Technologie-Zentrum Informatik

Technical Report 50

Classification of News Images Using Captions and a Visual Vocabulary

Ilies, Iulian
Jacobs, Arne
Wilhelm, Adalbert F.
Herzog, Otthein

TZI-Bericht Nr. 50
2009

Universität Bremen

TZI-Berichte

Herausgeber:
Technologie-Zentrum Informatik
Universität Bremen
Am Fallturm 1
28359 Bremen
Telefon: +49-421-218-7272
Fax: +49-421-218-7820
E-Mail: info@tzi.de
http://www.tzi.de

ISSN 1613-3773

Classification of News Images Using Captions and a Visual Vocabulary

Abstract

A recent technique for automatic image classification that does not rely on associated text data is to employ a vocabulary of visual features. We investigate here a combined approach, using both textual and image-feature information in the construction of the vocabulary. Our method allows the classification of novel images into text-derived categories using only image data. As an initial application, we report the promising results of a series of tests conducted in the setting of person recognition.

1 Introduction

Efficient classification and indexing of textual data using bag-of-words classifiers, inverted files, and related methods are well-researched and widely used (Baeza-Yates & Ribeiro-Neto, 1999). The application of similar procedures on image data is, on the other hand, rather counterintuitive, primarily since images lack words and sentences or similar structures. However, recent research has attempted to achieve such an extension, through the introduction of a so-called "visual vocabulary" (Sivic & Zisserman, 2003). The resulting approaches range from efficient visual search in videos (Sivic & Zisserman, 2003) to object detection (Leibe & Schiele, 2004). Within this line of research, we present here a method for person classification of images from news websites using a vocabulary based on SIFT features (Lowe, 1999).

2 Methods

2.1 Data

The full data set consisted of 110167 interest-point descriptors extracted from 757 unique images with text captions originating from the www.faz.de and www.tagesschau.de news websites. The images were harvested and preprocessed as explained in Jacobs, Hermes, and Wilhelm, 2007. In each image, interest points were identified with the SIFT algorithm (Lowe, 1999). Each interest-point descriptor consisted of 128 integer values (range 0-255), representing a four-by-four grid of edge direction histograms with eight directional bins each, calculated using grey-value pixel data from the neighborhood of the corresponding interest point.

A named entity detector was run on the image captions (see Drozdzynski et al., 2004). Based on the results, every image was associated with one to three person names (with 49 images having more than one person mentioned in the caption). In total, 13 distinct person names were present in the considered set of images, including personalities from politics (e.g. Angela Merkel, Nicolas Sarkozy), sports (e.g. Patrik Sinkewitz), and industry (e.g. Margret Suckale). These connections were extended from images to descriptors in a natural way. On average, there were 8474 descriptors associated with each person (range 4389-19066).

2.2 Preliminary tests

In order to assess the possibility of differentiating between persons using descriptors only, we first examined the frequencies with which the different person names occur in different clusters of descriptors. We constructed a "dictionary" of 100 clusters using the k-means algorithm, and then compared the observed person frequencies with the expected ones using the chi-square test. We repeated this procedure twice: the first time using the data from two persons only, and the second time using the entire data set.

2.3 Prediction tests

For the main experiment, we first split the set of images randomly into a training subset, of size 385, and a test subset, of size 372. Each subset contained approximately half of the images for each person, and approximately half of the total number of descriptors (56724 and 53443, respectively). We then partitioned the set of training descriptors into 100 clusters as before, using the k-means procedure. Subsequently, we predicted the person associations for the test images by extended the person associations as following (see also Figure 1):

- Every training descriptor inherited the person association from its parent image;
- For every cluster, we averaged over the included descriptors to find the observed person frequencies, and divided by the expected frequencies (i.e. those in the entire training set) to obtain relative values;
- Every test descriptor was assigned to the cluster with the nearest centroid, and inherited the relative frequencies from this cluster;
- For every test image, we calculated the average relative person frequencies over its descriptors, and picked as prediction the person with the highest value.

We then measured the accuracy of our predictions by comparing the provided associations with those inferred from the captions of the test images. In order to improve the results, we tried different averaging rules at the last step, including restricting the set of descriptors to those lying in clusters with high person discrimination ability (as measured with the chi-square test), and imposing lower and upper bounds on the relative frequencies. The latter modification proved to be especially effective, most likely since the 13 categories are of very different sizes (with the largest one having about 5 times more descriptors than the smallest).

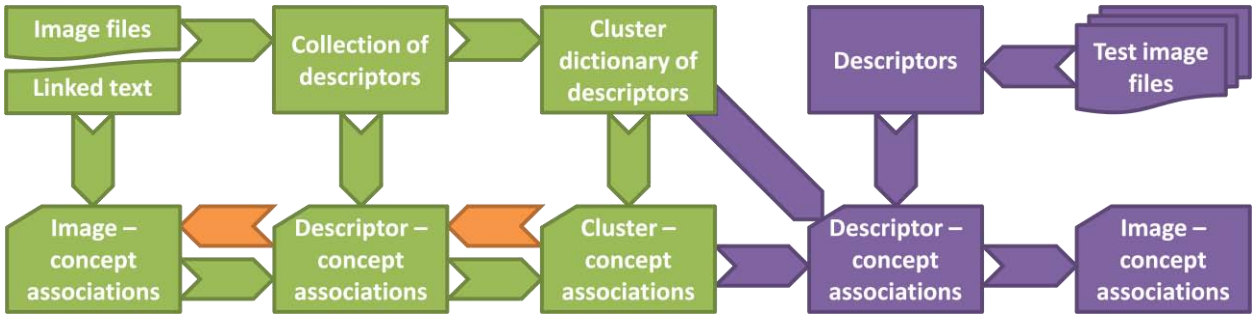

Figure 1: Schematic representation of the developed method (Section 2.3). Green – steps involving training data; purple – steps involving test data; orange arrows – additional steps in the training data optimization process (Section 2.4).

2.4 Optimization tests

In a similar setting, the accuracy of caption-based classifiers was found to be of about 65%, based on manually-annotated ground-truth data (see Jacobs et al., 2008).

Noting that the caption-based person associations are not entirely accurate, we conducted a further batch of tests, where we first tried to optimize the person associations in the training set. We used the caption-based associations as seeds, and used the descriptor clusters as a feedback mechanism in an iterative optimization process. The tests used the same cluster vocabulary and a similar methodology as above (see Figure 1):

- Every training descriptor was assigned with the person association of its parent image;
- For every cluster, we calculated the ratio between the observed and expected person frequencies;
- Every training descriptor was reassigned with the relative frequencies of its cluster;
- For every training image, we recalculated the person associations by averaging over its descriptors, and select the person with the highest value;
- If at the last step at least one image changed the person association, then returned to the first step, otherwise continues;
- Every test descriptor was assigned with the relative frequencies of the cluster with the nearest centroid;
- For every test image, we calculated the average relative person frequencies over its descriptors, and picked as prediction the person with the highest value.

All averages in steps 4 and 7 were performed using the best rule inferred in the previous experiment (bounding the relative frequencies to the 50-130% interval). We then compared both the stationary state of the training system and the predictions on the test images with the original caption-based associations.

3 Results

3.1 Preliminary tests

For the first set of tests, we considered the simplified task of discriminating between two categories only. To maximize the chances to observe significant effects, we chose two persons that would differ in both gender and clothing – the politician Angela Merkel, and the cyclist Patrik Sinkewitz. We applied the k-means algorithm with 100 clusters on the reduced data set consisting of the descriptors associated with either of these two persons. The convergence was fairly slow (97 iterations), and the clusters were of similar sizes (Figure 2, left panel) and evenly distributed across the data space, indicating that the obtained system of clusters is a vocabulary in the sense of Sivic and Zisserman (2003), rather than a representation of some hidden structure in the data. Within each cluster, we counted the number of descriptors associated with either of the two persons, and contrasted the observed values with the overall frequencies via chi-square tests. The frequency distributions were significantly different (at the 1% level) than the baseline in 16 (6) cases out of 100 (uncorrected / corrected for multiple comparisons, respectively) (Figure 2, middle and right panels), showing that it could be possible to use descriptor clusters for distinguishing between categories.

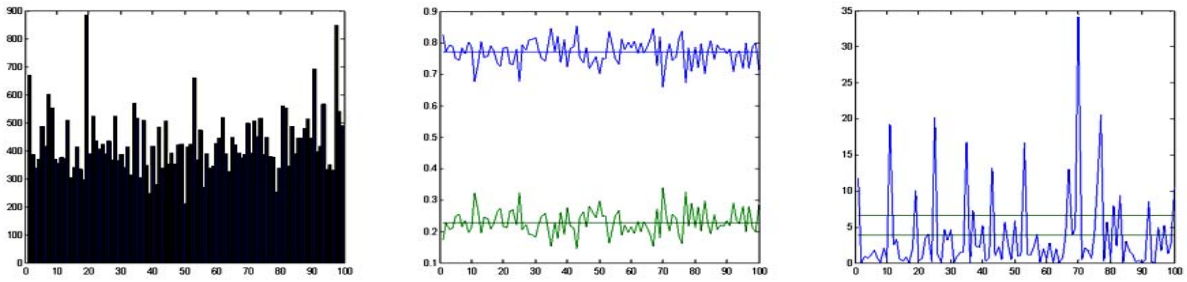

Figure 2: Preliminary tests on the reduced data set. Left – number of training descriptors in each cluster. Middle – Observed and expected (horizontal lines) frequencies of the two persons; blue – Merkel, green - Sinkewitz. Right – chi-square scores; horizontal lines are drawn at the corrected and uncorrected 0.01 significance thresholds.

In order to test whether the above observations hold in a more general context, we repeated the tests using the complete set of descriptors. We run the k-means algorithm with 100 clusters on the entire data set; the convergence was even slower (157 iterations), and the clusters were again of similar sizes (Figure 3, left panel) and similarly distributed across the data space. On average, each cluster contained descriptors belonging to half of the images (Figure 3, middle), and each of the 13 persons was represented in every cluster. The observed person frequency distributions were significantly different (at the 1% level) than the global ones for 96 (86) of the 100 clusters (uncorrected / corrected for multiple comparisons, respectively) (Figure 3, right panel). Basically, almost all clusters exhibited a bias towards one or several persons.

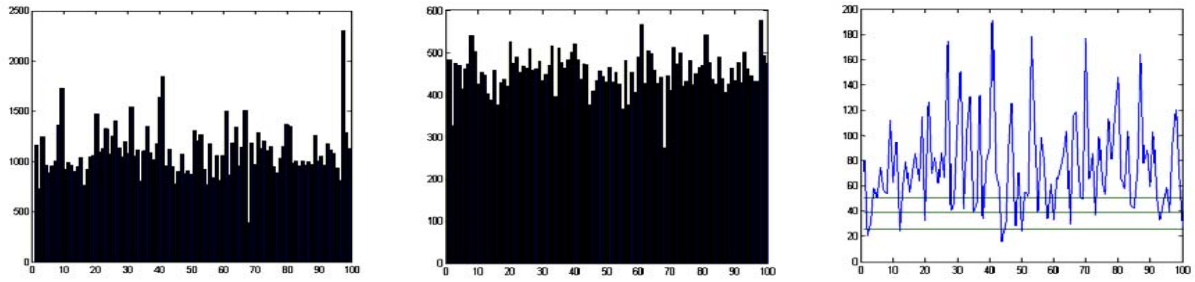

Figure 3: Preliminary tests on the full data set. Left – number of training descriptors in each cluster. Middle – number of training images represented in each cluster. Right – chi-square scores; horizontal lines mark the thresholds for the 0.0001, 0.01, and uncorrected 0.01 significance thresholds (from top to bottom).

3.2 Prediction tests

To start, we constructed a 100-cluster "dictionary" using the training data, using the k-means algorithm. The convergence was really slow (177 iterations, total running time of over two hours), and, correspondingly, the clusters were of similar sizes (Figure 4, left) and space covering. The observed person frequency distributions were significantly different (at the 1% level) than the expected values for 88 (71) of the 100 clusters (uncorrected / corrected for multiple comparisons, respectively) (Figure 4, middle). We then assigned each test descriptor to the nearest cluster; again, the distribution was rather uniform (Figure 4, right panel).

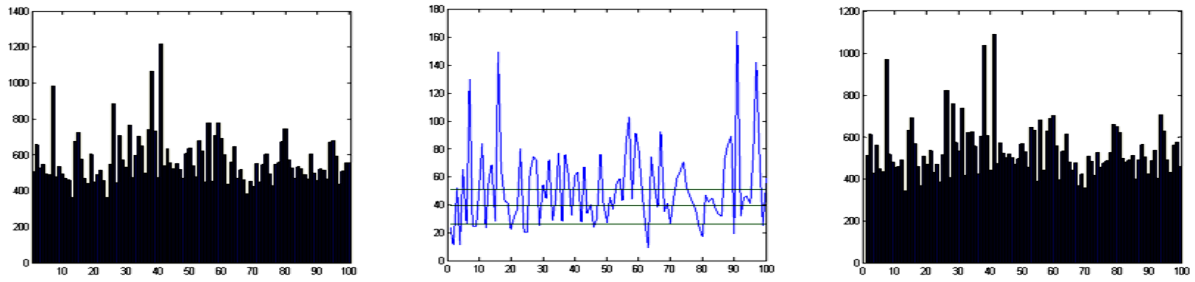

Figure 4: Characteristics of the cluster dictionary constructed for the prediction tests. Left – number of training descriptors in each cluster. Middle – chi-square scores; horizontal lines are drawn at the thresholds for the 0.0001, 0.01, and uncorrected 0.01 significance levels. Right – number of test descriptors assigned to each cluster.

Subsequently, we tested several methods for predicting the person associations of the test images using the relative frequencies of the clusters to which their descriptors were assigned (see Section 2.3). In the basic implementation, we simply averaged the relative frequencies over the descriptors belonging to each test image, and chose as prediction the person with the highest value. Compared to the available

caption-based person associations, our predictions agreed in 30% of the cases. Notably, the procedure did not predict Angela Merkel as the person in the image in any of the cases, even though this is the largest category in the present data, while two of the persons with the lowest incidences accounted for 44% of the predictions. This effect is presumably due to the increased likelihood of attaining high relative frequency variations for small categories; to counteract it, we tried to limit the amplitude of these differences. The best performance was obtained for a lower limit of 0.5 and an upper limit of 1.3 for the relative frequencies. In this case, 47% of the predictions agreed with the caption-based associations (range 24-85% over the 13 persons). Other modifications, such as considering only the descriptors belonging to the clusters with the top 50% chi-square scores, did not improve the results.

3.3 Optimization tests

For the final experiment, we first brought the images-clusters training system to a stationary state, as described in Section 2.4, and only then predicted the person associations for the test images. Convergence was really fast, with no more changes occurring after only 16 iterations. The agreement between the initial (caption-based) and final person associations of the training images was of 35%. Correspondingly, the predictions for test images agreed with their captions in 28% of the cases. If allowing the images to be associated with up to three persons, the cumulative degree of agreement with the captions increased to 50% in the training set, and 42% in the test set. Unfortunately, we could not assess whether these results represented an improvement over using the basic procedure (Section 3.2), since the true image-person associations were not known.

4 Discussion

4.1 Conclusions

In this paper, we have presented a new application of the image vocabulary method (Sivic & Zisserman, 2003) to a generalized image classification task. The images employed originated from an online news site, and therefore were not standardized with respect to factors such as size, background, location and position of the target person. In this setting, we developed a framework for predicting the person featured in an image using only interest point descriptors and caption information from other similar images. Given the heterogeneity of the data and the large number of classes, the results were excellent, with almost half of the predictions being correct when compared to the information extracted from the captions of the test images. Rather than stemming from a few salient features or, worse, from accidentally recurring background features, these predictions appeared to be a global, averaging effect (see Figure 5).

4.2 Future work

The method outlined in this paper is still in development. A possible improvement would be to find a faster alternative for the k-means algorithm used in constructing the cluster dictionary. More importantly, further tuning of the averaging parameters (Section 2.3) is needed to ensure the obtained

results are optimal. To achieve that, and to better quantify the performance of our method, the next step would be a ground-truth verification of the person associations for the images in the current data set. After the tweaking is finalized, several extensions of the method would be possible. On one hand, one could test other classification tasks on the same data and using the same dictionary, e.g. finding the gender. On the other hand, one could use the entire set for training, and then try to classify other images from the same source that do not have captions.

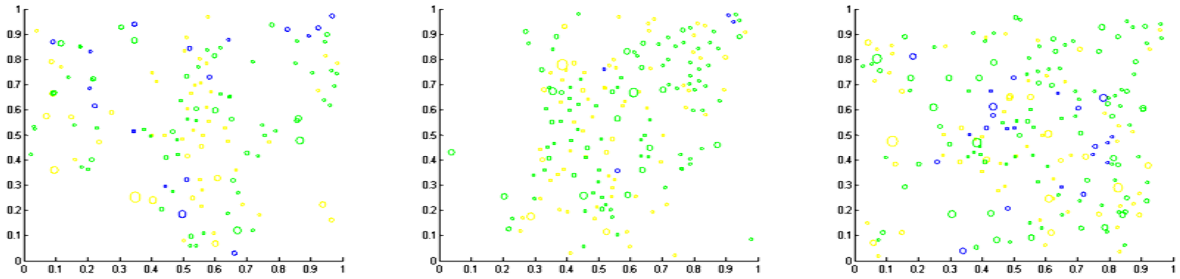

Figure 5: Size and position of descriptors in several test images. Blue – the descriptor had the highest relative frequency for the predicted person; green – the relative frequency for the predicted person was supra-unitary, but not the highest; yellow – the relative frequency for the predicted person was sub-unitary.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval.* New York, NY: ACM Press.

Drozdzynski, W., Krieger, H. U., Piskorski, J., Schäfer, U., & Xu, F. (2004). Shallow processing with unification and typed feature structures - foundations and applications. *Künstliche Intelligentz , 18* (1), 17-23.

Jacobs, A., Hermes, T., & Wilhelm, A. (2007). Automatic image annotation by association rules. *Electronic Imaging and the Visual Arts 2007*, (S. 108-112). Berlin, Germany.

Jacobs, A., Herzog, O., Wilhelm, A., & Ilies, I. (2008). Relaxation-based data mining on images and text from news web sites. *Proceedings of IASC2008*, (S. 736-743). Yokohama, Japan.

Leibe, B., & Schiele, B. (2004). Scale-invariant object categorization using a scale-adaptive mean-shift search. *Proceedings of the 26th DAGM Symposium on Pattern Recognition*, (S. 145-153). Tübingen, Germany.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the 7th IEEE International Conference on Computer Vision*, (S. 1150-1157). Kerkyra, Greece.

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