군집을 이루는 자궁 경부암 세포 인식에 관한 연구

최예찬[0] 김선아 김호영 김백섭
한 림 대 학 교 컴 퓨 터 공 학 과
{ycchoi, sakim, hykim, bskim}@myrinac.ce.hallym.ac.kr

A Study on Recognition of Clustered Cells in Uterine Cervical Pap-Smear Image

Yea-Chan Choi[0] Sun-A Kim Ho-Young Kim Baek-Seop Kim
Dept. of Computer Engineering, Hallym University

요 약

Pap Smear 테스트는 자궁 경부암 진단에 가장 효율적인 방법으로 알려져 있다. 그러나 이 방법은 높은 위 음성률 (false negative error, 15~50%)을 나타내고 있다. 이런 큰 오류율은 주로 다량의 세포 검사에 기인하여, 자동화 시스템의 개발이 절실히 요구되고 있다.

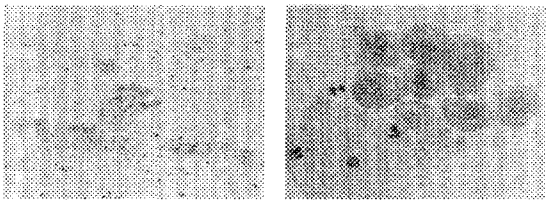
본 논문은 자궁 경부암의 특징인 군집을 이루는 암세포를 인식할 수 있는 시스템을 제안한다. 시스템은 두 부분으로 나누어 진다. 첫 단계에서는 저 배율(100배)에서 간단한 영상처리와 최소 근접 트리(Minimum Spanning Tree) 를 통해 군집을 이루는 세포를 찾는다. 두 번째 단계서는 고 배율(400배)로 확대하여 군집 세포들로부터 여러 가지 특징을 추출한 후 KNN(k -Nearest Neighbor)방법을 통해 인식하는 단계이다.

50개의 영상(640X 480, RGB True Color: 25 개의 100 배 영상, 25 개의 400 배 영상)이 실험에 사용되었다. 한 영상을 처리하는데 약 3 초(2.984초) 소요되었으며, 이는 region growing(20 초)나 split and merge (58 초) 방법 보다 덜 소요되었다. 100 배 영상에서 정상과 비정상의 두 그룹으로 나누었을 경우에는 96% 의 높은 인식율을 나타내었으나 비정상을 다시 5 개의 그룹으로 나누었을 때는 45% 로 나타내었다. 이는 영역추출(segmentation) 단계에서의 오류와 트레이닝 데이터의 비정확성에 기인한다. 400 배 영상에서는 각각 92% 와 30% 로 나타내었다. 이는 영역추출 단계에서 사용한 Watershed 방법의 오류로 기인한 것으로 본다.

1. 서론

전세계적으로 매년 50 만 건 이상이 보고 되는 여성 암으로서 가장 발생 빈도가 높으며 사망률도 매우 높은 질환인 자궁 경부암은 한국의 경우 여성암의 22% 을 차지하고 있으며, 여성 사망의 두 번째 빈도를 차지한다. 현재 자궁 경부암 진단 목적으로 가장 많이 쓰이는 Pap smear 방법은 그 자체의 높은 위 음성율(악성을 정상으로 판단하는 오류, False Negative Ratio : 15% ~ 50%) 로 인하여 여러 가지 문제점을 나타내고 있다[1]. 위음성율이 높은 이유는 표본 채취 상에서의 오류와 세포 검사 기사의 비정상 세포 판단상의 오류에 기인한다. 이 중 병리기사에 의한 비정상 세포에 대한 검색 오류의 원인은 한 사람의 기사가 하루에 약 500 만개의 세포를 검사하게 되고, 비 정상으로 진단이 내려진 슬라이드인 경우에는 단지 100 개 세포 중 1 개 미만의 비율로 비정상 세포가 출현하는 아주 적은 확률을 나타내기 때문이다. 문제를 더욱 악화시키는 점은 이렇게 지루하고 고된 일을 하는 기사를 키워내는데 많은 비용과 시간이 소요된다. 세포 인식 자동화는 이러한 병리기사의 비정상 세포 검색오류로 인한 위 음성률을 줄이고, 비용절감을 그 목적으로 한다.

자궁 경부 세포진 영상은 현미경에 카메라를 설치하여 Pap Smear 방법에 의해 얻어진 슬라이드를 통해 얻어진다. 자궁 경부암 검사의 자동화를 위해 제안하는 방법은 전문의에 의한 스캐닝 방법처럼 먼저 슬라이드를 100 배 배율에서 영상을 생성하여 초기 검사를 실시하고, 이상이 있는 부분에 대해서는 다시 400 배로 확대하여 정밀 검사를 실시하는 것이다. [그림 1]는 100 배와 400 배로 생성된 자궁 경부 세포진 영상을 보여주고 있다. 자궁 경부 세포진 영상의 경우, 핵과 세포질로 이루어진 자궁 경부 세포, 백혈구, 그리고 배경 등으로 구성된다.



[그림 1] 자궁 경부암 세포진 영상

주로 관심 대상이 되는 부분은 자궁 경부 세포의 핵 영역으로서 암을 판별하는 많은 요소가 집중되어 있다. 이러한 핵 영역을 추출하기 위해 많은 방법이 시도 되어져 왔다. 각 화소마다 독립적으로 특징 기준에 따라 의미를 부여하는 thresholding 방법과 주변 화소와의 관계를 계산하여 의미를 부여하는 region growing 방법과 split and merge 방법, edge detection 방법들이다[2][3]. 본 논문은 암 판별에 주요한 요인이 되는 군집 세포들의 영역을 추출하고 군집세포 내 세포들의 악성도를 인식하고자 하는 것이다. 먼저 군집세포를 간단한 thresholding 방법을 사용하여 100 배 영상에서 찾은 후 의심이 가는 부분을 400 배로 확대하여 그 지점에서의 핵 영역을 다시 간단한 영역 추출 방법을 사용한다. 이렇게 함으로서 한 슬라이드에서 항상 400 배 영상을 검사하는 것 보다 처리 시간을 단축할 수 있으며, 낮은 배율에서 최적의 특징 선택을 통해 효과적인 암세포 인식 시스템을 구성할 수 있으며, 추출된 영역에 대해 특징 추출과 인식기를 구성하여 세포의 정상 여부를 판단하도록 하였다

2. 본론

비정상 세포의 특징을 판별하는 대표적인 특징으로는 핵의 둥근정도와 핵과 세포질의 면적 비(N/C ratio) 혹은 핵 내부의 염색상의 균일성을

나타내는 거친 정도(texture feature) 등을 예로 들 수 있겠다. 그러나 이 외에도 표본 추출 시 많은 세포가 모여 군집을 이루는 경우에, 이런 군집 안에 비정상이 포함될 가능성이 크며, 군집 안의 세포핵들의 각 특징들의 변화정도를 측정함으로써 비정상을 판단할 수 있겠다. 그러나 이러한 군집에서 정확한 핵 영역을 찾기는 쉽지 않다. 본 연구는 이러한 군집 세포들에 대해 영상에서 군집을 찾는 방법과 군집 내에서의 단일 세포에 대한 인식과 군집 내의 세포들의 분산을 이용하여 인식하는 방법에 관하여 논한다.

2.1 시스템 구성

시스템은 크게 2 단계로 구성된다. 첫 단계는 100 배 영상에서의 전처리, 영역분할, 후처리, 군집세포 찾기(cluster cell detection), 특징 추출과 인식을 수행한다. 이 단계에서는 군집세포(cluster)를 찾는 과정 뿐만 아니라, 군집 세포에 대한 인식과 단위 세포(single cell)에 대한 인식 과정을 포함하고 있다.

그 이유는 단일세포에서의 암세포의 가능성을 배제할 수 없기 때문이다. 두 번째 단계는 100 배 영상에서 찾은 군집세포 영역을 400배로 확대하여 전처리, 영역분할, 후처리, 특징 추출, 인식 단계를 거쳐 보다 세부적인 인식을 수행한다.

각 방법의 이해를 돕기 위해 나머지 절에서는 RGB True Color, 640X480 영상인 [그림 1]을 초기 대상으로 실험 방법과 결과 영상을 실명하여 전체 시스템의 흐름을 설명한다.

2.2 전처리

2.2.1 명암 영상 생성

RGB(*Red, Green, Blue*) 모델에서 3 가지 색은 서로 높은 상관관계를 가진다. 이것은 영상처리 알고리즘을 그대로 실행하기는 힘들므로 이로부터 명암 영상을 얻는다.

본 실험에서 사용한 방법은 YIQ 모델 중 흑백TV 에서 방송되는 명암 성분인 Y 값을 얻기 위해 NTSC(National Television Standards Committee) 표준수식(식 2 -1)을 사용하여 3 채널의 칼라 영상에서 흑백 영상을 추출하였다.

$$Y = (0.299 * Red + 0.587 * Green + 0.114 * Blue) \quad (2\text{-}1)$$

2.2.2. 잡음 제거

데이터의 디지털화 전송 시에 잡음이 발생할 수 있다. 텔레비전 신호로부터 얻어지는 영상에서 흔히 볼 수 있는데, 이런 잡음은 보통 평균을 얻는smoothing을 사용하여 제거 될 수 있지만, 이는 가오시안 잡음제거에 알맞다. impulse 잡음이 있는 영상은 극대값이나 극적값을 가질 때가 많다. 메디안 필터링(median filtering)은 이런 *impulse* 잡음을 제거하는 데 좋고, 많은 정보를 갖는 에지(edge)를 그대로 유지하기 때문에 잡음 제거에 많이 쓰이는 방법이다. 이 방법은 영상 전체의 각각의 화소(pixel)에 대해서 수행되는데, 목적이 되는 화소를 중심으로 3X3 윈도우를 설정하여 윈도우에 포함되는 화소들에 대해 중앙값을 구한 후, 해당 화소의 값을 이 중앙값으로 대체한다.

2.3 영역 추출

영상으로부터 정보를 추출하는 영상 분석(*image analysis*) 단계 중 첫 단계는 일반적으로 영역 추출(image segmentation) 단계이다. 이 단계는 영상으로부터 관심 대상이 되는 객체의 부 영역을 찾는 단계이다. 이 단계는 영상 처리 분야에서 매우 어려운 분야 중 하나로 영상 분석의 성공 여부를 결정한다고 할 수 있다.

2.3.1 Kapur 방법에 의한 100 배 영상에서의 영역 추출

엔트로피란 정보량의 측정치이다. 즉, 어떤 symbol x 가 n 개 있다고 가정하면 symbol i 는 $p(X_i)$의 가능성을 가진다. 따라서 이 방법은 정보량을 최대로 하는 threshold 를 구하는 것이다. 엔트로피의 정의에 따라 몇 가지 변형이 있다. 본 실험에서는 Pun 의 방법을 수정한 Kapur 방법을 사용하였다. 먼저 Pun 의 방법을 살펴보면, 명암 값이 i 인 확률을 p_i 라고 하고 gray level 이 0 부터 t 까지 확률을 누적한 확률을 P_t 라고 하자. 원하는 객체의 엔트로피, H_o 와 배경의 엔트로피, H_b 를 각각

$$Ho = -\sum_{i=0}^{t} Pi \cdot \log(Pi), \qquad Hb = -\sum_{i=t+1}^{255} Pi \cdot \log(Pi)$$

로 정의하고 이 둘의 합을 최대로 하는 thresholding 방법을 제안하였다. 그러나 Pun 의 방법은 세포 영상에서 over - segmentation(원래 핵 영역보다 더 많은 부분이 나타남)하는 경향을 보였다. Kapur 는 Pun 이 제안한 엔트로피 함수를 다음과 같이 변경하여 사용하였다.

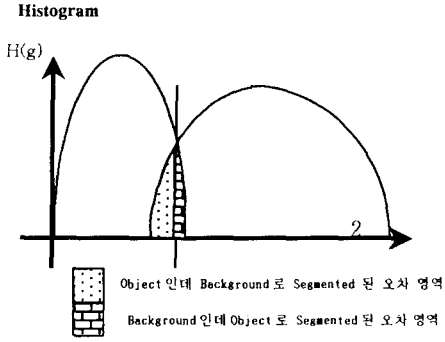
$$Ho = -\sum_{i=0}^{t} \frac{Pi}{Pt} \cdot \log(\frac{Pi}{Pt}), \qquad Hb = -\sum_{i=t+1}^{255} \frac{Pi}{1-Pt} \cdot \log(\frac{Pi}{1-Pt})$$

위 두 엔트로피의 합을 최대로 하는 t 값을 threshold 로 정한다. 이 방법을 사용한 이유는 핵의 영역보다는 핵의 위치에 초점을 맞추어볼 때 다른 방법보다 더 정확히 찾기 때문이다.

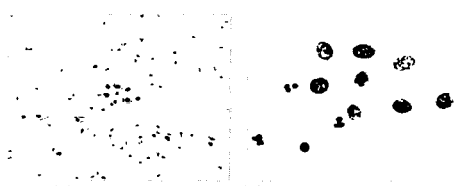
2.3.2 Minimum error 방법을 이용한 400 배 영상에서의 영역 추출

이 방법은 영상의 객체와 배경을 [그림 3] 과 같은 두 개의 histogram 을 정규 분포(normal distribution)로 가정하고 임의 threshold t 값을 기준으로 두 개의 정규분포로 나누었을 경우에 발생하는 이진화 오차(binarization error)를 최소화하는 t 값을 구하는 방법이다.

[그림 3] 에서 객체임에도 불구하고 배경으로 영역 분할된 영역과 배경임에도 불구하고 객체로 잘못 영역 분할된 영역의 합을 최소화하는 t 를 찾는 방법이다. 이는 두 개의 정규분포가 겹치는 부분이 적다는 것을 의미하고, 부정확하게 영역 분할 된 화소의 개수가 최소화되는 것을 의미하여, 결국은 다른 값에서 thresholding 작업을 하는 경우보다 영역 분할 결과의 신뢰도가 높다는 것을 나타내게 된다. 이 방법은 핵을 세포질과 배경으로부터 분리하는데 우수한 성능을 나타내었으며, 히스토그램에서 나타나는 두 개의 골(valley) 중에서 비교적 깊은 골이 세포질과 배경 사이의 골을 찾는 경우에 대해서 다른 방법들에 비해 안정적인 성능을 나타내었다



[그림 3] 두 개의 *Gaussian* 함수와 *Binarization error*



(a) Kapur Method (b) Minimum Error Method

[그림 4] 영역 추출

2.4 후처리

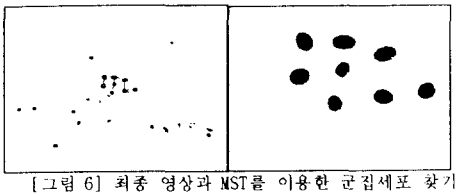
1 차 핵 분할에서 생성된 클러스터 영상에 대한 후 처리로는 boundary 에 접하는 영역을 제거하고, 홀을 채우는 작업(hole filling)을 수행한 후 핵의 gray level 과 비슷한 level 을 갖고 있으며 핵의 영역과 비슷한 넓이를 갖는 백혈구 영역을 제거했다. 이 때의 조건은 100 배 영상에서는 영역의 넓이가 46Pixel(정상핵의 평균 : 87.14404 - 정상핵의 표준편차 : 41.14519)이하를 갖는 Object 를 제거하였으며 400 배 영상에서는 영역의 넓이가 1066 Pixel 이하(정상핵의 평균 : 1770 - 정상핵의 표준편차 : 704)를 갖는 object 를 제거하였다. 400 배 영상에서는 Watershed algorithm[4]을 사용해 접하는(overlapped) 핵인 경우 자르는(split) 연산을 수행하였다. 400 배 영상인 경우 접하는 핵을 자른 이유는 over - segmentation 된 경우 그 핵의 영역이 너무 커서 악성으로 판단하는 위양성율(*false positive ratio*) 을 줄이고, 인식 단계에서 테스트 영상의 표본 데이터에 근접한 데이터를 추출하기 위함이다.

2.5 군집 세포 찾기

군집을 이루는 세포들은 크게 단일 세포들이 뭉쳐 있는 세포 그룹과 세포들이 뭉쳐 나와 판별이 어려운 세포 클러스터로 분류할 수 있다. 이것들은 암 조직에 서 떨어져 나올 수 있는 확률이 크므로 군집세포를

찾는 단계는 무척 중요하여 인식결과에 큰 영향을 줄수 있다. 군집 세포를 찾기 위한 방법으로는 각 세포핵간의 거리의 threshold 값을 사용하는 방법과 군집 핵을 포함하는 세포질을 찾는 방법 등이 있다.본 연구는 이들 중 세포가 군집을 이루지만 단일 세포를 분별할 수 있는 단일 세포 그룹이 있는 영상의 분석을 위한 시스템의 인식 파트를 위한 것이다.

100배 영상에서 후처리 후 얻어진 영상으로부터 군집 세포를 찾기 위해 먼저Dilation연산을 8 번 수행하였다. 이 때 군집 세포들을 찾기 위해서 MST(Minimum Spanning Tree) 를 사용하였다. 이 MST 방법은 군집세포를 찾는 방법뿐만 아니라, 군집세포의 특징으로도 좋은 성격을 갖는다. Dilation 후 영역 안에 있는 세포들 중 MST 의 최대거리가49 화소 내의 것들을 군집 세포로 정의하였다. 여기에서 사용한 MST 알고리즘은 Dijkstra/Prim 방법을 사용하였다.



[그림 6] 최종 영상과 MST를 이용한 군집세포 찾기

2.6 특징 추출

영역추출 후 얻어진 영상을 마스크로 하여 원 영상으로부터 특징을 추출한다. 어떤 특징 집합을 사용하느냐 하는 문제는 악성과 정상을 구별하는데 영역추출과 함께 중요한 요인으로 작용하지만, 이것은 매우 어려운 부분이다. 본 연구에서는 100 배 영상에서는 빠른 연산에 초점을 맞추었기 때문에, 명확한 비정상 세포를 찾는데 Geometric Feature(기하학적 특징)과 Densitometric Feature(밝기 특징)만을 사용하여 군집을 이루지 않는 세포 중에서의 비정상 세포(Rare Event Approach) 도 찾을 수 있도록 하였다. 또한 군집을 이루는 세포에 대한 악성도를 얻기 위해 군집 세포의 MACs(malignancy -associated changes) 를 나타내는 특징으로 핵의 방향 정보나 MST 정보를 사용하였다. 또한 400 배 영상에서는 보다 자세한 검사를 위해 실험에서 사용한 특징은 형태적인 특징(Morphmetry), 명암 특징(Densitometry), 질감 특징(Textural Feature)을 사용하였다.

2.7. 인식

100배 영상에서 추출된 특징으로부터 인식을 위한 방법으로는 간단한 방법으로서 MDC(Minimum -Distance Classifier)를 사용하였다. 이 방법은 유클리디안 거리(Euclidean Distance) 를 사용하여 원형 패턴과의 거리를 측정하여 가장 가까운 패턴의 타입으로 결정하는 방법으로 패턴 클래스 변수가 제한되어 나타날 때 문제 해결의 효과적인 도구로 알려져 있다.[5] 단일 세포를 위한 인식과 군집을 갖는 세포들의 인식을 각각 따로 구현하였으며 같은 MDC 알고리즘을 사용하였다. 트레이닝 데이터는 25 개의 영상에서 정상(WNL)에 대해 82 개, 비정상에 대해 67 개(악성도에 따라 ASCUS : 21, LSIL : 9 개, HSIL : 9 개, SCC : 28 개) Artifact에 대해 31 개를 표본 데이터로 사용하였다.400배 영상에서는 인식을 위해 k-Nearest Neighbor(kNN) 방법을 사용하였다. 이 방법은 Minimum Distance Classifier 가 가까운 한 개의 패턴의 타입으로 결정하는 반면, 기본적인k -Nearest Neighbor 방법은 가까운 거리에 있는k 개의 패턴 중 많은 패턴의 타입으로 결정하는 방법이다. 400 배 영상에 대해서는71 개의 영상으로부터 정상(WNL)에 대해 157 개, 비정상에 대해 743 개(ASCUS : 2, LSIL : 174 개, HSIL : 333 개, SCC : 234 개) Artifact에 대해 56 개를 표본으로 하였다. 본 실험에서는k 를 7 로 하였다.

2.8 실험 결과 및 평가

실험은 Intel Pentium 450 Dual Processor 에서 수행하였으며, 프로그램은 Visual C++ 6.0 을 사용하였다. 실험에 사용된 영상은 자궁 경부 세포진 영상으로 RGB True Color, 640X480 의 해상도를 갖는 100 배 영상 25 장과 400 배 영상 25 장을 사용하였다.

2.8.1수행 시간

[표 1] Time Duration on Each Step

	100X	400X
Preprocessing	281	297
Segmentation	140	94
Post Processing	1219	1218
Cluster Finding	406	
Feature Extraction	688	1563
Recognition	234	578
Total	2968	3250

평균 수행 시간은 100 배 영상에 대해서 3 초(평균 2984ms) 정도로 Region Growing(20 초)나 Split and Merge(1 분) 정도에 비해서 훨씬 적게 걸렸고, 400배 영상에 대해서 약 3 초(평균 3235ms) 정도 걸렸다.

2.8.2 인식률

100배 영상 25 장에 대해서 정상과 비정상에 대한 인식률은 96% 로 나타났지만,비정상에 대한 각 클래스의 인식률은 45% 로 비교적 낮게 나타났다. 100 배 영상에서의 형태적인 특징은 영역 추출을 위한 thresholding 값에 민감하기 때문에 정확한 영역 추출을 위한 알고리즘의 보완이 필요하다. 400 배 영상 25 장에 대해서 정상과 비정상에 대한 인식률은 92% 로 나타났지만, 비정상에 대한 각 클래스의 인식률은 30% 로 나타났다[표 2]. 이렇게 비정상에 대한 클래스 인식률이 저조한 원인은 영역 추출에 있어서 비정상에 대해 정확한 핵 영역을 찾지 못하였고, 트레이닝 데이터의 정확한 데이터베이스화의 실패에 기인한다.

[표 2] 100x ,400x Confusion Table

	Wnl	Ascus	Lsil	Hsil	scc	Wnl	Ascus	Lsil	Hsil	scc
Wnl	4	0	0	0	1	3	0	0	1	1
Ascus	0	1	1	2	1	0	0	0	4	1
Lsil	0	0	2	1	2	0	0	1	2	2
Hsil	0	0	0	2	3	0	0	0	0	5
Scc	0	0	0	1	4	0	0	0	0	5

3. 결론 및 향후 연구과제

Pap Smear 방법은 조기 검진을 위해 많은 효과를 봄에도 불구하고, 여러 가지 원인으로 인한 높은 위음성율로 신뢰성 있는 결과를 기대하기 어렵다. 본 연구는 이 문제를 해결하기 위한 자동화 시스템의 기반이 되는 부분으로, 특히 자궁경부 세포가 군집을 이룰 경우 이 군집세포를 검색하여 인식할 수 있는 시스템의 설계에 그 목적이 있다.

먼저100배 영상에서 군집세포를 찾는 단계로 간단한 thresholding 방법을 사용하여 영역을 추출하고, 군집이 될 가능성을 MST 를 사용하여 찾는다. 각 클러스터와 단일 세포에 대해 인식을 수행 후 이상이 있을 경우, 400 배로 확대하여 좀 더 세밀한 특징을 추출하여 인식할 수 있는 두 번째 단계이다. 평균 수행 시간은 100 배 영상에 대해서 3 초, 400 배 영상에 대해서 약 3 초(평균 3235ms) 정도 걸렸다. 100 배 영상 25 장에 대해서 정상과 비정상에 대한 인식률은 96% 로 나타났지만, 비정상에 대한 각 클래스의 인식률은 45% 로 비교적 낮게 나타났다. 400 배 영상 25 장에 대해서 각각 92%, 30% 로 나타났다. 세포의 핵들이 심하게 겹치는 경우 Watershed 알고리즘을 통해서 자른 경우 정확한 정상 세포핵의 크기 만큼 자르지 못했기 때문이다. 따라서 이 방법을 임상에 적용하기 위해서는 핵의 농도 변화에 따른 핵 영역을 정확히 찾을 수 있는 알고리즘이 개발되어야겠다. 또한, 정확한 세포영상의 데이터 베이스화를 위해 의사의 지식이 절실히 요구되며, 현미경의 정확한 군집세포로의 이동과 초점 자동화의 연구가 계속 이루어져야 하겠다

【참고 논문】

[1] Heinz K. Grohs and O.A. Nasseem Husain, Automated Cervical Cancer Screening, Igaku-Shoin, Ch.2, 1994

[2] R. L. Cahn, R. S. Poulsen and G. Toussaint, "Segmentation of cervical cell images ", The Journal of Histochemistry and Cytochemistry, Vol., 25, No. 7, pp681~688, 1977

[3] Joan S. Weszka, " SURVEY : A Survey of Threshold Selection Techniques", Computer Graphics and Image Processing, Vol. 7, pp259~265, 1978

[4] J. R. Parker, Practical Computer Vision using C, John Wiley & Sons, Ch.5, 1994

[5] S. Arya and D. M. Mount, "Approximate nearest neighbor queries in fixed dimensions ", Proceedings of the 4 th Symposium on Discrete Algorithms, 1993, pp271~280