and Pattern Recognition*, Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.05879.

[34] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR 2011*, Jun. 2011, pp. 529–534, doi: 10.1109/CVPR.2011.5995566.

[35] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 343–347, doi: 10.1109/ICIP.2014.7025068.
[36] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, Apr. 1998, doi: 10.1016/S0262-8856(97)00070-X.
[37] D. L. Spacek, "Faces94 a face recognition dataset," 2007.
[38] M. K. Bhowmik, P. Saha, A. Singha, D. Bhattacharjee, and P. Dutta, "Enhancement of robustness of face recognition system through reduced gaussianity in Log-ICA," *Expert Systems with Applications*, vol. 116, pp. 96–107, Feb. 2019, doi: 10.1016/j.eswa.2018.08.047.

BIOGRAPHIES OF AUTHORS

Imane Hachchane is a Ph.D. student in Image processing at the EEA and TI Laboratory, Hassan II University Casablanca, Faculty of Sciences and Technology of Mohammedia (FSTM) in Morocco. She received her Software Engineering Degree from the National School of Applied Sciences of Kenitra, Morocco in 2016. She's currently working on Facial Large Scale Image Retrieval under the supervision of Pr. A. Badri. Her main research interest is to enhance the accuracy and speed of largescale image and video face retrieval using neural networks and deep learning. She can be contacted at email: hachchaneimane@gmail.com.

Abdelmajid Badri is a holder of a doctorate in Electronics and Image Processing in 1992 at the University of Poitiers–France. In 1996, he obtained the diploma of the authorization to Manage Researches (Habilitation à Diriger des Recherches: HDR) to the University of Poitiers–France, on the image processing. He is a director at the Higher School of Technology (EST) at Casablanca and he is a University Professor (PES-C) at the University Hassan II-Casablanca-Morocco (FSTM). He is a member of the laboratory EEA and TI (Electronics, Energy, Automatic and information Processing) which he managed since 1996. He managed several doctoral theses. He is a co-author of several national and international publications. He is responsible for several research projects financed by the ministry or by the industrialists. He was member of several committees of programs of international conferences and president of three international congresses in the same domain. He is a member and co-responsible in several scientific associations in touch with his domain of research. He can be contacted at email: abdelmajid_badri@yahoo.fr.

Aïcha Sahel is a holder of a doctorate in Electronics and Image Processing in 1996 at the University of Poitiers-France. She is a university Professor at the University Hassan II-Casablanca-Morocco (FSTM) She is a member of the laboratory EEA and TI. The research works of A. Sahel concern the Communication and Information Technology (Electronics Systems, Signal/Image Processing and Telecommunication). She co-supervises doctoral theses and she is a co-author of several national and international publications. She is a member in financed research projects. She was a member of steering committees of three international congresses in the same domain of research. She can be contacted at email: sahel_ai@yahoo.fr.

Ilham Elmourabit is a holder of a doctorate in Telecommunication and information engineering in 2011 at the University Hassan II-Casablanca-Morocco (FSTM). She is a university Professor at the Hassan II University Casablanca, Faculty of Sciences and Technology of Mohammedia (FSTM) in Morocco. She is a member of the laboratory EEA and TI. The research works of I. Elmourabit concern the Communication and Information Technology. She co-supervises doctoral theses and she is a co-author of several national and international publications. She can be contacted at email: elmourabit.ilham@gmail.com.

Yassine Ruichek (Senior Member, IEEE) received the Ph.D. degree in control and computer engineering and the Habilitation à Diriger des Recherches (HDR) degree in physic science from the University of Lille, France, in 1997 and 2005, respectively. Since 2007, he has been a Full Professor with the University of Technology of Belfort-Montbéliard (UTBM). His research interests include computer vision, image processing and analysis, pattern recognition, data fusion, and localization, with applications in intelligent transportation systems and video surveillance. He can be contacted at email: yassine.ruichek@utbm.fr.