SNP 발견을 위한 TDGS (Two-Dimensional Gene Scanning) 영상의 분석

장환°, 박유나, 이복주
단국대학교 전자컴퓨터공학과
{hwan°, ypark, blee}@dankook.ac.kr

Automated Analysis of TDGS Image for SNP Discovery

Hwan Chang°, Youna Park, Bogju Lee
Dept. of Computer Engineering, Dankook University

요 약

게놈 프로젝트에 의해 인간 유전자 염기서열이 밝혀지면서 개개인의 유전자에 나타나는 SNP(Single Nucleotide Polymorphism)을 분석하여 질병의 진단과 예후, 치료와 예방이 미래에 가능하게 되었다. 본 논문은 그러한 SNP 분석을 위한 자동 분석 시스템의 영상 처리 과정으로서, 기존의 육안을 통해 분석하였던 TDGS 영상을 본 시스템의 자동적인 영상 처리 과정을 통해 SNP 분석을 위한 디지털 패턴을 추출한다. SNP 분석을 위해 사용되는 샘플은 대략 수백개가 되는데, 실험이라는 특성상 영상에 나타나는 불규칙한 요소들이 많고, 영상의 상태가 좋지 않은 경우 명암도가 낮은 반점들의 구분이 힘들게 된다. 본 논문에서는 TDGS 영상의 지역적 특성을 가장 잘 반영하기 위한 동적 이진화의 새로운 척도를 제안하였고, 영상에서 잡영과 배경을 제거한 후 남겨진 관심영역을 반점으로 판별하여 이를 디지털 패턴으로 추출한 결과를 보여준다.

1. 서 론

현재 생명 과학 분야에서 당면한 중요한 과제는 막대한 DNA 염기 서열 정보를 분석하여 어떻게 인간의 건강과 복지에 이용할 수 있느냐는 것이다 [1]. 사람 개개인의 특정 질병에 대한 차이와 다양한 약물에 관한 반응성 및 효과는 개개인 별로 게놈 상에 나타나는 미묘한 차이 때문인데, 특히 그들 중 가장 흔한 변이인 SNP (Single Nucleotide Polymorphism)를 분석하여 질병의 진단과 예후, 치료와 예방에 이용함으로써 유전학 연구의 강력한 수단으로 이용할 수 있다 [2]. 이러한 SNP의 중요성과 유용성 때문에, 최근 1997년부터 산업계와 학계에서는 SNP의 발굴을 위한 대규모의 노력을 진행하여 왔다. 특히 특정 SNP의 유용성은 인종적으로 상당한 차이를 나타내므로 한국인 특유의 유전적 배경을 바탕으로 한 SNP의 발굴이 절실히 요구되고 있다[1].

TDGS는 SNP를 발굴하기 위해 사용되는 전기영동 기술로서 다른 방법에 비해서 정확성과 재현성을 증가시키면서 저비용과 고효율로 특정 유전자에서 지금까지 알려지지 않은 새로운 변이를 발굴하는 것을 가능하게 하는 새로운 방법이다. TDGS를 거쳐 나타나는 영상은 반점(spot)들이 2차원에 걸쳐 분리되며, 서로 다른 유전자마다 모두 개수와 패턴이 다양하게 나타난다. 본 논문에서 실험에 사용한 유전자는 BRCA1이라는 유방암 발현 유전자이다.

유용한 SNP 발굴을 위해서는 수백명의 샘플들을 분석하여야 하는데, 기존에는 이러한 TDGS 영상의 분석을 전문가에 의한 육안으로 식별해 왔기 때문에 소비되는 시간비용과 사람마다 분석의 차이를 나타내는 주관적 판단이 문제가 되지 않을 수 없었다. 본 논문에서는 TDGS 영상 분석의 자동화를 통해 human error를 줄이고 시간 비용을 절약함으로써, 추후 SNP 분석에 관한 연구를 효율적으로 수행할 수 있는 TDGS 영상 자동 분석 시스템을 제안한다.

본 시스템에서는 전기영동을 통해 나타나는 2D 영상을 입력으로 하고, 영상 처리의 과정을 통해 영상 내의 반점으로 나타나는 디지털 패턴의 결과를 출력으로 하는 시스템이다. 2장은 영상 자동 분석 과정에 대해서 간단하게 설명하고, 3장은 문제의 정의 및 본 논문에서 사용한 접근 방법에 대해서 설명하고, 4장에서는 결론 및 향후과제에 대해 설명한다.

2. TDGS 영상 자동 분석 과정

TDGS 영상에서 디지털 패턴을 추출하는 과정을 간단하게 설명하면 다음과 같다. 우선 Gaussian Smoothing, 고주파 필터 등의 전처리를 통하여 영상 내의 잡영을 제거하고, 반점들을 부각시킨 후 영상에서 배경을 제거한 1차 관심영역을 추출한다. 그리고, 1차 관심영역을 대상으로 모든 반점들이 구분된 2차 관심영역을 추출하여 이를 디지털 패턴으로 변환하여 결과로 출력한다(그림 2).

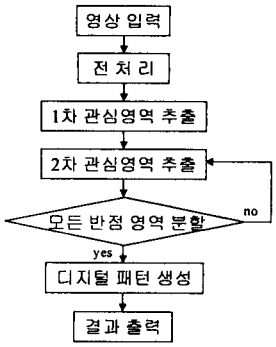


그림 2. 디지털 패턴 추출 과정

3. 문제 정의

TDGS 영상은 실험이라는 특성상 영상에 나타나는 불규칙한 요소들이 많고 영상의 상태가 좋지 않은 경우 명암도가 떨어지는 반점들의 구분이 힘들게 된다. 기존의 전문가의 육안에 의한 TDGS 영상 분석은 그러한 불안적 요소들에 대해 유연하게 대처할 수 있는 능력이 있었다. 하지만, 그러한 예외적인 경우를 컴퓨터가 처리하기 위해서는 영상의 지역적 상태에 맞는 융통성 있는 영상처리 과정이 필요하다.

3.1 1차 관심영역 추출

1차 관심영역은 배경에서 반점에 해당하는 부분을 분할 하기 위한 과정으로 영상 분할에 있어서 가장 일반적인 방법인 이진화 방법을 사용한다.

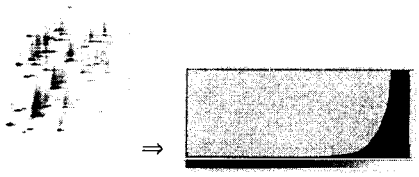
이진화는 다음과 같은 함수 T에 대한 연산으로 볼 수 있다.

$$T = T[\, x, y, p(x, y), f(x, y)\,] \qquad (1)$$

여기서 f(x, y)는 픽셀(x, y)의 명암도이며, p(x, y)는 해당 픽셀의 국부적 성질을 나타낸다. 만약, T가 오직 f(x, y)에만 의존할 경우 이를 전역적 이진화라 하고, f(x, y)와 p(x, y)에 모두 의존하게 되면 이를 동적 이진화라 한다.[3]

전역적 이진화

영상의 히스토그램을 이용한 방법으로 크게 단일 임계치와 다중 임계치에 의한 이진화로 나뉜다. 여기서 다중 임계치에 의한 방법은 그림 3의 히스토그램 분포도를 보면 알 수 있듯이 대상 영역을 효과적으로 분리시키는 여러 개의 임계치을 설정하는 것이 어렵다. 단일 임계치에 의한 방법은 P-Tile 방법, Mode 방법, Iterative Selection, Adaptive Thresholding 등의 방법이 있다[4]. 하지만, 단일 임계치에 의한 이진화 방법은 인접 픽셀과의 지역적 특성을 고려하지 못하여 반점의 명암도가 배경과 유사한 경우 이를 배경으로 인식하여 정확한 식별을 못하게 된다.



(a) (b)

그림 3. TDGS 영상과 히스토그램

지역적 특성을 고려한 동적 이진화

부분적인 특징이 강한 영상은 각 부분마다의 임계치를 설정해야 하는데 여러 가지의 개선된 알고리즘 중 많이 사용되어지는 방법으로 블록 이진화 방법이 있다. 일반적인 블록 이진화 방법은 해당 픽셀을 둘러싸는 블록 마스크를 이용하여 블록 내 명암도의 평균을 이용한 방법으로 지문 인식에서 주로 사용되어 진다. 하지만 블록 내 평균을 이용한 블록 이진화 방법은 TDGS 영상에 적용하기는 힘들었고 본 논문에서는 1차 관심영역 추출을 위해 블록 이진화 방법을 개선한 동적 이진화 방법을 사용하였다.

우선 영상을 일정한 크기의 여러 구역으로 나누고, 각 구역마다 otsu 알고리즘에 의한 지역적 임계치를 구한다[5]. 그리고, N×M 크기의 블럭 마스크를 이용한 동적 이진화를 하게 되는데 블럭 마스크 내의 지역적 특성을 검사하기 위한 척도(measure)로는 블럭 마스크 내의 모든 픽셀의 분산의 차이, 그리고 해당 픽셀이 위치한 구역의 지역 임계치를 사용한다.

그림 4의 (a)와 (b)는 영상에 나타나는 위치별 히스토그램을 보여주고 있다. ①은 블록 내에 반점이 존재하는 영역이고, ②는 블록 내에 반점이 존재하지 않는 배경에 해당하는 영역이다. 반점이 존재하는 영역은 히스토그램의 분포가 넓게 분포되어 있는 반면, 배경에 해당하는 영역은 고명도에 히스토그램의 분포가 집중되어 있다. 그림 4(c)는 영상 내의 각 픽셀들에 대한 분산의 분포도를 보여주는 그림으로 분산이 높게 나타나는 피크(peak)에 가까운 부분은 그림 4(a)와 같이 블록 내에 반점이 존재하는 부분이고 분산이 낮게 나타나는 벨리(vally)에 가까운 부분은 그림 4(b)와 같이 배경에 해당하는 부분이다. 그리고 두 부분을 구분하기 위한 특정값 V가 존재한다. 분산 척도에 관한 정의는 다음과 같다.

n = 블록 내 픽셀의 개수,
x = 블록 내 픽셀의 명암도
m = 블록 내 명암도 평균

$$p(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

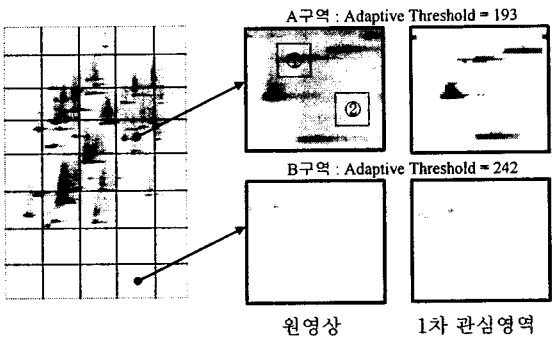
$$g(x, y) = \begin{cases} 1 \text{ if } p(x, y) > V \\ 0 \text{ if } p(x, y) \leq V \end{cases} \qquad (2)$$

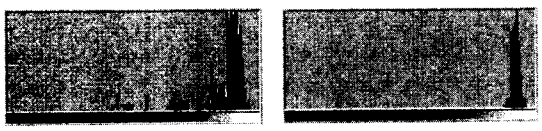
영상을 나누는 모든 구역들은 구역별로 픽셀들의 명암도의 분포가 다르므로 최적 임계치를 구하면 A구역과 B구역과 같이 지역적으로 다른 임계치를 갖게 된다. 지역별 이진화의 정의는 다음과 같다.

각 지역별 최적 임계치 T_i에 의해

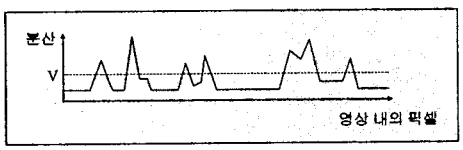
$$g(x, y) = \begin{cases} 1 \text{ if } f(x, y) > T_i \\ 0 \text{ if } f(x, y) \leq T_i \end{cases} \quad (3)$$

위의 (2)와 (3) 두 가지 정의를 이용하여 동적 이진화를 수행하게 되면 B구역과 같은 명암도가 배경과 유사한 반점도 식별이 가능하게 된다.





(a) ①영역의 히스토그램 (b) ②영역의 히스토그램



(c) 분산 분포도

그림 4. 동적 이진화에 의한 1차 관심영역

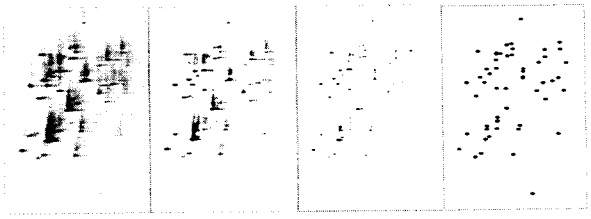
3.2 2차 관심영역 추출

반점들이 특정 위치에 밀집해 있는 경우 1차 관심영역의 추출을 통해서도 다수의 반점들이 하나의 영역 안에 속해 있을 수 있다. 디지털 패턴의 추출을 위해서는 관심영역 내의 반점들이 모두 별개의 영역으로 구분되어야 하므로 이를 검사하여 모든 반점들이 단일 영역을 갖도록 영역별 최적 이진화를 통하여 2차 관심영역을 추출한다.

3.3 디지털 패턴의 추출

최종적으로 나타나는 2차 관심영역은 반점들이 모두 별도의 영역으로 분할된 결과이다. 각 반점들은 디지털 패턴을 구성하는 객체들이 되므로 각 관심영역의 무게중심점을 구하여 이 점을 중심으로 하는 객체들로 구성된 디지털 패턴을 생성한다.

그림 5는 입력된 영상에서 디지털 패턴을 추출하는 과정을 그림으로 나타내었다.



(a) 원영상 (b) 1차관심영역 (c) 2차관심영역 (d) 디지털패턴

그림 5. 단계별 영상처리 결과

4. 결론 및 향후과제

본 논문에서는 2차원 그레이 영상인 TDGS 영상을 효과적으로 분석할 수 있는 영상 처리 과정을 제안하였다. 지역적 특성을 효과적으로 이용한 동적 이진화의 척도와 반복된 관심영역의 추출을 통해 명도가 낮은 객체도 정확한 구분이 가능하고, 수백개의 샘플을 전문가의 육안에 의해 분석을 해야 하던 기존의 작업을 영상의 자동 분석을 통해 많은 시간 비용을 줄일 수 있다.

디지털 패턴 생성은 SNP 분석을 위해 사용되는 후보 객체 생성 과정으로서 실제 유효한 패턴을 다시 구분해야 한다. 디지털 패턴의 객체 수가 많아지면 전체 SNP 분석을 위한 시스템의 부하가 커지게 되므로 유효한 디지털 패턴이 아닌 노이즈에 해당하는 객체를 구분하기 위한 연구가 요구된다.

참고문헌

[1] 서유신, Two-Dimensional Gene Scanning (TDGS)에 의한 SNP (Single Nucleotide Polymorphism) 발굴, 서울대학교 의과대학 암연구소, 2000

[2] N.J. van Orsouw, R.K. Dhanda, R.D. Rines, W.M. Smith, I. Sigalas, C. Eng, and J. Vijg, "Rapid design of denaturing gradient-based two-dimensional electrophoretic gene mutational scanning tests", Nucleic Acids Research, Vol. 26, No. 10, pp. 2398-2406, 1998

[3] Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing", Addison Wesley. pp 443-455, 1992.

[4] Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, "Machine Vision", McGraw-Hill Inc. pp 76-86, 1995.

[5] N. Otsu, "A threshold selection method from gray level histograms" IEEE Trans. Systems, Man, and Cybernetics, 9(1), p.62, 1979

[6] 권영희, 김진형, 윤수형, 윤세왕, "미생물 자동 분류 및 계수를 위한 영상 분석 시스템", 한국정보과학회논문집, 제29권, 제1호, pp. 607-609, 2002

[7] 장동혁, "Visual C++을 이용한 디지털 영상처리의 구현", 정보게이트. pp 258-260, 2001