Project proposal

1. Problem statement

Content based image retrieval (CBIR) has been an active topic in the last decades. One of the most important issues in image retrieval is the feature extraction and representation. Color, texture and shape features are the most straightforward ones in image feature extraction, such as color histogram/color correlograms, grey level co-occurrence matrices(GLCM), etc. Besides, most existing approaches have adopted the low-level visual features such as SIFT descriptors via bag-of-words encoding, vector locally aggregated descriptors (VLAD) or Fisher vectors (FV).

Recently, convolutional neural network (CNN) has become a powerful tool for tackling on the image classification problems. By training multiple layers of convolutional filters, CNNs are able to learn the complex features of the images automatically and demonstrated superior performance compared to the low-level features.

Nevertheless, there are still problems unaddressed in the implementation of image retrieval with CNNs. Firstly, the CNNs by default are trained for classification tasks which utilize the features obtained from the final layer for the decision making, however, this procedure does not preserve the local characteristics of the objects at the instance level well. And this has questioned that whether it is best to directly extract features from the final layer or higher layers for instance-level image retrieval, where the different objects from the same category should be separated. Secondly, the assumption that all the test images have the same size and scale is not valid. Therefore, the question is left that how to deal with the images with different scales and sizes when they are passing through the neural network.

2. Methodologies

2.1 Convolutional neural network

Generally the goal is applicable to various convolutional neural network architectures. There are multiple outstanding deep neural networks: GoogLeNet, OxfordNet, Tensorflow, etc.

The GoogLeNet is a 22-layer deep convolutional network which takes 224×224 pixels as input and then pass them through multiple convolutional layers and stacking the inception modules.

2.2 Convolutional features extraction

The final features $\{F^1, F^2, ..., F^L\}$ are obtained by processing the input image throughout the network.

2.3 VLAD encoding

During instant image retrieval there are no training data available. In this situation, a pre-trained network may fail to produce representations that are invariant to translation or any other viewpoint changes. The local features can be a good candidate for representing instance level information while generalizing to other object categories.

The VLAD encoding is effective for encoding local features into a single descriptor while achieving a favorable trade-off between retrieval accuracy and memory footprint.

2.4 Image retrieval

After extracting the convolutional features and encoding them into VLAD descriptors, image retrieval can be done by calculating the L2 distance between the VLAD descriptors of the query image and the images in the database. PCA is needed to compress the original VLAD descriptors to lower dimension for efficiency.

3. Datasets and Experiments

Various instance-level image datasets will be used:

- Holidays dataset: 1491 images of 500 categories
- Oxford dataset: 5062 images of landmarks in Oxford
- Paris dataset: 6412 images of Paris

For simplicity, the full images will be used for retrieval in the project rather than the queries in the datasets with specific rectangular regions which are the instances of interest to be retrieved.

The works will be within the scope of:

- performance of convolutional features from different layers
- the impact of changing scales
- the comparison between uncompressed representation and the low-dimensional feature extraction

References

[1] Joe Yue-Hei Ng et al., Exploiting Local Features from Deep Networks for Image Retrieval, 2015

[2] Ricardo da Silva Torres et al., Content-Based Image Retrieval: Theory and Applications, 2006