Video retrieval with CNN features

Imane Hachchane, Abdelmajid Badri, Aïcha Sahel, Yassine Ruichek

Video retrieval with CNN features

Imane HACHCHANE[1] , Abdelmajid BADRI[1], Aïcha SAHEL[1] and Yassine RUICHEK[2]
hachchaneimane@gmail.com
[1]Laboratoire d'Electronique, Energie, Automatique & Traitement de l'Information (EEA&TI), Faculté des Sciences et Techniques Mohammedia, Université Hassan II Casablanca, Mohammedia, Morocco.
[2] IRTES-Laboratoire SET, Université de Technologie de Belfort Montbéliard, France, Belfort

ABSTRACT : Convolutional neural network features are becoming the norm in instance retrieval. This work investigate the relevance of using an of the shelf object detection network like Faster R-CNN as a feature extractor. We build an Image-to-video face retrieval pipeline composed of filtering and re-ranking that uses the objects proposals learned by a Region Proposal Network (RPN) and their associated representations taken from a CNN. Moreover we study the relevance of features from a finetuned network. The results obtained are very promising.

keywords : Image Processing, Classification, Object Recognition, CNN, Faster R- CNN, Image-To-Video Instance Retrieval, Face Retrieval, Video Retrieval.

1 INTRODUCTION

Visual search applications especially video retrieval have gained a vast popularity recently due to the explosion of visual content that we are witnessing nowadays. This increase led to a proliferation of visual search applications like instance search. This is used to retrieve images or videos of a specific object from large databases. This work addresses a variant of this problem, the task of image-to-video instance retrieval witch is the task of identifying a video collection from a specific instance in a static image.

Traditionally, image-to-video retrieval methods[1]–[3] are based on hand-crafted features (SIFT [4], BRIEF[5] , etc.) and not much effort has been put so far into the adaptation of deep learning techniques, such as convolutional neural networks (CNN).

CNNs trained with large amounts of data can learn features generic enough to be used to solve tasks for which the network has not been trained[6] .For image retrieval, in particular, many works in the literature [7], [8] have adopted solutions based on standard features extracted from a pretrained CNN for image classification[9], achieving encouraging performances.

In this paper we try to fill this gap by exploring the relevance of on-the-shelf and fine-tuned features of an object detection CNN for image-to-video face retrieval.

2 RELATED WORK

Most work in visual search focus on image to image retrieval, were we use a query image and a database of images[10], [11] but this work focus on image to video retrieval were we search a database of videos using query images. But more precisely we are focusing on instance face retrieval.

Face retrieval remains a challenging task because conventional image retrieval approaches, such as bag of words, are difficult to adapt to the face domain[12]. This is mainly the result of using the traditional key point detection based descriptors like SIFT, that have tendency to fail due the smooth face surface. Early works, using a pretrained image classification convolutional neural network as a feature extractor, showed that fully connected layers for image retrieval were more suitable [13]. Razavian et al. [14], improved the results by combining fully connected layers extracted from different image sub-much. Later on, new works found that using convolutional layers significantly outperform fully connected layers during image recovery tasks [14], [15].

Many CNN-based object detection pipelines have been proposed, but we are more interested in the latest ones. Faster R-CNN [16] created by Ren et al. it uses a Region Proposal Network (RPN) that removes the dependence of object proposals from older CNN object detection systems. In Faster R-CNN, RPN shares features with the object-detection network in [17] to simultaneously learn prominent object propositions and their associated class probabilities. Although the Faster R-CNN is designed for generic object detection, Jiang et al. [18] Demonstrated that it can achieve impressive face detection performance specially when retrained on a suitable face detection training set[19].

In this work we exploit the features of a state of the art pre-trained object detection CNN Faster R-CNN. We use his end-to-end object detection architecture to extract global and local convolutional features in a single forward pass and test their relevance for image-to-video face retrieval.

3 METHODOLOGY

3.1 CNN-based Representations

We explore the relevance of using CNN features for face image to video face retrieval. The query instance is defined by a bounding box above the query image. We use the features extracted from Faster R-CNN pre-trained models[16] as our global and local features. Faster R-CNN has a region proposal network that give the locations in the image that have bigger probabilities of having an object, and a classifier that labels each of those object proposals as one of the classes in the learning dataset[15]. We will extract compact features from the activations of a convolutional layer in a CNN [15], [20]. Faster R-CNN is faster on a global and local scale. We build a global frame descriptor by ignoring all the layers

that works with object proposals and extract features from the last convolutional layer. Considering the extracted activations of a convolution layer for a frame, we group the activations of each filter to create a frame descriptor with the same dimension as the number of filters in the convolution layer. We aggregate the activations of each window suggestion in the RoI Pooling layer to create regional descriptions[19].

3.2 Video retrieval

This section describes the three ranking strategies we used:

Filtering step. We create image descriptors for query and database frames. At testing time, the descriptor of the query is compared to all items in the database, which are then ranked according to a similarity measure. At this stage, the entire frame is considered as a query.

Spatial re-ranking. After the filtering step, the N upper elements are analyzed locally and re-ranked.

Query expansion (QE). We average the frame descriptors of the N higher elements of the first ranking with query descriptor to carry out a new search.

3.3 Fine-tuning Faster R-CNN

Fine tuning the Faster R-CNN network allows as to obtain features specific to face retrieval and should help improve the performance of spatial analysis and re-ranking. To achieve this, we choose to fine-tune Faster R-CNN to detect the query faces to be retrieved by our system. The resulting networks will be used to extract better local and global representations, and will be used to perform spatial reranking.

4 EXPERIMENTS

4.1 Datasets exploited

We evaluate our methodologies using the following datasets:

- YouTube Celebrities Face Tracking and Recognition Data (Y-Celeb) [21] : The dataset contains 1910 sequences of 47 subjects. All videos are encoded in MPEG4 at 25fps rate.
- YouTube Faces Database [22] : The data set contains 3,425 videos of 1,595 different people. All the videos were downloaded from YouTube. An average of 2.15 videos are available for each subject. The shortest clip duration is 48 frames, the longest clip is 6,070 frames, and the average length of a video clip is 181.3 frames.

The datasets used to finetune the network:

- FERET [23]: 3528 images, including 55 Query images. A framing box surrounding the target face is provided for query images.
- FACES94 [24]: 2809 images 2809 images, including 55 Query images. A framing box surrounding the target face is provided for query images.
- FaceScrub [25]: 55127 images.

4.2 Experimental Setup

We use the VGG16 architecture of Faster R-CNN to extract the global and local features. We chose the VGG16 architecture because it performs better. That has been showmen in previous works in the literature [15], [19]were the capabilities of deeper networks achieve better performance. The global descriptors are extracted from the last convolution layer "conv5_3" and are of dimension 512. The local features are grouped from the Faster R-CNN RoI clustering layer. All experiments were performed on a Nvidia GTX GPU.

4.3 Off-the-shelf Faster R-CNN features

We carried out a comparative study of the sum and max-pooling strategies of the image-wise and region-wise descriptors. Table 1 summarizes most our results. According to our experiments, the sum-pooling gives better performance than the max-pooling. It also shows the performance of Faster R- CNN with a VGG16 architectures trained on two different datasets (Pascal VOC and COCO), VGG16 trained on COCO performed better because the dataset is bigger and more diverse . Moreover, it present the impact of spatial reranking and query expansion. Using the global features of Faster R-CNN on their own without any reranking strategy give the best results. Spatial reranking & QE had no impact or a negative one on the results

Table 1 Mean Average Precision (MAP) of pre-trained Faster R-CNN models with VGG16 archi-tectures. (P) and (C) denote whether the network was trained with Pascal VOC or Microsoft COCO images, respectively. With a comparison between sum and max pooling strategies. When indicated, QE is applied with M = 5

Network	Pooling	Y-Celeb			YouTube Faces Database		
		Ranking	***Reranking***	***QE***	***Ranking***	***Reranking***	***QE***
VGG16 (P)	max	0.888	0.860	0.550	0.892	0.877	0.882
	sum	0.915	0.846	**0.600**	0.897	0.886	0.891
VGG16 (C)	max	0.911	**0.888**	0.522	0.892	0.878	0.889
	sum	**0.926**	0.807	0.512	**0.903**	0.882	**0.896**

Table 2 Mean Average Precision (MAP) of pre-trained Faster R-CNN models with VGG16 archi-tectures. (F-S) AND (F-F) denote whether the network was trained with FaceScrub or Feret & Faces94 images, respectively. With a comparison between sum and max pooling strategies. When indicated, QE is applied with M = 5

Network	Pooling	Y-Celeb			YouTube Faces Database		
		Ranking	***Reranking***	***QE***	***Ranking***	***Reranking***	***QE***
VGG16 (F-S)	max	0.809	0.777	0.457	0.848	0.834	0.838
	sum	0.917	0.843	0.578	0.882	0.873	0.874
VGG16 (F-F)	max	0.915	0.874	0.554	0.894	0.884	0.887
	sum	**0.924**	**0.899**	**0.621**	**0.896**	**0.892**	**0.893**

4.4 Fine-tuning Faster R-CNN

We evaluate the impact of fine-tuning a pre-trained network on recovery performance with the query objects to retrieve. We chose to refine the model VGG16 Faster R-CNN, pre-trained with the objects of Pascal VOC, with two deferent datasets. The first network was refined using FERET and Faces94 datasets, we combine them to create one dataset. We modify the output layer in the network to return 422 class probabilities (269 people in the FERET dataset plus 152 people in the Faces94 dataset, plus one additional class for the background) and their corresponding bounded bound box coordinates[19]. This new refined network will be called VGG(F-F). The second network was refined using FaceScrub dataset. we modify the output layer in the network to return 530 class probabilities (530 people, plus one additional class for the background)
and their corresponding bounded bound box coordinates. Our second refined network will be called VGG(F-S)[19]
We kept the Faster R-CNN original parameters described in [19], but due to the our smaller number of training sample we decreased the number of iterations from 80,000 to 20,000.
We use the refined networks of the tuning strategy (VGG(F-F) & VGG(F-S)) on all datasets to extract image and region descriptors to perform a face retrieval. Those results are also presented in Table 2. The refined features slightly exceeded the raw feature in the spatial reranking and the QE stages. But still, the global features of Faster R-CNN from VGG16 trained on COCO used without any reranking strategy give the best results.

5 CONCLUSION

This article explores the use of features from an object detection CNN for image-to-video face retrieval. It uses Faster R-CNN features as global and local descriptors. We have shown that the common similarity metric give similar results. We also found that sum-pooling performs better than max-pooling in this case, and contrary to our previous work [19] fine tuning does not improve the results. In general we found that applying the similarity measure on the CNN feature gave the best results. In future work we will work on reducing the feature extraction time.

Bibliographie

[1] R. Arandjelovic and A. Zisserman, “Three things everyone should know to improve object retrieval,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2911–2918.

[2] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, “Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation,” Mar. 2015.

[3] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” Nov. 2014.

[4] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[5] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, “BRIEF: Binary Robust Independent Elementary Features,” Springer, Berlin, Heidelberg, 2010, pp. 778–792.

[6] J. Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 4694–4702, 2015.

[7] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, “Visual Instance Retrieval with Deep Convolutional Networks,” Dec. 2014.

[8] A. Araujo and B. Girod, “Large-Scale Video Retrieval Using Image Queries,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. XX, no. c, pp. 1–1, 2017.

[9] G. De Oliveira Barra, M. Lux, and X. Giro-I-Nieto, “Large scale content-based video retrieval with LIvRE,” *Proc. - Int. Work. Content-Based Multimed. Index.*, vol. 2016-June, 2016.

[10] L. Zheng, Y. Yang, and Q. Tian, “SIFT Meets CNN: A Decade Survey of Instance Retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, 2018.

[11] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, “Query Specific Rank Fusion for Image Retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 803–815, Apr. 2015.

[12] C. Herrmann and J. Beyerer, “Fast face recognition by using an inverted index,” 2015, vol. 9405, p. 940507.

[13] A. Babenko, A. Slesarev, A. Chigorin, and V.

Lempitsky, “Neural Codes for Image Retrieval,” Apr. 2014.

[14] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN Features off-the-shelf: an Astounding Baseline for Recognition,” Mar. 2014.

[15] A. Salvador, X. Giro-I-Nieto, F. Marques, and S. Satoh, “Faster R-CNN Features for Instance Search,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 394–401.

[16] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[17] R. Girshick, “Fast R-CNN,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[18] H. Jiang and E. Learned-Miller, “Face Detection with the Faster R-CNN,” *Proc. - 12th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2017 - 1st Int. Work. Adapt. Shot Learn. Gesture Underst. Prod. ASL4GUP 2017, Biometrics Wild, Bwild 2017, Heteroge*, pp. 650–657, 2017.

[19] 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*, vol. 66, 2019, pp. 113–120.

[20] G. Tolias, R. Sicre, and H. Jégou, “Particular object retrieval with integral max-pooling of CNN activations,” Nov. 2015.

[21] Minyoung Kim, S. Kumar, V. Pavlovic, and H. Rowley, “Face tracking and recognition with visual constraints in real-world videos,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[22] L. Wolf, T. Hassner, and I. Maoz, “Face recognition in unconstrained videos with matched background similarity,” in *CVPR 2011*, 2011, pp. 529–534.

[23] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, “The FERET database and evaluation procedure for face-recognition algorithms,” *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, Apr. 1998.

[24] D. L. Spacek, “Faces94 a face recognition dataset,” 2007.

[25] H.-W. Ng and S. Winkler, “A data-driven approach to cleaning large face datasets,” in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 343–347.