Large-scale image-to-video face retrieval with convolutional neural network features

Imane Hachchane¹, Abdelmajid Badri², Aïcha Sahel³, Yassine Ruichek⁴

¹,²,³Laboratory of Electronics, Energy, Automation & Information Processing, Faculty of Sciences and Techniques
Mohammedia, University Hassan II Casablanca, Mohammedia, Morocco
⁴IRTES-Laboratory SET, University of Technology Belfort Montbéliard, France

ABSTRACT
Convolutional neural network features are becoming the norm in instance retrieval. This work investigates the relevance of using an of the shelf object detection network, like Faster R-CNN, as a feature extractor for an image-to-video face retrieval pipeline instead of using hand-crafted features. We use the objects proposals learned by a Region Proposal Network (RPN) and their associated representations taken from a CNN for the filtering and the re-ranking steps. Moreover, we study the relevance of features from a finetuned network. In addition to that we explore the use of face detection, fisher vector and bag of visual words with those same CNN features. We also test the impact of different similarity metrics. The results obtained are very promising.

Keywords: BOVW, Classification, CNN, Faster R-CNN, Face retrieval, FV, Image processing, Image-to-video instance retrieval, Object recognition, Video retrieval

1. INTRODUCTION
In this work we address the task of image to video face retrieval. With billions of images and videos created each day, it is essential to build tools for accessing and retrieving multimedia content efficiently. In the context of retrieval, image-to-video face retrieval is the task of identifying a specific frame or scene in a video or a collection of videos from a specific face instance in a static image.

On one hand, image-to-video retrieval is an asymmetric problem. Images only contain static information but videos have much richer visual information, like optical flow. Due to the lack of temporal information, standard techniques used for extracting video descriptors [1-4] cannot be directly used on static images. But, standard features for image retrieval [5-8] can be applied to video data by processing each frame as an independent image. Temporal information is usually compressed either by reducing the number of local features or by encoding multiple frames into a single global representation. On the other hand, face retrieval remains a challenging task because conventional image retrieval approaches, such as bag of visual words (BOVW), are difficult to adapt to the face domain [9].

Traditionally, image-to-video retrieval or face retrieval methods [10-12] are based on hand-crafted features (SIFT [13], BRIEF[14], 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 [15]. For image retrieval, in particular, many works in the literature [7, 16] have adopted solutions based on standard features extracted from a pretrained CNN for image classification [17], achieving encouraging performances. Many CNN-based object detection pipelines have been proposed, but we are more interested in the latest ones. Faster R-CNN [18] 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 [19] to simultaneously learn prominent object propositions and their associated class probabilities. Although the Faster R-CNN is designed for generic object detection [20], demonstrated that it can achieve impressive face detection performance especially when retrained on a suitable face detection training set [21].

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. We exploit the features of a state-of-the-art pretrained object detection CNN called 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. We also explore the use of face detection, Fisher Vector (FV) [4] and BOVW words with those same CNN features. The rest of this paper is organized as follows: Section 2 presents our research method, including our features extraction method and the ranking and re-ranking strategies. Section 3 presents our results and discussions. Finally, we present our conclusions in Section 4.

2. METHODOLOGY
2.1. Datasets exploited

We evaluate our methodologies using the following datasets:
- YouTube Celebrities Face Tracking and Recognition Data (Y-Celeb) [22]: The dataset contains 1910 sequences of 47 subjects. All videos are encoded in MPEG4 at 25fps rate.
- YouTube Faces Database [23]: 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 fine-tune the network:
- FERET [24]: 3528 images, including 55 Query images. A framing box surrounding the target face is provided for query images.
- FACES94 [25]: 2809 images 2809 images, including 55 Query images. A framing box surrounding the target face is provided for query images.
- FaceScrub [26]: 55127 images

2.2. Video retrieval strategy:

This section describes the three major steps in our pipeline, we used:
1. Filtering step. We create image descriptors for query and database frames using CNN features. 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.
2. Spatial re-ranking. After the filtering step, the N upper elements are analyzed locally and re-ranked.
3. Query expansion (QE). We average the frame descriptors of the M higher elements of the first ranking with query descriptor to carry out a new search.

2.3. 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 [18] as our global and local features. Faster R-CNN has a region proposal network that gives the locations in the image which have higher probabilities of having an object, and a classifier that labels each of those object proposals as one of the classes in the learning dataset [27]. We extract compact features from the activations of a convolutional layer in a CNN [27-28]. Faster R-CNN is faster on a global and local scale. We build a global frame descriptor by ignoring all the layers that work 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, to do so both max and sum pooling strategies are considered and compared in section 3. We aggregate the activations of each window suggestion in the RoI Pooling layer to create regional descriptions [21].
We use the VGG16 architecture of Faster R-CNN to extract the global and local features. We choose that architecture because it performs better. It has been shown in previous works in the literature [21, 27] that 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.

2.4. 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. The resulting networks will be used to extract better local and global representations, and will be used to perform spatial reranking.

We chose to refine the model VGG16 Faster R-CNN, pre-trained with the objects of Pascal VOC, with two different datasets. The first network was refined using FERET and Faces94 datasets, we combine them to create one bigger 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 [21]. This new refined network will be called VGG(F-F), the training process took 2 hours 47 minutes. 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)[21], the training took 2 hours 30 minutes.

We kept the Faster R-CNN’s original parameters described in [19], but due to our smaller number of training samples we decreased the number of iterations from 80,000 to 20,000. We use the refined networks of the tuning strategy (VGG(F-S) & VGG(F-F)) on all datasets to extract image and region descriptors to perform a face retrieval.

2.5. Faster R-CNN features & Face detection

We evaluate the use of Faster R-CNN for feature extraction and the ranking and reranking strategies as described previously.

2.6. Faster R-CNN features & FVs

To explore the relevance of using FVs on CNN feature, for the image-to-video face retrieval task, we first extract the CNN features of each frame. We then apply Principal Component Analysis (PCA), Gaussian mixture model (GMM), L2 normalization on those features before using our FV function. Finally, as described before, we compute the similarity measure and use the ranking and reranking strategies.

2.7. Faster R-CNN features & BOVW

To explore the relevance of using BOVW with CNN feature, for the image-to-video face retrieval task, we first extract the CNN features of each frame. Then we apply the clustering, vector quantization and inverted indexing steps. Finally, as described before, we compute the similarity measure and use the reranking strategies.

3. RESULTS AND DISCUSSION

We evaluate the use of Faster R-CNN features for face image to video face retrieval. We experimented with six different similarity metrics. The results were similar and close but overall cosine performed better. Table 1 shows an example of our results when using features from an on the shelf network with VGG16 architecture trained on pascal dataset.

We carried out a comparative study of the sum and max-pooling strategies of the image-wise and region-wise descriptors. Table 2 summarizes most of 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 presents the impact of spatial reranking and query expansion. Using the global features of Faster R-CNN on their own without any reranking strategy gives the best results. Spatial reranking & QE had no positive impact on the results. We should note that in average the offline feature extraction took 29.7 minutes while the online ranking steps took 3.7 seconds and the reranking strategy took 7 minutes for Y-Celeb dataset. For YouTube Faces Database, the offline feature extraction took 20 hours while the online ranking steps took only 85 seconds and the reranking strategy took 21 minutes.
Large-scale image-to-video face retrieval with convolutional neural network features (Imane Hachchane)
Table 4. Mean Average Precision (mAP) of pre-trained Faster R-CNN models with VGG16 architectures. (C) denote that the network was trained with Microsoft COCO. With a comparison between sum and max pooling strategies. When indicated, QE is applied with $M = 5$.

<table>
<thead>
<tr>
<th>Network</th>
<th>Pooling</th>
<th>Ranking</th>
<th>Y-Celeb Reranking</th>
<th>QE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG16 (C)</td>
<td>max</td>
<td>0.032</td>
<td>0</td>
<td>0.097</td>
</tr>
<tr>
<td></td>
<td>sum</td>
<td>0.032</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Finally, we can clearly see in that the raw faster R-CNN features largely outperformed the other strategies with a mAP of 92.6%. Table 5 show comparison with State-of-the-art.

Table 5. Comparison with State-of-the-art. Results provided as mAP.

<table>
<thead>
<tr>
<th>Method</th>
<th>Y-Celeb</th>
<th>YouTube Faces Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN [23]</td>
<td>-</td>
<td>0.145</td>
</tr>
<tr>
<td>O-SBoF [29]</td>
<td>-</td>
<td>0.471</td>
</tr>
<tr>
<td>RN-BOF [30]</td>
<td>-</td>
<td>0.465</td>
</tr>
<tr>
<td>Faster R-CNN features</td>
<td>0.926</td>
<td>0.903</td>
</tr>
<tr>
<td>Faster R-CNN features + FV</td>
<td>0.097</td>
<td>0.006</td>
</tr>
<tr>
<td>Faster R-CNN features + BOVW</td>
<td>0.032</td>
<td>0.001</td>
</tr>
</tbody>
</table>

4. 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 gives similar results. We also found that sum-pooling performs better than max-pooling in most cases, and contrary to our previous work [21] fine tuning does not improve the results. More importantly, we found that applying the similarity measure directly on the CNN feature of an off-the-shelf CNN trained on a large and diverse dataset gave the best results, and that using FVs or BOVW is memory consuming and is not suitable for CNN features in this case.

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

Large-scale image-to-video face retrieval with convolutional neural network features (Imane Hachchane)