Combining Text and Image Processing in an Automatic Image Annotation System

Iulian Ilieș (SHSS, Jacobs University)

Joint work with Arne Jacobs, Otthein Herzog (TZI, Universität Bremen), and Adalbert Wilhelm (SHSS, Jacobs University)

Supported by the Deutsche Forschungsgemeinschaft (DFG)

Overview

- Motivation and approach
- Current work:
  - Framework of concept propagation
  - Data and algorithms employed
  - Comparison of different classifiers
  - Effect of visual vocabulary size
- Summary and outlook

Motivation

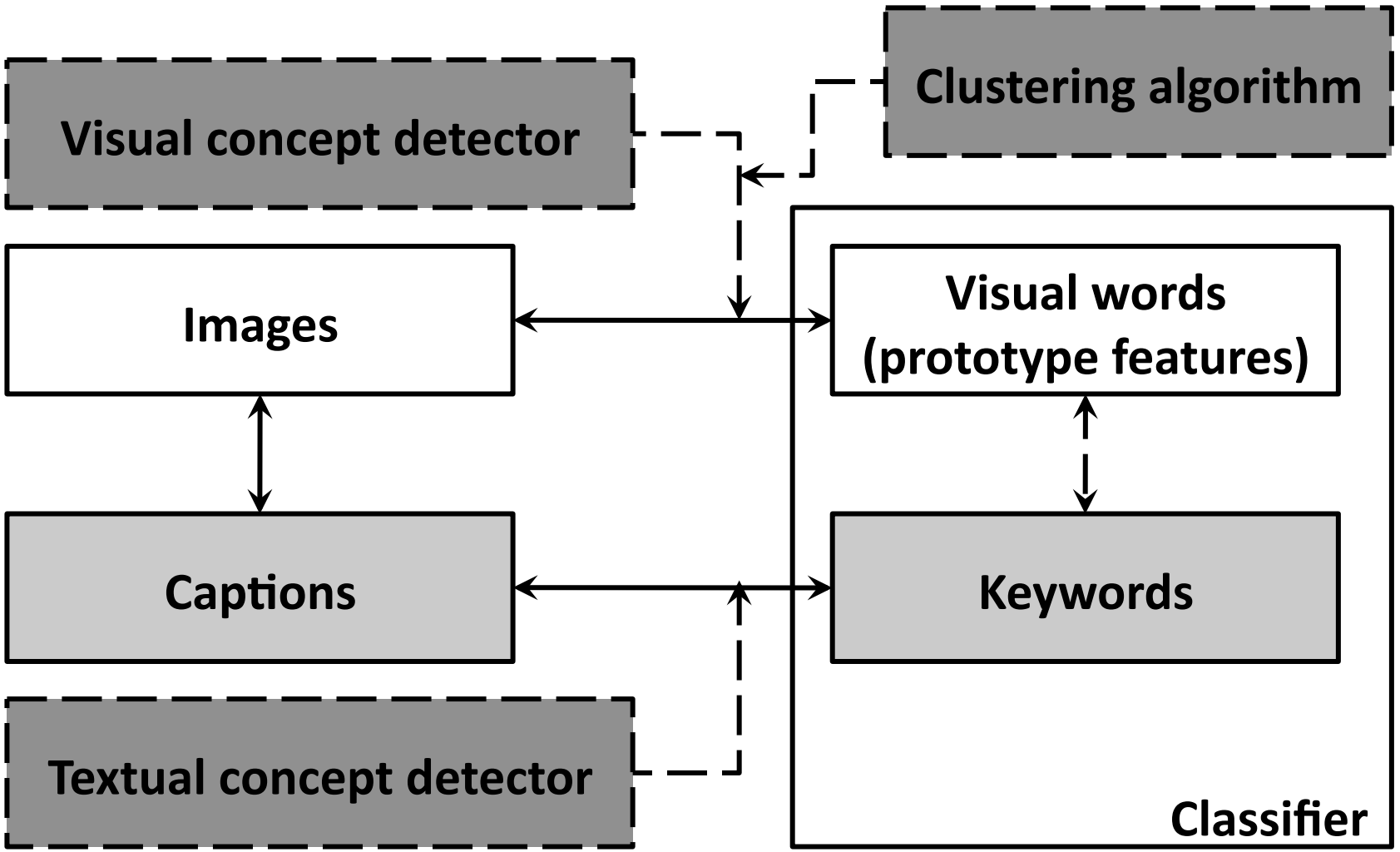
- Continuously increasing quantity of image data available on the Internet, which necessitates efficient classification and indexing methods for easy access and usage

- Existing methods, especially mainstream, do not exploit all available information:
  - Text-based search, using file names and/or captions
  - Pure visual search, relying only on image features
  - Semantic search, via image understanding techniques

Approach

- Combine the advantages of these different viewpoints into an integrated framework, which would allow the classification of images using keywords, features, or both
- Focus on the construction of a dual-layered linkage scheme between images, based on the co-occurrence of keywords, and on similarities between visual features
- Define visual words, and associate them to keywords

Framework


Clustering algorithm
Visual concept detector
Images
Visual words
(prototype features)
Captions
Keywords
Textual concept detector
Classifier

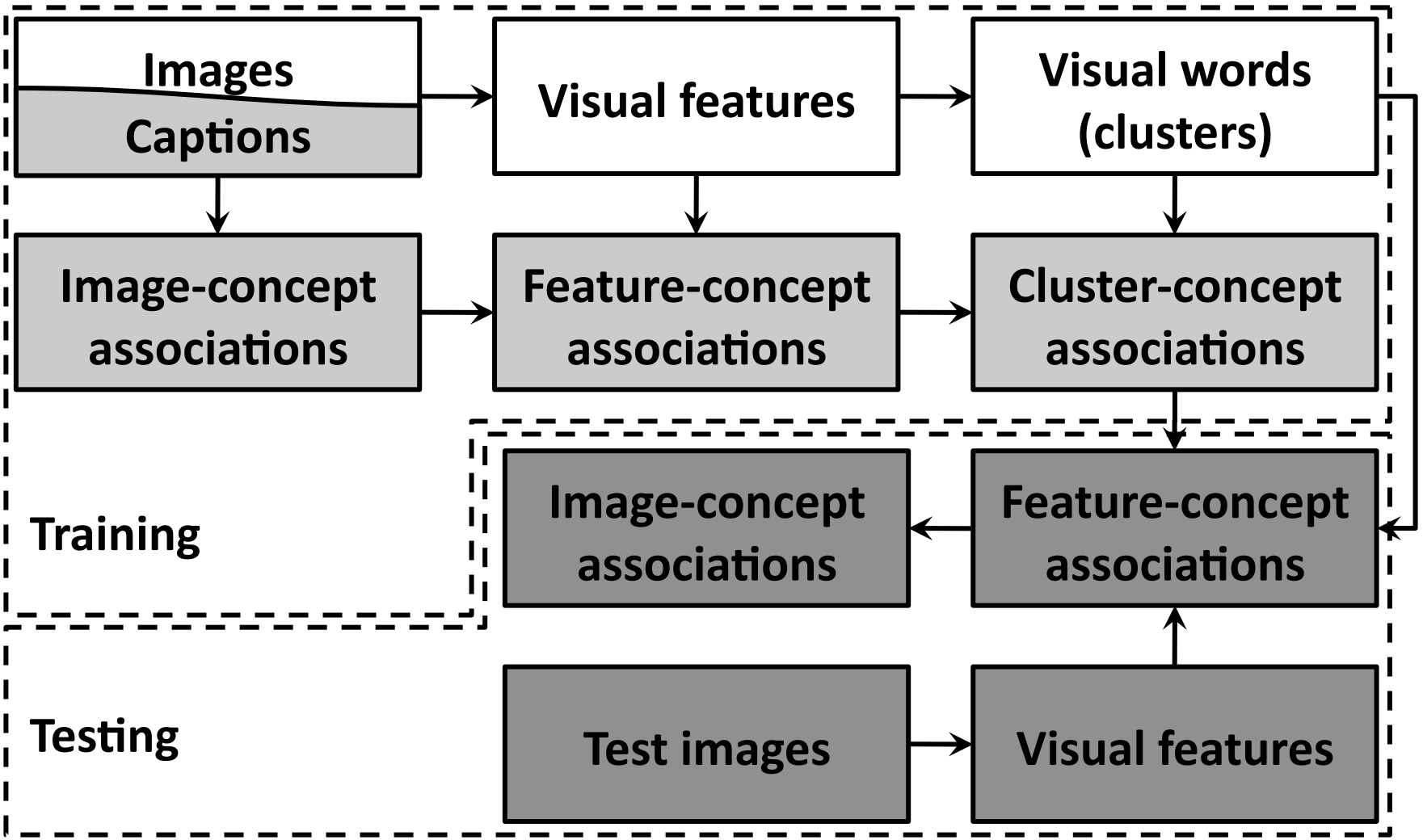
Concept propagation

- Directly transfer the associations with keywords from captions to related images, and further to the visual features found in these images

- For each visual word, average across the visual features that have it as prototype, and contrast the obtained value with the corresponding global average

  - These operations can be performed in reversed order!

Classifier


Images
Captions
Visual features
Visual words (clusters)
Image-concept associations
Feature-concept associations
Cluster-concept associations
Training
Image-concept associations
Feature-concept associations
Testing
Test images
Visual features

Data employed

- Images and related text (e.g. captions, titles) harvested from news websites

- Strongly structured articles, that can be parsed automatically

Concept detectors

- Specialized keyword detector:
  - Person names extracted from captions by a named entity recognizer (NER; Drozdzynski et al. 2004), complemented by manual annotations

- Generic visual feature detector:
  - Interest-point descriptors extracted from images by the SIFT algorithm (Lowe 1999), clustered into a vocabulary of visual words (Sivic & Zisserman 2003)

Data set

- Approx. 1000 images (some duplicated) and associated captions, harvested from German news websites
- Over 50 different person names detected in the captions by the NER algorithm:
  - 81% precision and 87% recall vs. ground-truth
- Approx. 175000 interest point descriptors extracted from the images with the SIFT algorithm

Current experiments

- Used a standard classification procedure:
  - Partitioned the data set into 6 stratified subsets – 5 cross-validation sets, and a test-only set
- Trained with respect to the F1-measure (the harmonic average of precision and recall)
- Using the simplex search algorithm of Lagarias et al. (1998) for objective function maximization

Transfer functions

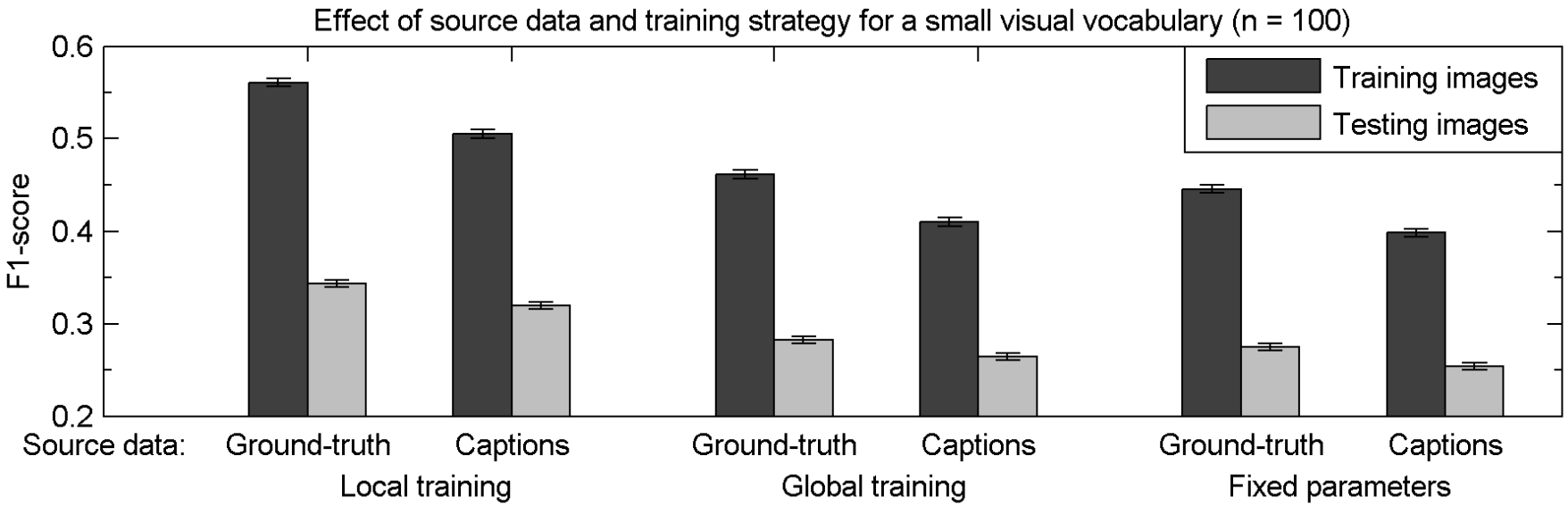
- Defined several methods for calculating association probabilities between keywords and visual prototypes:
  - Use the significance of the chi-square test contrasting the within-cluster (-prototype) and global averages
  - Apply a sigmoid function to the ratio of these averages
  - Apply a sigmoid to the logarithm of the ratio
  - Simply truncate the ratio to an interval centered at or near 1, and then map to the unit interval

Experiment 1 - classifying procedures

- Used visual vocabularies of 100 words (clusters), obtained with the k-means algorithm
- Tested the four methods for calculating the degrees of association between visual prototypes and keywords
- Tested three training strategies – for each keyword separately, globally, and with predefined parameters
- Trained using ground-truth or caption-based associations

Experiment 1 – results

- Minor differences between the four averaging methods
- Best results obtained when using ground-truth data, and training each concept separately:
  - F1-score of 56% at training and 34% at testing

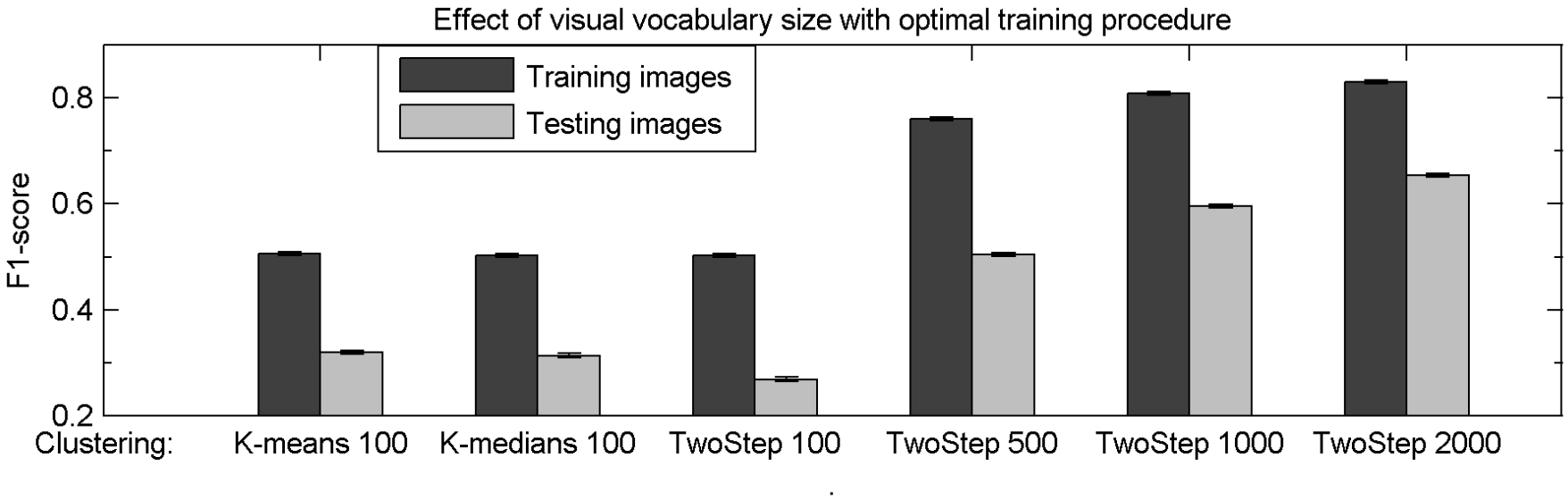


Experiment 2 – vocabulary size

- Different clustering algorithms and numbers of clusters:
  - K-means with 100 clusters (6 hrs)
  - K-medians with 100 clusters (10 hrs)
  - TwoStep (SPSS algorithm for large data sets) with 100, 500, 1000, and 2000 clusters (10 min – 2 hrs)
- Using caption-based data only (realistic setting), and training each concept separately (best performance)

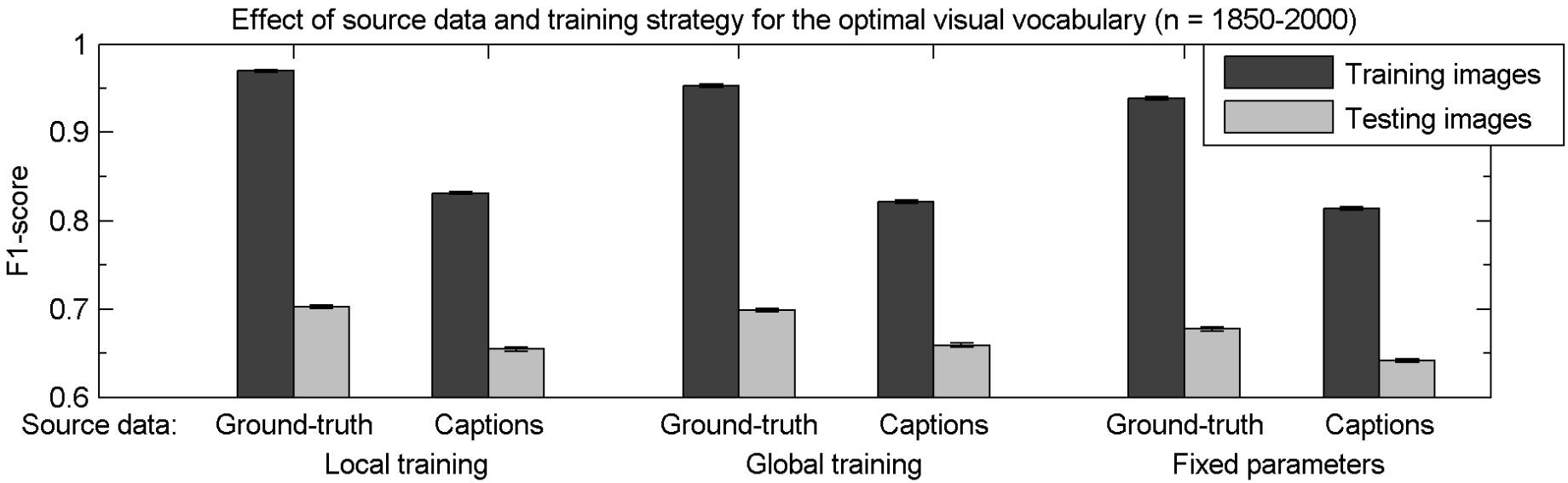
Experiment 2 – results

- Performance increased with the number of clusters, with close to perfect training at approximately 2000 clusters
  - (Data did not have enough variance to produce more clusters with the default settings for TwoStep)



Experiment 3

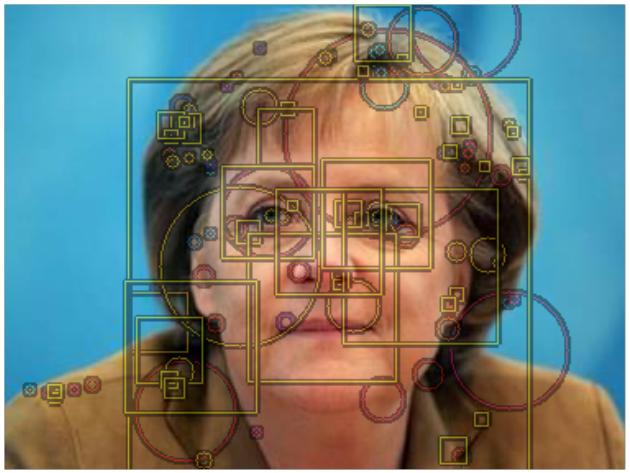
- Repeated the first experiment (testing different classifiers) at the optimal vocabulary size:
  - Significantly improved results, with F1-scores on the test images of 65% – 71% and close to perfect training

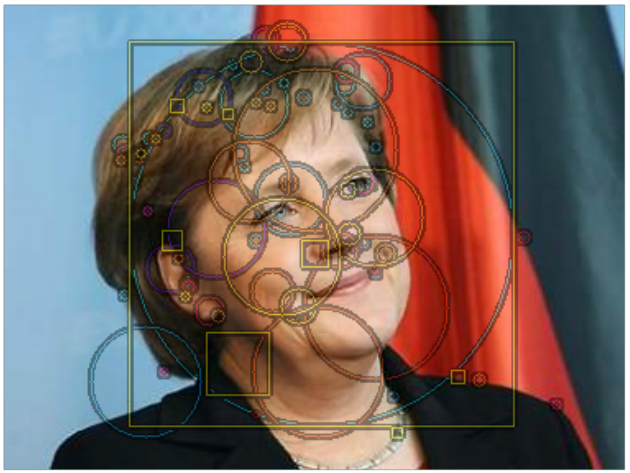


Experiment 3 – further results

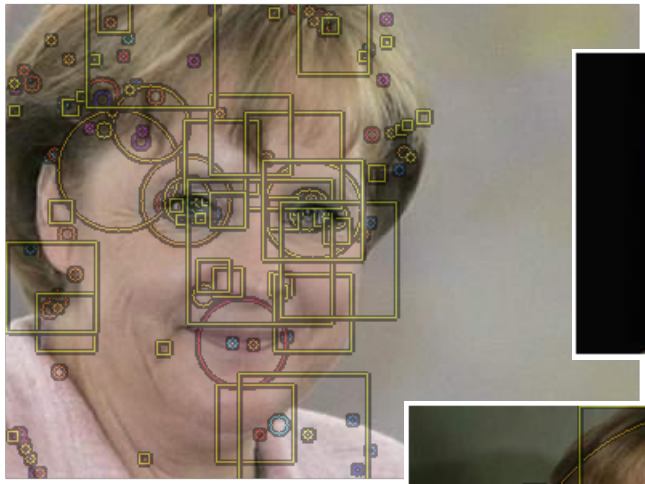
- Best performance using ground-truth data, training each concept separately – F1-score of 71% on test images
- No difference between training each concept separately and training globally when using the captions as source data or measuring the performance on test images
- The impact of training data (ground-truth vs. captions-based) is significantly reduced on testing images

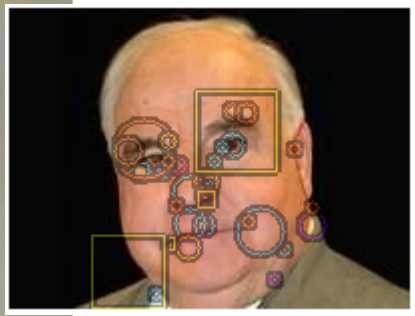
Some examples

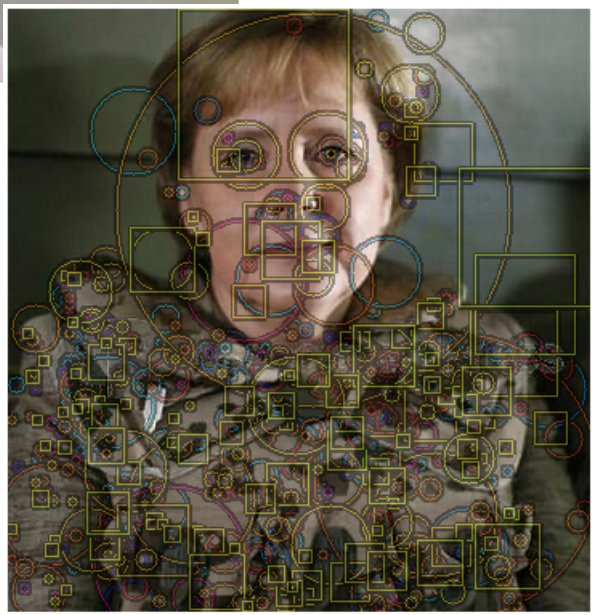




Training







Testing

Summary

- Presented a novel image classification approach, relying on keywords extracted from captions, and on a visual vocabulary of features extracted from the actual images
- Described a method of propagating associations with keywords from training images to visual prototypes, and subsequently to test images
- Applied successfully the developed methodology in a person classification task

Future work

- Generic keyword detector (in progress):
  - Clusters of stemmed words, obtained via latent semantic analysis (LSA, Deerwester et al. 1990) performed on the TF-IDF weighted term-image matrix
- Using an enlarged data set of approx. 150000 images

- Include color data and spatial information about the (relative) position of (groups of) visual features
- Further generalization with respect to concepts (e.g. news category) and data (unstructured websites)

References

Drozdzynski W., Krieger H.-U., Piskorski J., Schäfer U., and Xu F. (2004). Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications. Künstliche Intelligenz, vol. 1, pp. 17-23.

Lagarias J. C., Reeds J. A., Wright M. H., and Wright P. E. (1998). Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. SIAM Journal of Optimization, vol. 9(1), pp. 112-147.

Lowe D. G. (1999). Object Recognition from Local Scale-Invariant Features. Proceedings of the International Conference on Computer Vision, vol. 2, p. 1150

Salton G., and Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, vol. 24, pp. 513-523.

Sivic J., and Zisserman A. (2003). Video Google: A text retrieval approach to object matching in videos. Proceedings of the 9th IEEE International Conference on Computer Vision, Nice, France, pp. 1470-1477.

Preliminary results in:

Jacobs A., Herzog O., Wilhelm Adalbert F.X., and Ilies I. (2008). Relaxation-based data mining on images and text from news web sites. 4th World Conference of the IASC, Yokohama, Japan, pp. 736–743.

Ilies I., Jacobs A., Wilhelm, A.F.X., and Herzog, O. (2009). Classification of News Images Using Captions and a Visual Vocabulary. Technical Report No. 50, TZI, Universität Bremen.

Thank you!

Any questions?