Proceedings of International Conference on Image
Processing and Character Recognition

Segmentation of Renal Tubules by Knowledge-guided Image Analysis

Qi Chen and Andrew J. Taylor

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052 AUSTRALIA
{chen,andrewt}@cse.unsw.edu.au

Abstract

In this paper we present an approach for automatically recognising tubules and epithelial cell nuclei in renal biopsy sections. This approach employs image processing methods, histological knowledge and machine learning techniques. Potential applications of this work include the diagnosis of tubulitis, renal allograft rejection and other renal lesions.

1 Introduction

The analysis of histopathologic sections is an extremely important component in the diagnosis of many diseases. Manual examination of sections is routine practice in pathology laboratories, but has a number of significant drawbacks. It requires expensive training and extensive experience and expertise to analyse sections effectively [1]. Also, the pathologists who perform such analyses are subject to all of normal stresses encountered by human beings, leading to potential inconsistency or inaccuracy in diagnoses. In short, the process of manual pathological diagnosis is sometimes more subjective than one might like.

Computer-assisted diagnosis has the potential to provide cheap, reproducible and objective diagnoses. It also has the potential to provide better quantitative measures of the extent of a disease. However, computer-assisted diagnosis has proved difficult to achieve. The main reason appears to be that images from histopathologic sections are too poor quality to be segmented using simple image processing techniques [2]. Some systems have overcome this by requiring manual identification of the areas on the images where pathological problems occur [3, 4, 5]. Success with an expert systems approach for the detection of lesions in colonic, breast and prostate lesions has been reported at the Optical Sciences Centre of the University of Arizona [2, 6, 7, 8, 9, 10]. Also [11] describes a system to check for cancerous cells automatically in PAP smears, presumably a simpler problem than histopathologic sections.

The local segmentation of renal biopsy sections appears to pose a set of problems. Unfortunately these problems are sufficiently diverse that a single approach is unlikely to be successful. In this paper, we describe an approach that uses a combination of manually encapsulated histological knowledge and a decision tree automatically induced by a machine-learning system to successfully recognise tubules and epithelial cell nuclei in renal biopsy sections.

2 Data

2.1 Renal biopsy sections

The clinical material for this study comprises sections of needle renal biopsies, stained with periodic acid schiff (PAS). The sections were obtained from the Department of Anatomical Pathology, Prince Henry Hospital, Sydney. The sections were scanned through a laser scanning microscope (Olympic LSM-GB200) and each image was recorded in an array of 768x1024 with 256 grey levels.

2.2 Renal tubules

The human kidney contains a great number of functional units called nephrons. Each nephron consists of a glomerulus and renal tubules. Together with the glomerulus, renal tubules play an important role in maintaining the waste disposal system of the body [12, 13].

Each renal tubule has 4 segments, the proximal tubule, Henle's loop, the distal tubule and the collecting duct. All 4 segments are tubes of various lengths and diameters. The cavity at the centre of a tubule is usually referred to as its lumen. It is the proximal and distal tubule segments which are mostly seen in renal biopsy sections. An image of a renal biopsy section can be seen in Figure 1.

Tubules are lined by epithelial cells. The nuclei of these epithelial cells are obvious as small dark discs within tubules in Figure 1. Proximal tubules have microvilli lying around the inner part of the epithelium,

resulting in a dark layer in sections usually referred to as a brush border. The spaces between tubules is usually referred to as interstitium. Blood vessels also occasionally appear in renal biopsy sections.

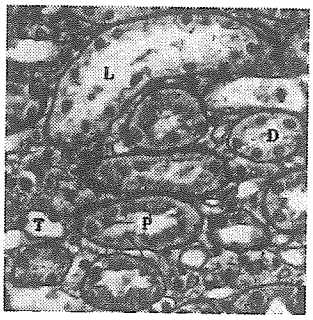


Figure 1: A renal biopsy section showing proximal tubules(P) and distal tubules(D). Also labelled are the lumen(L) of a tubule, either Henle's loop or a collecting duct(T) and packed microvilli (arrow).

2.3 Tubules in renal biopsy sections

Renal tubules are like a soft tube which may be squashed or twisted. Their morphology in a section varies considerably. As can be seen in Figure 1, the grey level of tubule boundaries in sections also varies considerably. The thickness of these boundaries in sections varies too, primarily due to cut angle. Typically a distinct bright lumen can be seen at the centre of each tubule but in some cases this is reduced or absent because of the cut angle or because the tubule is squashed or twisted. The epithelium lies between the tubule boundary and the lumen. The nuclei of epithelial cells are a distinctive feature in renal biopsy sections. As can be seen in Figure 1, their grey level varies considerably and they may touch the tubule boundary and have a similar grey level to this boundary. There is also some variation in nuclei size and shape.

3 Method

Our prototype system consists of two largely independent components: one to identify tubules, the other to identify epithelial cell nuclei. We describe each component separately.

3.1 Identifying tubules

We begin our processing of the image by using k-means clustering [14] to convert the 256 grey level image I into a 3-tone image I_3. This use of k-means clustering makes the subsequent processing somewhat independent of the overall brightness of individual slides. In other words it provides a degree of scale-invariance. The black areas of this 3-tone image will include tubule boundaries and brush borders and occasionally blood cells. The grey areas of this 3-tone image will include interstitium and epithelium. The nuclei of epithelial cells may be entirely black in the 3-tone image but many will contain grey areas as well. Occasionally they are mostly grey. The white areas of this 3-tone image are mostly the lumina of tubule and occasionally blood vessels. An example of a 3-tone image can be seen in Figure 7(b).

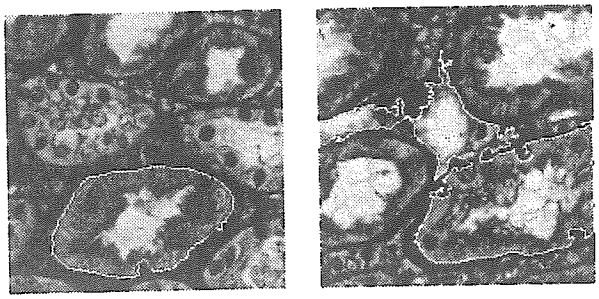
The first step is the identification of potential lumina. This is done by searching for large contiguous white regions. Initially we assume all such regions are tubule lumina. This white region is then enlarged to incorporate the entire tubule body. This is done in two steps. First any black pixels immediately adjacent to the white region are included. These are assumed to be part of a brush border. Region growing [15] is then used to incorporate any surrounding grey pixels. This should result in the region including most of the surrounding epithelium. However many of the nuclei of epithelial cells will not be incorporated by this region growing. They will remain as black regions, either isolated or touching the tubule boundary.

The black pixels surrounding an extracted tubule region form the initial estimate of the boundary of the tubule. Further processing is necessary as cell nuclei touching the tubule boundary or indistinct boundaries may make this initial boundary inaccurate.

Epithelial cell nuclei are typically roughly circular and 6-15 pixels in diameter in our images. Tubule boundaries are relatively smooth at this scale. The distortion of the boundaries caused by the epithelial cell nuclei is that the boundary is excavated and concavities are formed along the boundary. We can hence detect where epithelial cell nuclei touching the tubule boundaries have affected our initial estimate of the boundary by searching for the concavities of this shape and size. Such a concavity can be seen at the bottom of Figure 2(a).

We find these concavities by searching for three points on initial tubule boundaries which form a triangle of appropriate dimension. We also check that the triangle points towards the centre of the tubule. This check is necessary because otherwise the part of the boundary between two epithelial cells touching the tubule boundary could be incorrectly removed. When such a triangle is detected, we change the boundary to follow the base of the triangle.

Sometimes part of the boundary of a tubule will be indistinct. This means it will not appear as black pixels in our 3-tone image. As a result, the region-growing described above will then incorporated an



(a) Epithelial cell producing incorrect boundary

(b) Indistinct boundary

Figure 2: Problematic boundaries

area larger than the tubule body. The tubule may be merged with a neighbouring tubule or a non-tubule region. An example of this can be seen just to the right of the centre of Figure 2(b).

Experimentally we found that such gaps in tubule boundaries will be short relative to the size of tubules. This allows us to detect them and correct the boundary. We search for pairs of points lying on our initial tubule boundary which are only a short distance apart but lie a long distance apart measured along the boundary in either direction. When such a case is detected, we divide the region into two at this point.

Large continuous white regions in the images can be produced by features other than tubule lumina such as the interstitium. Examination of our data suggested that shape could be used to eliminate these non-tubule regions. We found experimentally that a compactness metric ($compactness = \frac{perimeter^2}{area}$) was useful [16]. The value of this metric for tubules is typically below 50 whereas for non-tubules it is almost always higher than 100. This metric is applied in the last step of tubule recognition to eliminate non-tubule regions.

The entire procedure of tubule identification can be summarised in Figure 3.

3.2 Identifying epithelial cell nuclei

Epithelial cell nuclei are dark, almost circular regions in the epithelium. They may appear disjoint from other objects, but often appear touching tubule boundaries or touching to each other. Some of them have a centre region that is somewhat lighter than the rest of the nucleus because of nuclei components.

We found identifying epithelial cell nuclei a difficult problem. We tried a variety of standard image processing techniques without success. The variation

```
/* input: a 3-tone image I3 */
foreach potential lumen {
        if enclosed by a brush border
                dilate (lumen);
        tubule_body = grow_region(lumen);
        boundary = trace(tubule_body);
}
foreach boundary repeat {
        if boundary contains multiple regions
                break(boundary);
} until no more corrections;

foreach boundary repeat {
        if a cell touches boundary
                correct(boundary);
} until no more corrections;

foreach boundary {
        if shape factor check ok
                keep boundary;
        else ignore boundary;
}
```

Figure 3: Tubule boundary identification

in shape and brightness of the nuclei defeated all the methods we tried. We also explored writing special purpose code to recognise epithelial cell nuclei but found there were a bewildering array of features that might be used to detect such nuclei and it was difficult to estimate the likely efficacy of each feature. This led us to the more principled approach of making an automatic search to find a subset of effective attributes which could be efficiently employed.

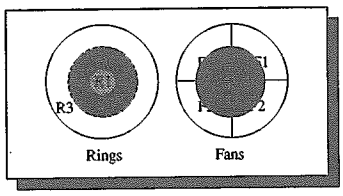


Figure 4: Regions used in nuclei attribute extraction

The set of attributes we began with are all constructed with respect to a single pixel which might be the centre of an epithelial cell nuclei. We constructed attributes from the region surrounding this pixel by calculating a number of measures such as means, standard deviations and gravities of disks, rings and fans centered at the pixel as illustrated in Figure 4. This produced about 70 attributes. Note both the original

image and the 3-tone image were used in constructing attributes. We also automatically constructed attributes from pairwise arithmetic relations of the above 70 attributes producing another about 20,000 attributes.

We then used a greedy search similar to what [17] term `forward selection` to choose a subset of the attributes which gives the smallest error rate of classification on the training data. 14 of these attributes were selected for our system to employ. These 14 attributes are listed in Figure 5.

No.	Attribute Description
1	standard deviation of pixel value of rings $R_{2,3}$
2	standard deviation of pixel value of rings $R_{1,2,3}$
3	standard deviation of the ring R_2 only
4	number of dark pixels in the rings $R_{2,3}$
5	number of light pixels in the rings $R_{2,3}$
6	mean of pixel intensity of the ring R_1
7	mean of pixel intensity of the ring R_2
8	mean of pixel intensity of the ring R_3
9	difference between 7 and 6
10	displacement of the centre of gravity of ring R_3
11	diff. between means of rings $R_{1,2}$ and R_3
12	diff. between fraction of dark_pixels in R_2 & R_3
13	diff. between fraction of light_pixels in R_2 & R_3
14	deviation of numbers of dark_pixel in fans

Figure 5: A set of attributes used for classifying epithelial cells

The search is done using a simple wrapper around Quinlan's machine learning system, C4.5 [18]. C4.5 is a supervised learning system which, given a set of classified cases and a number of attributes for each case as training data, produces a classifier in the form of a decision tree to classify further cases. Figure 6 shows a part of the decision tree that we used to identify epithelial cell nuclei.

A decision tree is constructed by the divide-and-conquer algorithm [18] from the set of training cases. The algorithm splits the training cases recursively into subset, based on a single attribute, until all the cases in a subset belong to a single class. A decision tree classifies a case by starting at its root and moving through it until a leaf node is encountered. At non-lead node, the test on the a single attribute value is carried out and determines to move on to the subtree corresponding the test outcome. In our case, the decision tree can be interpreted as

```
C4.5 [release 8] decision tree generator
----------------------------------------

  Options:
    File stem < tub >
Read 2196 cases (14 attributes) from tub.data

Decision Tree:
14_diff_g2Dark_g3Dark <= 35.37 :
  10_distance_ct_cg3 <= 19.53 :
    11_R2_SD <= 1.08 :
      3_disk_Rg2_mean > 2.18 : non_epi
      3_disk_Rg2_mean <= 2.18 :
        4_disk_Rg3_mean <= 1.86 : non_epi
        4_disk_Rg3_mean > 1.86 :
          5_disk_Rg2_SD <= 20 : non_epi
          5_disk_Rg2_SD > 20 : epi
    11_R2_SD > 1.08 :
      3_disk_Rg2_mean > 2.91 : non_epi
      3_disk_Rg2_mean <= 2.91 :
        8_R2_mean <= 87.63 : non_epi
        8_R2_mean > 87.63 : epi
  10_distance_ct_cg3 > 19.53 :
    3_disk_Rg2_mean > 2.67 : non_epi
    3_disk_Rg2_mean <= 2.67 :
      9_R3_mean <= 114.05 : non_epi
      9_R3_mean > 114.05 :
        10_distance_ct_cg3 > 29.47 : epi
        10_distance_ct_cg3 <= 29.47 : non_epi
14_diff_g2Dark_g3Dark > 35.37 :
  5_disk_Rg2_SD <= 25 : non_epi
  5_disk_Rg2_SD > 25 :
    14_diff_g2Dark_g3Dark <= 45.37 :
      8_R2_mean <= 77.55 : non_epi
      8_R2_mean > 77.55 :
        4_disk_Rg3_mean <= 1.8 : non_epi
        4_disk_Rg3_mean > 1.8 : epi
    14_diff_g2Dark_g3Dark > 45.37 :
      5_disk_Rg2_SD <= 59 : epi
      5_disk_Rg2_SD > 59 :
        12_diff_8_9 <= 31.83 : epi
        12_diff_8_9 > 31.83 : non_epi
```

Figure 6: Part of the decision tree for recognition of epithelial cell nuclei

```
if attr_14 <= 35.37 then
  if attr_10 <= 19.53 then
    if attr_11 <= 1.08 then
      if attr_3 > 2.18 then
        class non epithelial nucleus
      else if attr_3 <= 2.18 then
        if attr_4 <= 1.86 then
          class non epithelial nucleus
```

```
else if attr_4 > 1.86 then
  if attr_5 <= 20 then
    class non epithelial nucleus
  else if attr_5 > 20 then
    class epithelial nucleus
        ......
```

The training data for C4.5 was constructed by manually labelling the roughly 200 epithelial cell nuclei in an image. These were used to produce 200 positive cases. A further 2000 negative training cases were constructed by randomly choosing points in the image which were separated from any of the 200 labelled nuclei by at least a minimum distance.

A much larger number of negative cases is necessary because there is, naturally, much more variation in things that are not epithelial cell nuclei and the training data must reflect this variation.

Given only a set of 14 attributes, C4.5 produces a more accurate and smaller classifier. This is not unique to C4.5 or decision tree induction systems; the performance of many other classification systems declines as poor attributes are added. The computational time for constructing attributes is also much reduced by eliminating poor attributes.

4 Results

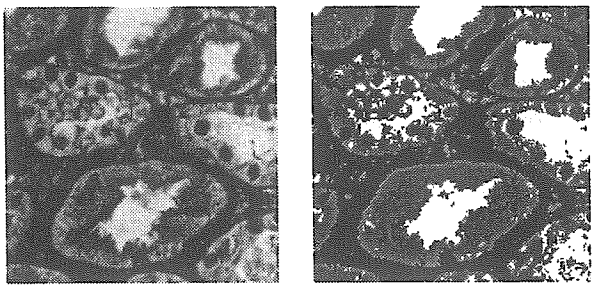
Figure 7(a) shows a typical tubule image. The initial estimate of tubule boundaries are affected by some epithelial cell nuclei because the cells are very close to or even touch the boundaries. Figure 7 also shows the corrected tubule boundaries and the result superimposed on the original image. Figure 8 shows an example where the image has a tubule with an indistinct part of the boundary. It was broken down first and then all the boundaries underwent concavity checking and compactness metric examination.

Figure 9 present some results of identification of epithelial cell nuclei. The centres of the nuclei are marked by white cross.

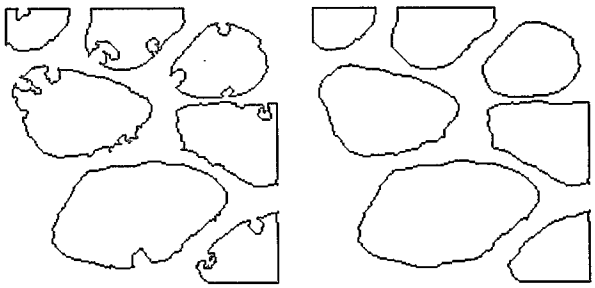
We have used renal biopsy sections from 4 patients as test data. This data was not used previously in the development of our software. We manually identified 293 renal tubules cross sections in these 4 slides. We then used the software described above to automatically label renal tubules and epithelial cell nuclei.

Our software correctly recognised 263 of the 293 tubule cross sections. It incorrectly labelled 5 regions as tubule cross sections. We then checked the accuracy of boundaries marked for each tubule cross section.

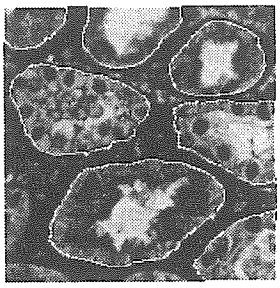
It was also noticed that among these 263 recognised tubule cross sections, there are 22 cross sections with a small part of the boundary not well located.



(a) original image (b) 3-tone image

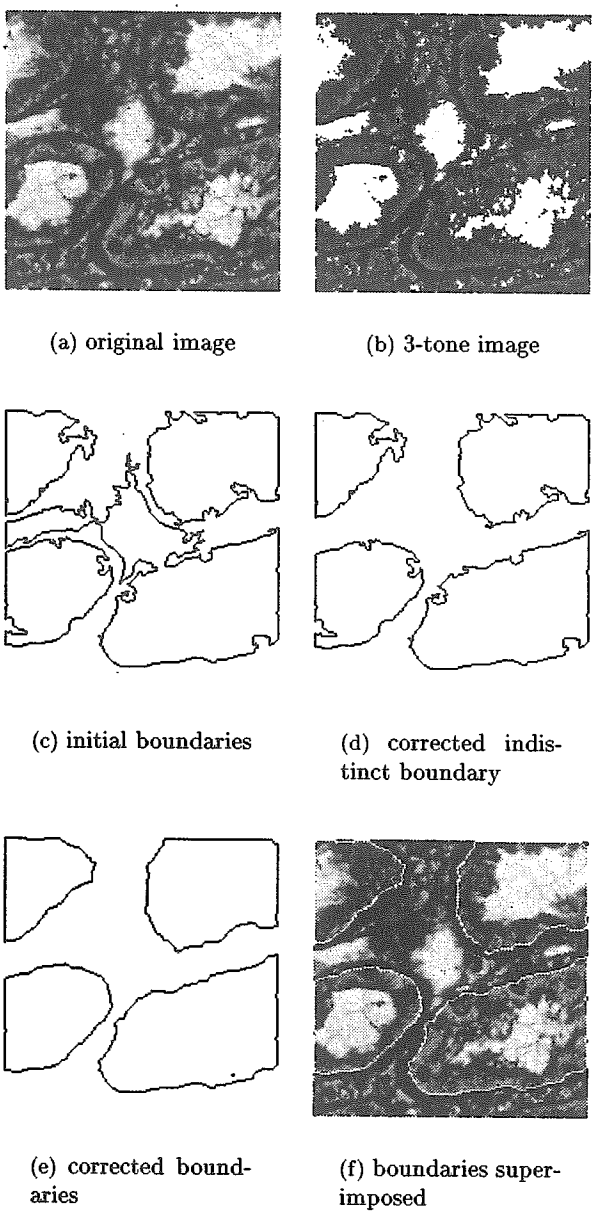


(c) initial boundaries (d) corrected boundaries



(e) boundaries superimposed

Figure 7: A common example



(a) original image (b) 3-tone image

(c) initial boundaries (d) corrected indistinct boundary

(e) corrected boundaries (f) boundaries superimposed

Figure 8: Indistinct boundary correction

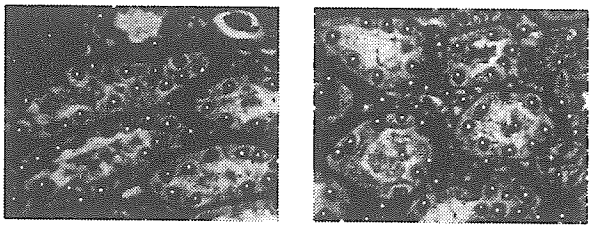


Figure 9: Examples of labelled nuclei

Evaluating the accuracy of epithelial cell recognition is more time consuming. We were able to evaluate the accuracy for the epithelial cell nuclei contained within 180 of the 263 tubule cross sections. Each of these tubules contained 10-12 epithelial cells. In 97 of the 183 tubule cross sections all the epithelial cell nuclei were accurately labelled. In 73 of the 183 cross sections 1 or 2 nuclei were not labelled. In 10 cross sections 3-5 nuclei were not labelled. In addition, in each cross section on average 2 nuclei were incorrectly labelled as epithelial cell nuclei.

5 Conclusion

These preliminary results are very promising. We feel they establish that these methods could be used in computer-assisted diagnosis. Our software already recognises tubules well and marks their boundaries accurately. Currently the complexity of the algorithm for boundary correction is $O(N^2)$, where N is the length of the boundary. We also feel we can modify the software to significantly improve on the performance above.

The recognition of epithelial cell nuclei is a more challenging problem but the employment of machine learning techniques and attribute selection has provided useful results which we had been unable to obtain otherwise. We would prefer to reduce the false positive rate and are sure we can do so to some degree. We intend to examine other possible attributes and also to enlarge the size of training data we provided.

References

[1] G. Brugal. Pattern recognition, image processing, related data analysis and expert systems integrated in medical microscopy. In *IEEE 9th International Conference on Pattern Recognition*, pages 286–293, 1988.

[2] D. Thompson, P.H. Bartels, H.G. Bartels, and R. Montironi. Image segmentation of cribriform gland tissue. *Analytical & Quantitative Cytology & Histology*, 17(5):314–322, October 1995.

[3] C.A. Beltrami. Structure analysis of breast lesions using neighbourhood graphs. *Analytical & Quantitative Cytology & Histology*, 17(2):143–150, April 1995.

[4] J. Linder. Automation in cytopathology. *American Journal of Clinical Pathology*, 98(4 suppl 1):314–322, October 1992.

[5] A.R. Brown. Combined immunocytochemical staining and image analysis for the study of lym-

phocytes specificity and function in situ. *Journal of Immunological Methods*, 130:111–121, 1990.

[6] P.H. Bartels, M. Bibbo, A. Graham, S. Paplanus, R.L. Shoemaker, and D. Thompson. Image understanding system for histopathology. *Analytical Cellular Pathology*, 1:195–214, 1989.

[7] A.R. Graham, S.H. Paplanus, and P.H. Bartels. A diagnostic expert system for colonic lesions. *American Journal of Clinical Pathology*, 94(4 suppl 1):S15–S18, 1990.

[8] H.E. Dytch, M. Bibbo, J.H. Puls, P.H. Bartels, and G.L. Wied. Software design for an inexpensive, practical, microcomputer-based dna cytometry system. *Analytical & Quantitative Cytology & Histology*, 8(1):8–18, 1986.

[9] J.E. Weber JE and P.H. Bartels. Colonic lesion expert system. evaluation of sensitivity. *Analytical & Quantitative Cytology & Histology*, 11(4):249–254, 1989.

[10] M. Bibbo, P.H. Bartels, T. Pfeifer, D. Thompso n, C. Minimo, and H.G. Davidson. Belief network for grading prostate lesions. *Analytical & Quantitative Cytology & Histology*, 15(2):124–135, 1993.

[11] A. Bykat. Computer automated pap smear diagnosis: A survey of recent work. *Journal of Clinical Computing*, 10(3-4):94–114, 1992.

[12] R.V. Krstić. *Illustrated Encyclopedia of Human Histology*. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1984.

[13] Thomas S. Leeson, C. Roland Leeson, and Anthony A. Paparo Inke. *Text/Atlas of Histology*. W.B. Saunders Company, Philadelphia, London, Toronto, Montreal, Sydney, Tokyo, 1988.

[14] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium*, pages 281–297, 1967.

[15] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison-Wesley Publishing Company, 1992.

[16] Dana H. Ballard and Christopher M. Brown. *Computer Vision*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1982.

[17] G.H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Mac hine Learning*, pages 121–129. Morgan Kauffman, 1994.

[18] J.R. Quinlan. *C4.5 : programs for machine learning*. Morgan Kaufmann Publisher, San Mateo, Calif, 1993.