AUTOMATIC ANNOTATION OF HISTOPATHOLOGICAL IMAGES USING A LATENT TOPIC MODEL BASED ON NON-NEGATIVE MATRIX FACTORIZATION

Cruz-Roa A., Diaz G., Romero E., González F. - Universidad Nacional de Colombia
{aacruzr,gmdiazc,edromero,fagonzalezo}@unal.edu.co

UNIVERSIDAD NACIONAL DE COLOMBIA SEDE BOGOTÁ

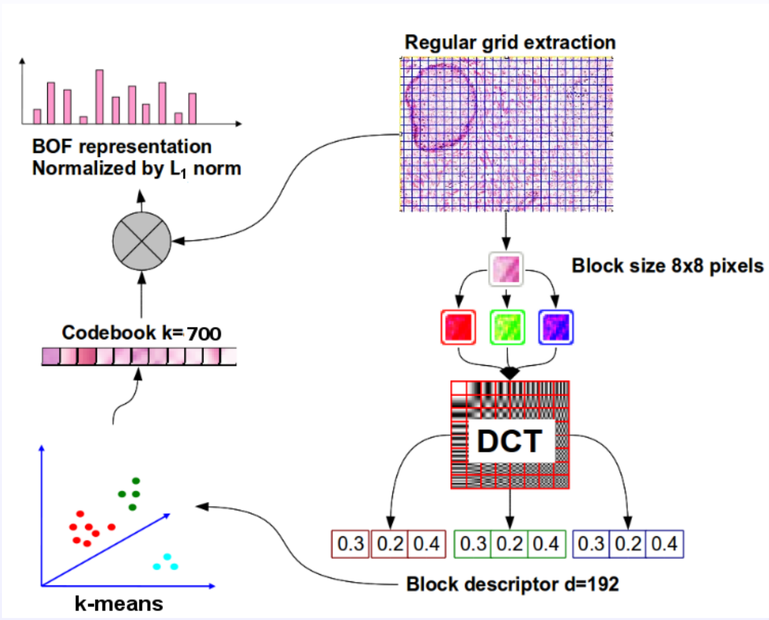
BioIngenium Research Group

Abstract

Histopathological images are an important resource for clinical diagnosis and biomedical research. Automatic annotation of these images is particularly challenging from an image understanding point of view. This paper presents a novel method for automatic histopathological image annotation based on three complementary strategies, first, a part-based image representation, called the bag of features, which takes advantage of the natural redundancy of histopathological images, second a latent topic model, based on non-negative matrix factorization, which is in charge of capturing the high-level visual patterns, and, third, a probabilistic annotation model that connects visual patterns with the semantics of this problem. The method was evaluated using 1604 annotated images of basal cell carcinoma, a collection with different types of skin cancer. The preliminary results demonstrate an improvement on precision and recall of 24% and 64% against support vector machines.

Bag of Features (BOF)

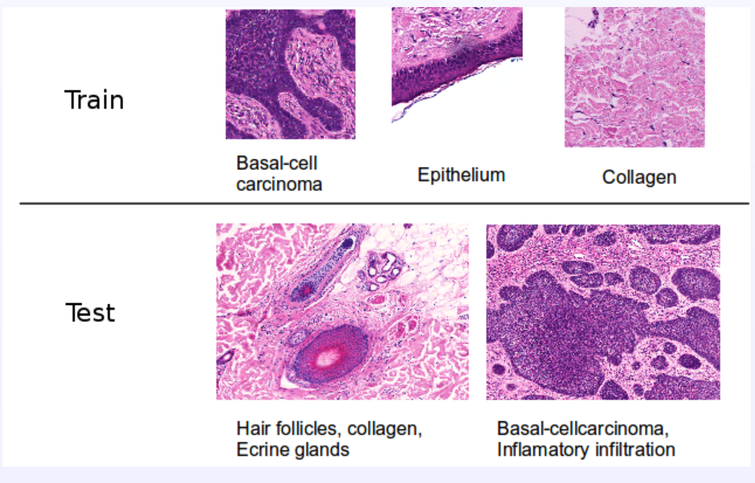
The visual representation of histopathological images is obtained as a bag of features (BOF). The below image depicts the setup used here.



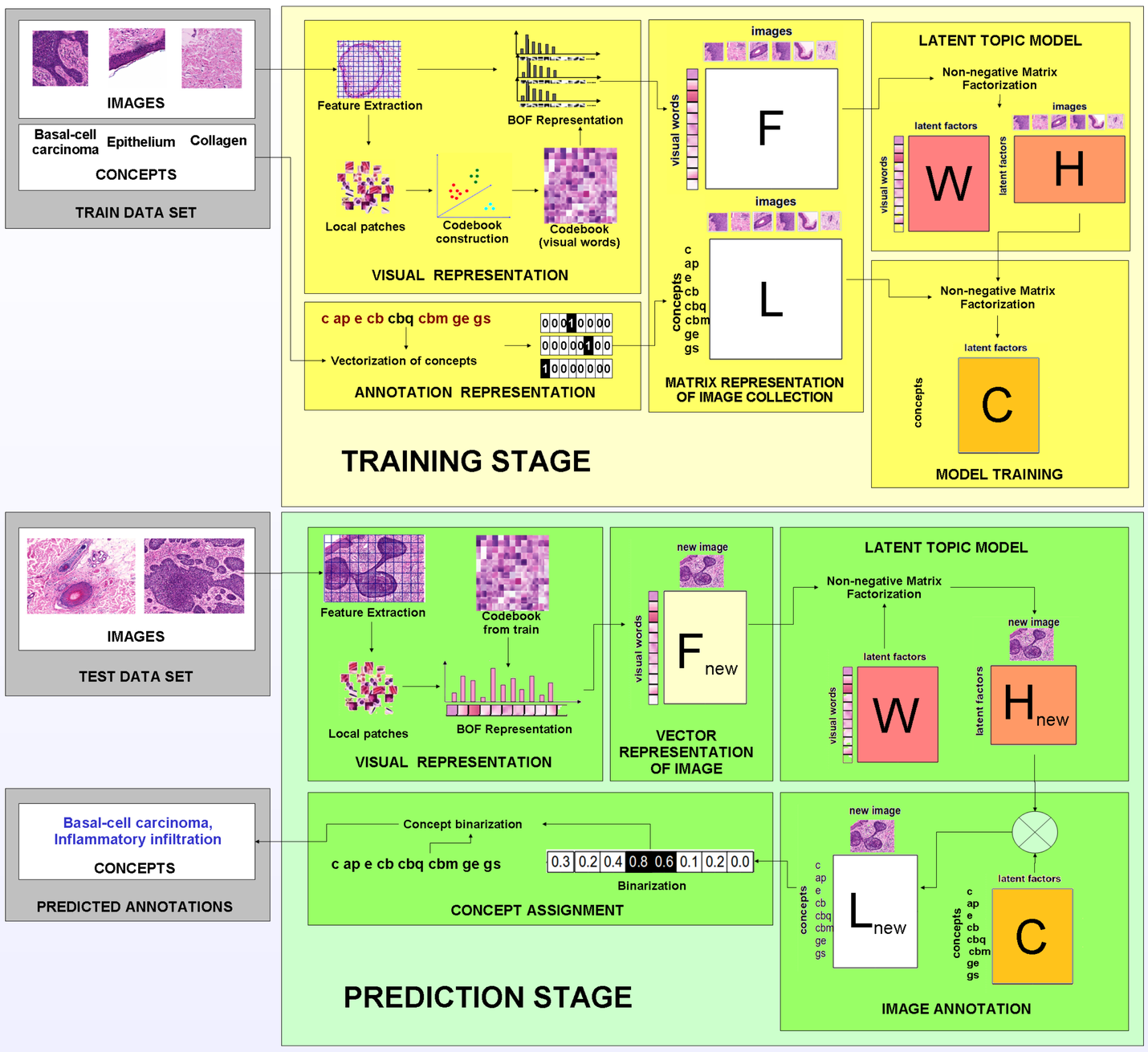
Histopathology dataset

The image dataset used here corresponds to a skin cancer known as basal cell carcinoma stained with hematoxylin-eosin (HE). This dataset is composed by two set of images, 1466 (training) and 138 (testing). The training image set (mono-label) comprises subimages of 300 × 300 pixels, each annotated with only one of the 10 concepts present in the collection, whereas the test image set (multi-label) comprises larger images of 1024 × 768 pixels, which, in general, are annotated with more than one concept.

Example images of each data set are shown below:



Annotation Model Based on NMF and BOF (A2NMF)



Results and Conclusion

The below Table shows the average performance in the test dataset using Accuracy (Acc), Precision (Pr), Recall (Rc) and F-measure (F).

	1-concept images				n-concepts images			
Method	**Acc**	**Pr**	**Rc**	**F**	**Acc**	**Pr**	**Rc**	**F**
SVM-RBF	**0.96**	**0.84**	**0.69**	**0.76**	0.70	0.26	0.10	0.11
A2NMF	0.92	0.67	0.46	0.51	**0.76**	**0.5**	**0.74**	**0.55**

The evaluation was performed in two scenarios: a simple mono-label annotation task, corresponding to using only training images (80%-20% split for training and testing), and the original complex multi-label annotation task. In both cases the method was compared with a Support Vector Machine (SVM) with RBF kernel choosing the best parameters by 10-fold cross-validation.

Histopathological images are particularly challenging to analyze because of their high variability and complex visual structure. These results suggest that latent semantic characterization of the visual structure is a viable alternative to build competitive histopathology annotation models.