Extracting keywords from images

Bag-of-visual-words enriched with graph techniques

Gjorgji Madjarov[1] and Sanda Martincic-Ipsic[2]

[1]FCSE, Ss. Cyril and Methodius University, Skopje, Macedonia

[2]Department of Informatics, University of Rijeka, Rijeka, Croatia

What are the keywords and the task of keyword extraction?

- Keywords are the important topics in one content and can be used to index data, generate tag clouds or for searching

- Keywords have become primary means for searching information in documents, images and videos on the WWW

- The task of keyword extraction is to automatically identify a set of terms that best describe the document

Keyword extraction

- State-of-the-art keyword extraction approaches are based on statistical methods which require learning from hand-annotated data sets

- Now, the focus of research has shifted toward unsupervised methods, mainly network or graph enabled keyword extraction

Keyword extraction in text document

- In a network (graph) based keyword extraction the source (document, text, specific data etc.) is transformed into network of:
  - words - nodes of the network and
  - relations - represented with links
- Two words are linked if they are adjacent in a window of maximum *n* words
- Links are weighted according to the co-occurrence frequencies of the words they connect

Keyword extraction in text document

- Graph-based methods for keyword extraction
  - do not require advanced linguistic knowledge or processing,
  - are domain independent
  - are language independent

What about images?

- State-of-the-art methods use Bag of Visual Words (BoVW) representation of images.
- In BoVW models, a vocabulary (or codebook) of visual words is obtained by clustering local image descriptors extracted from images.
- An image is then represented as a BoVW, which is a sparse vector of occurrence counts of the visual words in the vocabulary.

Keyword extraction in images

- We want to represent the images as a complex network of linked visual words:
  - each individual visual word could be a node and interactions amongst visual words could be links
  - co-occurrence networks exploit *global location costs* of visual words and the *adjacency cost* of local descriptors in the database as weights of the links between the visual words
    - Those metrics were proposed on CVPR 2014 for image reconstruction from BoVW*

* Kato, H.; Harada, T., "Image Reconstruction from Bag-of-Visual-Words," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* , vol., no., pp.955-962, 23-28 June 2014

Keyword extraction in images

- Adjacency cost is defined as the negative logarithm of the normalized histogram of co-occurrences of pairs of visual words in a neighboring region
  - Only pairs which are in *m*-neighbor distance (window *n* x *n* = *m*) are taking into account and their relative positions are using.
- Global location cost is defined as the negative logarithm of the normalized histogram of the occurrence of a certain visual word at a certain location

Potential outcome

- Using network model and measures used in graph theory, we can represent the images on higher level (e.g. construct a layer with a semantic view of the image)
- We expect to identify representative parts of images, patterns or even detect and describe objects in the images.
- In this case the keyword extraction, representation, retrieval, clustering, searching of the images could be improved.