Visual Mining in Histology Images Using Bag of Features

Angel Cruz-Roa, Juan C. Caicedo, Fabio González

SIPAIM 2010

Bioingenium Research Group, 2010
Outline

Introduction
   Histology Image Dataset
   Motivation
   Problem

Proposed Method
   Collection-based Image Representation
   Visual Mining using Feature Selection and Co-clustering Analysis
   Automatic Annotation in Histology Images

Conclusion
Outline

Introduction

Histology Image Dataset
Motivation
Problem

Proposed Method
Collection-based Image Representation
Visual Mining using Feature Selection and Co-clustering Analysis
Automatic Annotation in Histology Images

Conclusion
Image dataset

Histology dataset

- Normal tissues
- Four fundamental tissues (epithelial, connective, muscular and nervous)
- Different stains (HE, PAS, trichrome of Masson, etc.)
- 2,828 images
Histology Dataset

Figure: Sample images of four fundamental tissues from histology image dataset.
Outline

Introduction
  Histology Image Dataset
  Motivation
  Problem

Proposed Method
  Collection-based Image Representation
  Visual Mining using Feature Selection and Co-clustering Analysis
  Automatic Annotation in Histology Images

Conclusion
Motivation

Image analysis $\Rightarrow$ image collection analysis (as a whole).
Data and text mining

Bioimage informatics: a new area of engineering biology

Hanchuan Peng
Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia, USA

Received on June 6, 2008; revised on July 1, 2008; accepted on July 2, 2008
Advance Access publication July 4, 2008
Associate Editor: Jonathan Wren

ABSTRACT

In recent years, the deluge of complicated molecular and cellular microscopic images creates compelling challenges for the image computing community. There has been an increasing focus on developing novel image processing, data mining, database and visualization techniques to extract, compare, search and manage the biological knowledge in these data-intensive problems. This emerging new area of bioinformatics can be called 'bioimage informatics'. This article reviews the advances of this field from several aspects, including applications, key techniques, available tools and resources. Application examples such as high-colors which may correspond to various molecular reporters, (3) 4D spatio-temporal information for developing tissues or moving cells, (4) various co-localized biological signals such as mRNA expression levels of different genes (Lein et al., 2007; Long et al., 2007b; Peng et al., 2007) or (5) other screening experiments related to RNA interference (RNAi), chemical compounds, etc. (Echeverri and Perrimon, 2006; Moffat et al., 2006; Sepp et al., 2008). Analyzing these images is critical for biologists to seek answers to many biological problems, such as differentiating cancer cell phenotypes (Long et al., 2007a), categorization of neurons (Jefferis et al., 2007), etc.
Outline

Introduction
Histology Image Dataset
Motivation
Problem

Proposed Method
Collection-based Image Representation
Visual Mining using Feature Selection and Co-clustering Analysis
Automatic Annotation in Histology Images

Conclusion
Problem definition

How to **extract knowledge** in an **automatic way** from medical image databases?

The visual content in medical images is difficult to characterize and to associate with their semantics, because the medical images are heterogenous (acquisition techniques, anatomical variability, points of view, etc.)

To extract knowledge in medical images is particularly challenging!
Problem definition

How to **extract knowledge in an automatic way** from medical image databases?

The visual content in medical images is difficult to characterize and to associate with their semantics, because the medical images are heterogenous (acquisition techniques, anatomical variability, points of view, etc.)

**To extract knowledge in medical images is particularly challenging!**
How to extract knowledge?

- How to characterize relationships between images?
- How to find common and distinctive characteristics among them?
- How to find implicit categories or groups that could be identified in the collection?

How to relate visual content with semantic content?
How to extract knowledge?

- How to characterize relationships between images?
- How to find common and distinctive characteristics among them?
- How to find implicit categories or groups that could be identified in the collection?

**How to relate visual content with semantic content?**
Proposed Method

Image Collection → Image Blocks → Feature Extraction → BOF Representation → BOF Image Collection

Collection-based Image Representation

Codebook Construction (k-means) → Codebook

Visual Patterns → Feature Selection and Analysis → Coclustering Analysis

Discriminant codewords

Classifier Training

New Image → Annotation Model → Annotated Image

Image Collection Automatic Annotation
Outline

Introduction
- Histology Image Dataset
- Motivation
- Problem

Proposed Method
- Collection-based Image Representation
- Visual Mining using Feature Selection and Co-clustering Analysis
- Automatic Annotation in Histology Images

Conclusion
Question

How to represent the visual content in an image collection?
Collection-based Image Representation

Figure: Overview of the Bag of Features.
Visual words (or image patches)

In BOF, image patches are the visual equivalents of individual “words” and the image is treated as an unstructured set (“bag”) of these [Nowak 2006]. Visual words are 8x8 sized blocks, described using:

- Raw-blocks (texture)
- SIFT (texture)
- DCT (texture & color)
Codebook examples

Figure: Comparison of visual words in the dictionaries of size 500 based on blocks (left) and DCT (right) sorted by their occurrence.
Question

How is the distribution of visual words in an image collection?
Zipf’s Law in Language Codebooks

Figure: Comparison of Zipf curves for English, Spanish, Irish and Latin. [Ha2006]
Zipf’s law in Visual Codebook

Figure: The frequency of visual words against their rank for 1000-size codebook based on blocks, SIFT and DCT in histology dataset.
Outline

Introduction
Histology Image Dataset
Motivation
Problem

Proposed Method
Collection-based Image Representation
Visual Mining using Feature Selection and Co-clustering Analysis
Automatic Annotation in Histology Images

Conclusion
Question

How to select the most discriminant visual words from a visual codebook?
Feature Selection

What is feature selection?

- Is a method to choose a subset of features with high information content.
- There are several methods (BLogReg, CFS, Chi-square, FCBF, Fisher score, Gini Index, Information Gain, Kruskal-Wallis, ReliefF, ... and so on).
- A State-of-the-Art method is Minimum Redundance Maximum Relevance Feature Selection (mRMR) [Peng2005¹].

# mRMR Feature Selection

## Max-Relevance criteria

\[
\max_W D(W, c_j) = \max_W \frac{1}{|W|} \sum_{w_i \in W} I(w_i; c_j), \quad (1)
\]

## Min-Redundance criteria

\[
\min_W R(W) = \min_W \frac{1}{|W|^2} \sum_{w_i, w_j \in W} I(w_i; w_j) \quad (2)
\]

## mRMR optimization criteria

\[
\max_W \Phi(W, c_j) = \max_W D(W, c_j) - R(W) \quad (3)
\]
Visual words selected by mRMR

Figure: 100 visual words selected by mRMR method in histology dataset.
Question

What are the most relevant visual words per concept?
Codewords with highest conditional probabilities

| Concept    | #Words | max $P(C_j|w_i)$ | Visual Words |
|------------|--------|------------------|--------------|
| Muscular   | 18     | 1                | ![Visual Words](Muscular) |
| Epithelial | 21     | 0.569792         | ![Visual Words](Epithelial) |
| Nervous    | 58     | 1                | ![Visual Words](Nervous)  |
| Connective | 3      | 0.5              | ![Visual Words](Connective) |

| Concept    | #Words | max $P(C_j|w_i)$ | Visual Words |
|------------|--------|------------------|--------------|
| Muscular   | 24     | 0.821853         | ![Visual Words](Muscular) |
| Epithelial | 31     | 0.971094         | ![Visual Words](Epithelial) |
| Nervous    | 26     | 0.938613         | ![Visual Words](Nervous)  |
| Connective | 19     | 0.863061         | ![Visual Words](Connective) |
Question

Can we locate the blocks in an image that belong to the most relevant visual words?
Location of Relevant Visual Words in an Image

Figure: Original images annotated with *muscular* tissue.
Location of Relevant Visual Words in an Image

Figure: Spatial location of visual codewords according with high conditional probabilities from DCT-based codebook.
The previous analysis relates individual visual words and concepts.

**Question**

How to relate groups of visual words and images with concepts?
The previous analysis relates *individual* visual words and concepts.

**Question**

How to relate *groups* of visual words and images with concepts?
Coclustering in Gene expression analysis

Figure: Graphical representation (Heat map) for genes expression analysis. Rows are the patients (healthy or not) and columns are genes.
Coclustering in histology images

- G1: epithelial tissue
- G2: nervous tissue
- G3: epithelial tissue

---

Proposed Method

- Clustering algorithm
- Predictive model

---

Conclusion
Outline

Introduction
Histology Image Dataset
Motivation
Problem

Proposed Method
Collection-based Image Representation
Visual Mining using Feature Selection and Coclustering Analysis
Automatic Annotation in Histology Images

Conclusion
Question

How affects the codebook size and visual word type the automatic annotation performance?
Automatic Annotation Performance

Table: Automatic annotation performance for both datasets.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Precision</th>
<th>Recall</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fundamental tissues dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(k = 150)</td>
<td>0.60</td>
<td>0.61</td>
<td>0.68</td>
<td>0.65</td>
<td>0.74</td>
<td>0.66</td>
</tr>
<tr>
<td>(k = 500)</td>
<td>0.52</td>
<td>0.27</td>
<td>0.52</td>
<td>0.31</td>
<td>0.49</td>
<td>0.36</td>
</tr>
<tr>
<td>(k = 1000)</td>
<td>0.84</td>
<td>0.83</td>
<td>0.89</td>
<td>0.87</td>
<td>0.91</td>
<td>0.88</td>
</tr>
</tbody>
</table>

**Proposed Method**
Conclusion

• Is possible to extract knowledge from medical image databases!, this approach is just an idea for performing visual mining in histology images.

• BOF representation is useful to do image analysis in different ways.

• Blocks-based and DCT-based visual words capture different aspects (appearance/semantic) of histology images.

• Visual mining could be a powerful tool to support biomedical image research!
Thanks for your attention!
Questions?
References

Manfred Auer, Hanchuan Peng, and Ambuj Singh.
Development of multiscale biological image data analysis: Review of 2006 international workshop on multiscale biological imaging, data mining and informatics, santa barbara, USA (BII06).

Kristian Kvilekval, Dmitry Fedorov, Boguslaw Obara, Ambuj Singh, and B. S. Manjunath.
Bisque: a platform for bioimage analysis and management.

H. Peng.
Bioimage informatics: a new area of engineering biology.

Bioimage informatics for experimental biology*.

Jason R. Swedlow and Kevin W. Eliceiri.
Open source bioimage informatics for cell biology.