LEUKO:
Computer Decision Support System for Leukemia Diagnosis
Daniela M. Ushizima
Motivation

• Leukemia = 30% cancer.
• Pattern recognition: low-cost alternative solution;
• Human visual analysis:
  • 79.3% to 92.3% correct (lymphocytes);
  • 150 cells >> 10 minutes for an hematologist;
• Why lymphoproliferative diseases?
• Software for cell analysis and classification:
  • computer-aided automation;
  • educational purposes;
Synopsis

1. Human blood cells
2. Pattern recognition
3. Feature selection
4. Results
5. Latest developments
1

Human blood
1.1. Where leukocytes come from?

Primary Lymphoid Organs

Secondary Lymphoid Organs

Thymus

Tonsils and adenoids

Spleen

Peyer’s patches

Appendix

Lymph nodes

Lymphatic vessels

Bone marrow

Peripheral blood
1.2. Normal leukocytes

- Monocyte
- Lymphocyte
- Eosinophil
- Neutrophil
- Basophil
1.3. Malignant cells

- Prolymphocyte
- Hairy-cell
- CLL
2 Pattern Recognition
2.1. Steps in PR design

- Digital Image
- Preprocessing
- Segmentation
- Feature extraction
- Classification
- Human expertise
- Classes
- Examples
- Feature selection
- Criterium function

\[ \tilde{f} = \{f_0, f_1, \ldots, f_{n-1}\} \]
2.2. Segmentation

- Regions of interest: nucleus, cytoplasm, red blood cells and background
2.2.1. Methods

• G channel thresholding: where is the nucleus?

• Color segmentation: where are the leukocyte and RBCs?
2.2.2. Thresholding

<table>
<thead>
<tr>
<th>Green levels</th>
<th>Absolute frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>180 200 220</td>
<td>100 98 90 95 95</td>
</tr>
<tr>
<td>180 230 221</td>
<td>160 82 95 93 90</td>
</tr>
<tr>
<td>230 200 220</td>
<td>100 99 85 82 96</td>
</tr>
<tr>
<td>10 200 220</td>
<td>100 50 30 35 25 10</td>
</tr>
<tr>
<td>230 221 160</td>
<td>32 35 23 10 20</td>
</tr>
<tr>
<td>200 220 100</td>
<td>50 30 35 25</td>
</tr>
<tr>
<td>10 200 22</td>
<td>100 50 30 35 25</td>
</tr>
<tr>
<td>230 22 160</td>
<td>32 35 23 10 13</td>
</tr>
<tr>
<td>100 120 200</td>
<td>15 0 35 25 0</td>
</tr>
</tbody>
</table>

30-May-07

Computer Vision in Leukemia Diagnosis

12/41
Fixed threshold

Katz (2000)

30-May-07 Computer Vision in Leukemia Diagnosis 14/41
Dynamic threshold

threshold > 150

threshold > 142

threshold > 150

threshold > 152

threshold > 156

threshold > 122

threshold > 140

threshold > 124

threshold > 132

threshold > 136

threshold > 152

threshold > 124

threshold > 150

threshold > 136

threshold > 136

threshold > 122

threshold > 125

threshold > 136

threshold > 134

threshold > 126

threshold > 144

threshold > 152

threshold > 144

threshold > 146


30-May-07

Computer Vision in Leukemia Diagnosis

15/41
2.2.3. Color segmentation
2.3. Feature extraction

- Image properties;
- Nucleus and cytoplasm attributes;
- Shape;
- Size;
- Color;
- Texture.

- Traditional measures
- Subjective descriptions
<table>
<thead>
<tr>
<th>Id</th>
<th>Feature description</th>
<th>Id</th>
<th>Feature description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N perimeter</td>
<td>32</td>
<td>N energy Blue</td>
</tr>
<tr>
<td>2</td>
<td>N area</td>
<td>33</td>
<td>N entropy Blue</td>
</tr>
<tr>
<td>3</td>
<td>N max area</td>
<td>34</td>
<td>N/C ratio</td>
</tr>
<tr>
<td>4</td>
<td>N num elements</td>
<td>35</td>
<td>C inertia1</td>
</tr>
<tr>
<td>5</td>
<td>N factor form</td>
<td>36</td>
<td>C entropy1</td>
</tr>
<tr>
<td>6</td>
<td>N num concavity</td>
<td>37</td>
<td>C energy1</td>
</tr>
<tr>
<td>7</td>
<td>N num convexity</td>
<td>38</td>
<td>C correlation1</td>
</tr>
<tr>
<td>8</td>
<td>N bending energy</td>
<td>39</td>
<td>C homogeneity1</td>
</tr>
<tr>
<td>9</td>
<td>N num holes</td>
<td>40</td>
<td>C inertia2</td>
</tr>
<tr>
<td>10</td>
<td>N area holes</td>
<td>41</td>
<td>C entropy2</td>
</tr>
<tr>
<td>11</td>
<td>N inertia</td>
<td>42</td>
<td>C energy2</td>
</tr>
<tr>
<td>12</td>
<td>N entropy</td>
<td>43</td>
<td>C correlation2</td>
</tr>
<tr>
<td>13</td>
<td>N energy</td>
<td>44</td>
<td>C homogeneity2</td>
</tr>
<tr>
<td>14</td>
<td>N correlation</td>
<td>45</td>
<td>C mean Red</td>
</tr>
<tr>
<td>15</td>
<td>N homogeneity</td>
<td>46</td>
<td>C std Red</td>
</tr>
<tr>
<td>16</td>
<td>N mean Red</td>
<td>47</td>
<td>C skewness Red</td>
</tr>
<tr>
<td>17</td>
<td>N std Red</td>
<td>48</td>
<td>C kurtosis Red</td>
</tr>
<tr>
<td>18</td>
<td>N skewness Red</td>
<td>49</td>
<td>C energy Red</td>
</tr>
<tr>
<td>19</td>
<td>N kurtosis Red</td>
<td>50</td>
<td>C entropy Red</td>
</tr>
<tr>
<td>20</td>
<td>N energy Red</td>
<td>51</td>
<td>C mean Green</td>
</tr>
<tr>
<td>21</td>
<td>N entropy Red</td>
<td>52</td>
<td>C std Green</td>
</tr>
<tr>
<td>22</td>
<td>N mean Green</td>
<td>53</td>
<td>C skewness Green</td>
</tr>
<tr>
<td>23</td>
<td>N std Green</td>
<td>54</td>
<td>C kurtosis Green</td>
</tr>
<tr>
<td>24</td>
<td>N skewness Green</td>
<td>55</td>
<td>C energy Green</td>
</tr>
<tr>
<td>25</td>
<td>N kurtosis Green</td>
<td>56</td>
<td>C entropy Green</td>
</tr>
<tr>
<td>26</td>
<td>N energy Green</td>
<td>57</td>
<td>C mean Blue</td>
</tr>
<tr>
<td>27</td>
<td>N entropy Green</td>
<td>58</td>
<td>C std Blue</td>
</tr>
<tr>
<td>28</td>
<td>N mean Blue</td>
<td>59</td>
<td>C skewness Blue</td>
</tr>
<tr>
<td>29</td>
<td>N std Blue</td>
<td>60</td>
<td>C kurtosis Blue</td>
</tr>
<tr>
<td>30</td>
<td>N skewness Blue</td>
<td>61</td>
<td>C energy Blue</td>
</tr>
<tr>
<td>31</td>
<td>N kurtosis Blue</td>
<td>62</td>
<td>C entropy Blue</td>
</tr>
</tbody>
</table>

N = nucleus, C = cytoplasm

Shape and size

Texture

Color
2.3.1. Shape and size

- Nucleus-cytoplasm ratio ($NC$):
  - $NC = \frac{A_{nuc}}{A_{nuc} + \text{Area}_{cit}}$

- Factor form=compactness measure ($C$):
  - $C = \frac{P^2}{A}$

- Curvature ($k$):
  - Concave and convexity points;

- Bending Energy ($B$):
  - $B = \frac{1}{P} \sum_{t=1}^{P} k^2(t)$
2.3.2. Fractal dimension

- **Costa**: multiscale fractal dimension (2D) curve and its peak ($FD_{\text{max}}$) for complexity description of neurons;
Spatial distribution of the graylevels
2.3.2. $\text{FD}_{\text{max}}$ - the method

- Multiscale fractal dimension
  - Peak of max fractality
  - Scale of max fractality
  - Width of max fractality

- $f_1 = \ln(f_1)$
- $f_2 = \ln(f_1)$
- $f_3 = 3 - d(f_2)$
2.3.2. Gray Level Coocurrence Matrix

- Joint probability distribution of 2 pixels given a displacement (direction+distance) and a window;

$$I(m,n) = \text{gray levels}$$

$$d = (1,0) & w = m \times n$$

$$GLCM(i,j) = \text{gray level transitions rate}$$

- Spatial distribution and spatial dependence.
2.3.2. GLCM

\[ \text{Inertia} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j)(i - j)^2 \]

\[ \text{Entropy} = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \log(p(i, j)) \]

\[ \text{Energy} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j)^2 \]

\[ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left( i \cdot j \cdot p(i, j) \right) - \mu_x \mu_y \]

\[ \text{Correlation} = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left( i \cdot j \cdot p(i, j) \right) - \mu_x \mu_y}{\sigma_x \sigma_y} \]

\[ \text{Homogeneity} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{1}{1 + (i - j)^2} p(i, j) \]

*Haralick (1976)*
2.3.3. Color


\[ P(b) = \frac{N(b)}{M} \]

\[ S_M = \bar{b} = \sum_{b=0}^{255} b \times P(b) \]

\[ S_D = \left[ \sum_{b=0}^{255} (b - \bar{b})^2 \right]^{1/2} \]

\[ S_S = \frac{1}{\sigma^3} \sum_{b=0}^{255} (b - \bar{b})^3 P(b) \]

\[ S_k = \frac{1}{\sigma^4} \sum_{b=0}^{255} (b - \bar{b})^4 P(b) - 3 \]

\[ S_N = \frac{1}{\sigma^3} \sum_{b=0}^{255} [P(b)]^2 \]

\[ S_E = -\frac{1}{\sigma^3} \sum_{b=0}^{255} P(b) \times \log_2[P(b)] \]
2.4. Leuko classification

- Learning: *supervised*;
- Decision rule: *maximum likelihood*;
- Pdf estimation: *parametric*;
- Pdf: *Gaussian*;
- Error estimation methods regarded:
  - Resubstitution, holdout, cross-validation, leave-one-out.
Leuko

START

1. Blood slide image

2. Train computer to differentiate ROI colors

3. Resize image using G channel

4. Segment color image in ROI

User interaction

5. Extract features

6. Select features

7. Train computer leukocyte classes

8. Classify leukocyte

STOP

Get next image

yes

no
Leuko - leukemia diagnosis using pattern recognition

Area Distribution

- Nucleus: 43.15%
- Cytoplasm: 35.27%
- RBC: 15.14%
- Background: 6.45%

Subclass: granulocyte
Class: neutrophil

File Edit Window Settings Tools Help

clad regions: N=2

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Perim</td>
<td>533</td>
</tr>
<tr>
<td>2. Area</td>
<td>10336</td>
</tr>
<tr>
<td>3. MaxArea</td>
<td>3422</td>
</tr>
<tr>
<td>4. N elem</td>
<td>2</td>
</tr>
<tr>
<td>5. S/A/P</td>
<td>0.59429</td>
</tr>
<tr>
<td>6. Conv pixs</td>
<td>6</td>
</tr>
<tr>
<td>7. Conc pixs</td>
<td>14</td>
</tr>
<tr>
<td>8. Bending</td>
<td>4.5113E-3</td>
</tr>
<tr>
<td>9. N holes</td>
<td>1</td>
</tr>
<tr>
<td>10. Area holes</td>
<td>21</td>
</tr>
<tr>
<td>11. N Entia</td>
<td>1.2376E+0</td>
</tr>
<tr>
<td>12. N Entropy</td>
<td>9.694E-1</td>
</tr>
<tr>
<td>13. N Energy</td>
<td>2.2178E-1</td>
</tr>
<tr>
<td>14. N Con</td>
<td>2.1541E-1</td>
</tr>
<tr>
<td>15. N Hori</td>
<td>7.863E-1</td>
</tr>
<tr>
<td>16. N nuR</td>
<td>1.4003E+2</td>
</tr>
<tr>
<td>17. N axR</td>
<td>2.9547E+2</td>
</tr>
<tr>
<td>18. N skR</td>
<td>2.783E-5</td>
</tr>
<tr>
<td>20. N knR</td>
<td>3.256E-1</td>
</tr>
<tr>
<td>21. N enR</td>
<td>-2.7713E-1</td>
</tr>
<tr>
<td>22. N nuG</td>
<td>3.0545E+1</td>
</tr>
<tr>
<td>23. N axG</td>
<td>3.4654E+2</td>
</tr>
<tr>
<td>25. N knG</td>
<td>7.4011E-6</td>
</tr>
</tbody>
</table>

Classify | Save
3 Feature selection
3. Feature selection

- Problems of dealing with many features:
  - Overfitting \((\text{decision surface} = \text{many details})\)
  - Overtraining \((\text{decision surface} = \text{few details})\)
  - Curse of dimensionality \((\text{lacking samples})\)
  - Peaking phenomena.

- \(\therefore\) Define feature subsets using strategies:
  - Exhaustive;
  - Heuristics;
  - Random;
3. Feature selection

- Labeled samples
  - Searches
    - Exhaustive
    - FSS
    - BSS
  - Distance
    1. Euclidean
    2. Mahalanobis
  - Classifier
    1. Bayesian
  - Criterion function
    - Mean distance
    - Accuracy rate
4 Results
4. Results

- Differentiation among 6 types of leukocytes, including the CLL (the most common leukemia in adults from Western);

<table>
<thead>
<tr>
<th>%</th>
<th>baso</th>
<th>eosin</th>
<th>linnf</th>
<th>mono</th>
<th>neut</th>
<th>LLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>baso</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eosin</td>
<td></td>
<td>86</td>
<td></td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>linnf</td>
<td></td>
<td>92</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>mono</td>
<td>3</td>
<td>13</td>
<td>81</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>neut</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>LLC</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>86</td>
<td></td>
</tr>
</tbody>
</table>

- Project and development of pattern recognition and feature selection tools.
5

Latest developments
5.1. Leuko + SVM

- SVM extension to multiclass task, applying tree-based strategy;
- SVM and minimum spanning tree:
  - Adapted version of Kruskal algorithm
    - Generates a tree in a bottom-up iterative way
    - Polynomial complexity
      - Including weight assignment
    - Allows the **automatic** determination of the multiclass tree.
### 5.1. Leuko accuracy

<table>
<thead>
<tr>
<th>Technique</th>
<th>Baso</th>
<th>Eosi</th>
<th>Lymp</th>
<th>Mono</th>
<th>Neut</th>
<th>CLL</th>
<th>PL</th>
<th>HCL</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAA</td>
<td>92.5</td>
<td>96.4</td>
<td>90.8</td>
<td>90.6</td>
<td>97.6</td>
<td>86.8</td>
<td>79.5</td>
<td><strong>82.9</strong></td>
<td>91.5 (3.0)</td>
</tr>
<tr>
<td>AAA</td>
<td>91.0</td>
<td>93.6</td>
<td>91.2</td>
<td>89.6</td>
<td>97.1</td>
<td><strong>89.4</strong></td>
<td>78.0</td>
<td>72.9</td>
<td>90.6 (2.5)</td>
</tr>
<tr>
<td>ECOC</td>
<td><strong>95.0</strong></td>
<td>94.6</td>
<td><strong>94.0</strong></td>
<td><strong>91.7</strong></td>
<td><strong>98.8</strong></td>
<td>84.1</td>
<td>79.0</td>
<td>81.4</td>
<td><strong>92.1</strong> (1.9)</td>
</tr>
<tr>
<td>DDAG</td>
<td><strong>95.0</strong></td>
<td>95.5</td>
<td>92.2</td>
<td>91.6</td>
<td>97.2</td>
<td>89.3</td>
<td>73.5</td>
<td>75.7</td>
<td>91.5 (3.0)</td>
</tr>
<tr>
<td>ADAG</td>
<td><strong>95.0</strong></td>
<td>95.5</td>
<td>91.9</td>
<td>91.6</td>
<td>97.2</td>
<td>89.3</td>
<td>73.5</td>
<td>77.1</td>
<td>91.5 (3.0)</td>
</tr>
<tr>
<td>Tree</td>
<td>88.5</td>
<td>94.6</td>
<td>90.8</td>
<td>87.9</td>
<td>97.6</td>
<td>84.8</td>
<td><strong>80.0</strong></td>
<td>77.1</td>
<td>90.1 (2.3)</td>
</tr>
<tr>
<td>WW</td>
<td>90.0</td>
<td><strong>96.4</strong></td>
<td>89.4</td>
<td>86.0</td>
<td>96.3</td>
<td>84.8</td>
<td><strong>71.5</strong></td>
<td><strong>82.9</strong></td>
<td>89.6 (2.2)</td>
</tr>
<tr>
<td>CS</td>
<td>92.5</td>
<td><strong>96.4</strong></td>
<td>91.2</td>
<td>88.0</td>
<td>95.5</td>
<td>85.4</td>
<td>73.0</td>
<td>78.6</td>
<td>90.1 (3.0)</td>
</tr>
<tr>
<td>NB</td>
<td>90.5</td>
<td><strong>92.5</strong></td>
<td>63.7</td>
<td>73.1</td>
<td>70.2</td>
<td>72.8</td>
<td>74.0</td>
<td>75.7</td>
<td><strong>73.0</strong> (4.6)</td>
</tr>
</tbody>
</table>

30-May-07  
Computer Vision in Leukemia Diagnosis  
38/41
5.2. Level set

- \( \gamma(t) = \text{boundary of the nucleus in } \Phi(x,t=0) \)
- \( \Phi_t + F|\nabla \Phi| = 0 \)

\[
F = 1/(1 + \nabla (G*I))
\]
5.2.1. Colourful speed function

- Given C, the set of points of the nucleus contour and \( p \in C \): \( D_C(p) = \min(d(p,q)) \)
- If \( 0 < d(p,q) < \delta \),
  
  \[ \text{ColorSamplesSet} \leftarrow \text{RGB2Lab}(\rho(x,y)) \; ; \]
- Lightness and undersampling?
- Maximum likelihood:
  
  \[ ML = \arg\max_{h \in H} P(S|h) \]
Computer Vision in Leukemia Diagnosis

(a) Original image

(b) Evolution of the front

(c) Original image

(d) Evolution of the front

(e) Original image

(f) Evolution of the front
Acknowledgments:

- Unisantos: Profa. Marta Rosatelli, Daniella, Rodrigo, Aline, Rafael
- NIH: Dr. Edgar Rizzatti e Dr. Rodrigo Calado
- Instituto Fleury: Dr. Sérgio Martins
- Universidade de Sao Paulo: Prof. André Carvalho e Ana
- Universidade Federal de Santa Maria: Prof Marcos e Luciana
- Universidade Federal do Ceará: Profa Fátima Medeiros

Contact:

usdani@unisantos.br

http://www.unisantos.br/mestrado/informatica/daniela