ISSN 1975-8359(Print) / ISSN 2287-4364(Online)
The Transactions of the Korean Institute of Electrical Engineers Vol. 66, No. 8, pp. 1272~1277, 2017
http://doi.org/10.5370/KIEE.2017.66.8.1272

거리 기반 유사도 측정을 통한 유방 초음파 영상의 내용 기반 검색 컴퓨터 보조 진단 시스템에 관한 연구

A Study of CBIR(Content-based Image Retrieval) Computer-aided Diagnosis System of Breast Ultrasound Images using Similarity Measures of Distance

김 민 정* · 조 현 종†
(Min-jeong Kim · Hyun-chong Cho)

Abstract – To assist radiologists for the characterization of breast masses, Computer-aided Diagnosis(CADx) system has been studied. The CADx system can improve the diagnostic accuracy of radiologists by providing objective information about breast masses. Morphological and texture features were extracted from the breast ultrasound images. Based on extracted features, the CADx system retrieves masses that are similar to a query mass from a reference library using a k-nearest neighbor (k-NN) approach. Eight similarity measures of distance, Euclidean, Chebyshev(Minkowski family), Canberra, Lorentzian(F_2 family), Wave Hedges, Motyka(Intersection family), and Cosine, Dice(Inner Product family) are evaluated by ROC(Receiver Operating Characteristic) analysis. The Inner Product family measure used with the k-NN classifier provided slightly higher performance for classification of malignant and benign masses than those with the Minkowski, F_2, and Intersection family measures.

Key Words : Computer-aided Diagnosis(CADx), Breast Cancer, Ultrasound Images, Similarity Measures

1. 서 론

전 세계적으로 여성들의 사망 원인으로 유방암을 꼽을 수 있다. 그림 1과 그림 2를 통하여 국내뿐만 아니라 국외에서도 유방암이 암 발생의 상위권에 위치하고 있음을 확인할 수 있다. 유방암으로 인한 암 발생 및 사망률을 감소시키기 위한 가장 효과적인 방법으로는 정확한 검사와 진단을 통하여 초기에 암을 발견하여 치료하는 것이다. 초기에 암을 발견하여 치료하기 위하여 정확한 검사와 진단이 이루어져야 한다[1].

유방 초음파 영상의 컴퓨터 보조 진단(Computer-aided Diagnosis, CADx) 시스템은 객관적인 정보를 제공함으로써 영상의학과 전문의들이 암을 진단하는데 도움을 주기 위하여 연구되고 있는 분야이다. 현재 X선(X-ray)과 초음파 영상을 활용한 연구가 주로 진행되고 있다. 그런데 유방암 검진을 위하여 사용되는 X선이 유방암 발생에 영향을 미친다는 연구 결과에 따라 초음파 영상을 활용한 연구에 대한 중요성이 증가하고 있다[2]. 초음파 영상은 유방 종괴의 특성을 구분하기 위한 유용한 진단 방식이다[3]. 그림 3과 4는 각각 유방 초음파 영상의 악성 종괴와 양성 종괴를 나타낸다. 일반적으로 악성 종괴는 흐릿하거나 종괴의 경계가 명확하지 않으며 양성 종괴는 부드럽고 윤곽이 분명하다[4].

유방 초음파 영상의 CADx 시스템을 통하여 각 종괴의 특징들을 추출한 후 유사도 측정 방법을 적용하여 종괴의 특성을 추정할 수 있다. 각 유사도 측정 방법을 이용하여 종괴의 특성을 추정하여 영상의학과 전문의에게 객관적인 정보를 제공함으로써 암 진단에 대한 정확도가 향상될 수 있다. 유방 초음파 영상의 CADx 시스템을 설계하여 영상의학과 전문의에게 보다 객관적이

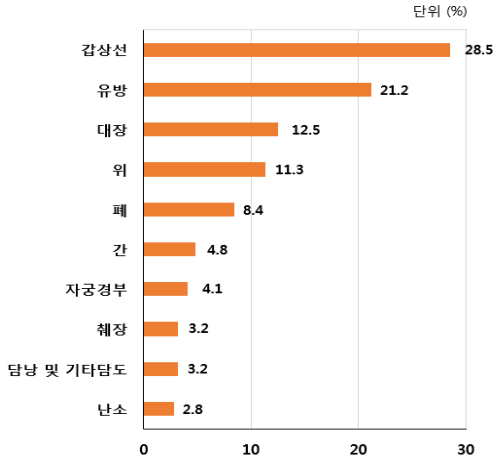


그림 1 여성 암 종별 발생률 (2014년도, 한국)
Fig. 1 Cancer occurrence of female(2014, Korea)

† Corresponding Author : Division of Electrical & Electronic Engineering and Interdisciplinary Graduate Program for BIT Medical Convergence, Kangwon National University, Korea.
E-mail : hyuncho@kangwon.ac.kr
* Interdisciplinary Graduate Program for BIT Medical Convergence, Kangwon National University, Korea.
Received : May 31, 2017; Accepted : July 28, 2017

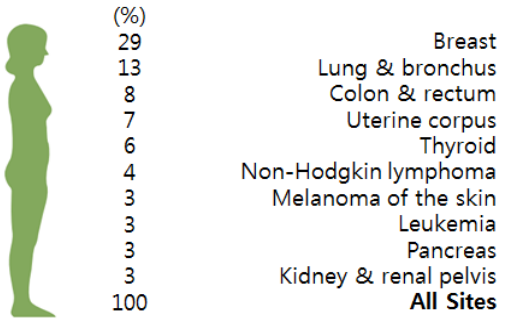


그림 2 여성 암 종별 발생률 (2016년도, 미국)
Fig. 2 Female cancer type incidence rate (2016, US)

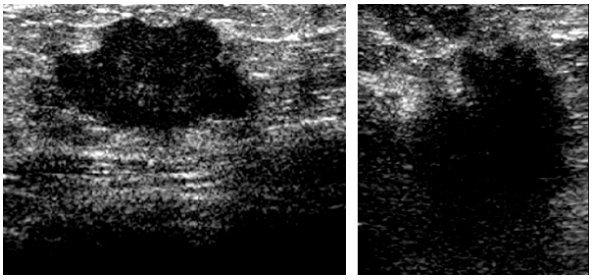


그림 3 유방 초음파 영상의 악성 종괴 예시
Fig. 3 Examples of malignant masses from ultrasound

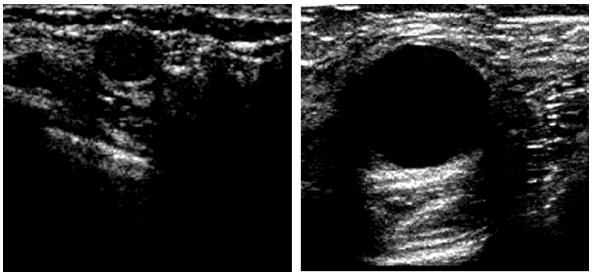


그림 4 유방 초음파 영상의 양성 종괴 예시
Fig. 4 Examples of benign masses from ultrasound

고 정확한 정보를 전달하는 것이 본 연구의 주요 목표이다.

2. 본 론

2.1 데이터베이스

본 연구에서는 유방 초음파 영상의 CADx 시스템을 설계하기 위하여 대학병원 영상의학과에서 검사를 받은 유방 초음파 데이터를 사용하였다. 모든 데이터는 조직검사에서 병리학적으로 입증되었다. 96개의 악성 유방 종괴와 154개의 양성 유방 종괴를 포함하는 총 250명 환자의 데이터를 취득하였다[5].

본 연구를 위하여 취득한 데이터를 임의로 두 집합 S_1과 S_2로 분류하였다. 각각의 집합은 각 종괴에 대하여 종괴를 가장 잘 표현하는 영상의학과 전문의가 선택한 두 가지 직교 초음파 영상들로 구성되었다. 이 중 일부는 직교 초음파 영상을 볼 수 없는 경우가 있는데 이러한 경우에는 하나의 초음파 영상만을 선택하였다. S_1 집합은 258개 영상(악성 종괴 55가지, 양성 종괴 74가지)을 포함하고 있으며, S_2 집합은 230개 영상(악성 종괴 41가지, 양성 종괴 80가지)을 포함하고 있다[1].

2.2 특징 추출 및 선택

본 연구에서는 초음파 이미지에서 유방의 종괴 부분을 분류하여 유방 초음파 영상의 CADx 시스템을 설계하기 위하여 이전에 설계된 활성 윤곽선 모델을 사용한 자동화 방법을 사용하였다[6]. 이는 종괴의 중심으로부터 활성 윤곽선 모델을 이용하여 종괴의 윤곽을 자동으로 추정하는 방법이다. 분류한 종괴에 대하여 여러 가지 특징을 추출하여 CADx 시스템이 질의 종괴의 특성을 추정하는 데에 사용하였다[6].

CADx 시스템을 설계하기 위하여 분류한 종괴의 형태학 특징(Morphological Features)과 질감 특징(Texture Features)을 추출하여 사용하였다[1]. 형태학 특징에는 악성 종괴에 대한 좋은 지표가 될 수 있는 종괴의 가로와 세로의 비율(종괴의 앞뒤로 긴 모양)과 후방 음영 특징(종괴의 내부와 가장 어두운 후방 부분의 정규화 된 평균 계조의 차이)이 있다. 이는 악성 종괴와 양성 종괴를 구별하기 위한 좋은 지표가 될 수 있다. 질감 특징으로는 공간 계조 의존성 배열 혹은 동시 발생 행렬에서 추출한 6가지(Information of Correlations 1 and 2, Difference Entropy, Entropy, Energy, and Sum Entropy)를 사용하였다[7].

선택된 특징들의 집합은 각 종괴에 대한 특징 벡터로 사용되며 두 집합 S_1과 S_2에 대하여 교차 검증 방법을 이용하여 훈련 집합(Training set)과 실험 집합(Test set)으로 분류하여 사용하였다.

2.3 유사도 측정

초음파 이미지에서 분류된 각 종괴를 특정할 수 있는 특징 벡터들이 참조 라이브러리에 저장되어 있다. 질의 종괴와 유사 종괴를 추출하기 위하여 내용 기반 영상 검색(Content-based Image Retrieval, CBIR) 시스템에 입력되면 참조 라이브러리에서 추출된 동일한 특징 벡터를 질의 종괴로부터 추출한다. 추출한 데이터의 특성을 이용하여 질의 종괴와 유사한 참조 라이브러리의 데이터를 추출하기 위하여 유사도를 측정하여 유사도 점수를 비교함으로써 질의 종괴와 유사한 종괴를 추출할 수 있다[5].

본 연구에서는 질의 종괴와 유사한 종괴를 추출하기 위한 유사도 측정 방법으로 거리에 기반한 8가지 방법을 적용하고 비교하였다. 8가지 유사도 측정 방법은 민코스키 부류(Minkowski family), F_2 부류(F_2 family), 교차점 부류(Intersection family), 내적 부류(Inner Product family)의 범주로 분류할 수 있다[8].

2.3.1 민코스키 부류(Minkowski family, F_1)

민코스키 거리란 유클리드 거리와 맨해튼 거리를 일반화하여 표현한 것으로 식 (1)과 같이 표현할 수 있다[9].

$$D_{Mks} = \sqrt[p]{\sum_{i=1}^{d} |q_i - r_i|^p} \tag{1}$$

이 때, q_i은 i번째 질의 종괴의 특징, r_i은 i번째 참조 라이브러리 종괴의 특징을 나타내며 d은 특징 공간의 차원을 의미한다.

민코스키 부류에는 유클리드 거리(Euclidean distance, D_{Eu}), 맨해튼 거리(Manhattan distance, D_{Man}), 체비쇼프 거리(Chebyshev distance, D_{Cheb})가 있다. 유클리드 거리는 다차원 공간상에서 두 지점 사이의 거리를 계산하기 위하여 가장 흔히 사용되는 방법으로 식 (2)과 같이 표현할 수 있다[10]. 맨해튼 거리는 두 지점의 좌표 간의 절대 값 차이의 합으로 계산할 수 있으며 식 (3)과 같이 표현할 수 있다[11]. 식 (1)에서 $p = \infty$인 경우를 체비쇼프 거리로 정의하며 식 (4)과 같다[9].

$$D_{Eu} = \sqrt{\sum_{i=1}^{d} (q_i - r_i)^2} \tag{2}$$

$$D_{Man} = \sum_{i=1}^{d} |q_i - r_i| \tag{3}$$

$$D_{Cheb} = \max |q_i - r_i| \tag{4}$$

2.3.2 F_2 부류(F_2 family, F_2)

민코스키 부류 중 절대 차이로 표현되는 맨해튼 거리에서 확장한 것이 F_2 부류이다. F_2 부류로는 캔버라(Canberra, D_{Can})와 로렌시안(Lorentzian, D_{Lo})이 있다. 캔버라는 두 벡터의 절대 차이를 벡터의 합으로 나눈 것으로 식 (5)와 같다[12]. 로렌시안 역시 두 벡터의 절대 차이를 포함하며 자연로그가 적용된 것으로 식 (6)과 같다[8].

$$D_{Can} = \sum_{i=1}^{d} \frac{|q_i - r_i|}{|q_i| + |r_i|} \tag{5}$$

$$D_{Lo} = \sum_{i=1}^{d} \ln(1 + |q_i - r_i|) \tag{6}$$

2.3.3 교차점 부류(Intersection family, F_3)

교차점 부류는 두 특징 벡터 사이의 교차점을 이용하는 방법으로 웨이브 헤지(Wave Hedges, D_{Wave}), 모티카(Motyka, D_{Mot}) 방법이 있으며 각각 식 (7), (8)과 같이 표현할 수 있다[8].

$$D_{Wave} = \sum_{i=1}^{d} \frac{|q_i - r_i|}{\max(q_i, r_i)} \tag{7}$$

$$D_{Mot} = \frac{\sum_{i=1}^{d} \max(q_i, r_i)}{\sum_{i=1}^{d} (q_i + r_i)} \tag{8}$$

2.3.4 내적 부류(Inner Product family, F_4)

유사도 측정의 마지막 범주로 식 (9)과 같이 표현되는 두 특징 벡터의 내적을 이용한 내적 부류가 있다. 내적 부류에는 식 (10)과 같이 두 특징 벡터 사이의 각도를 측정하는 코사인(Cosine, D_{Cos})과 식 (11)과 같이 표현되는 다이스(Dice, D_{Dice}) 방법이 있다[8].

$$D_{IP} = \sum_{i=1}^{d} q_i r_i \tag{9}$$

$$D_{Cos} = \frac{\sum_{i=1}^{d} q_i r_i}{\sqrt{\sum_{i=1}^{d} q_i^2} \sqrt{\sum_{i=1}^{d} r_i^2}} \tag{10}$$

$$D_{Dice} = \frac{\sum_{i=1}^{d} (q_i - r_i)^2}{\sum_{i=1}^{d} q_i^2 + \sum_{i=1}^{d} r_i^2} \tag{11}$$

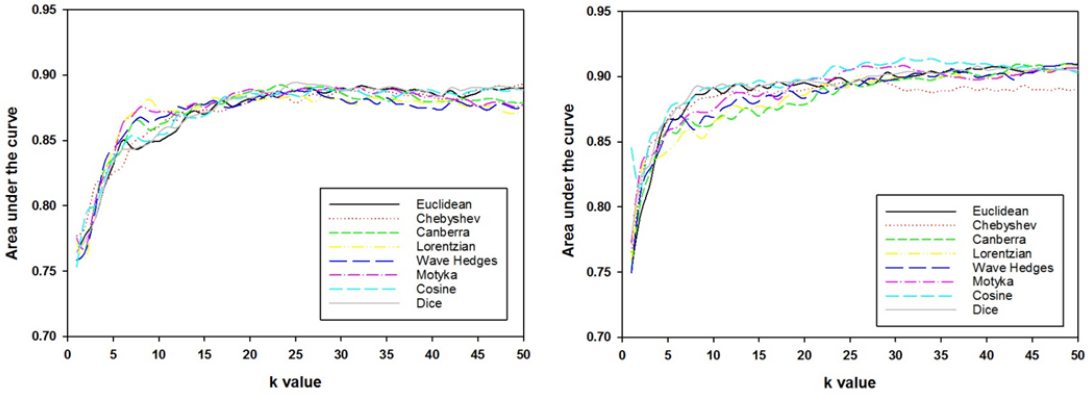
위와 같이 표기적 유사성에 의하여 다양한 유사도 측정법을 분류하였다[8]. 민코스키 부류는 유클리드 거리와 맨해튼 거리의 일반화된 방법이며, F_2 부류는 절대 차이를 이용한 방법이다. 교차점 부류는 두 특징 벡터 사이의 교차를 이용하는 방법으로 유사도 측정에 널리 사용되는 형태이다. 내적 부류는 내적의 형태를 포함하고 있는 유사성 척도를 사용한 부류이다.

3. 결 과

유방 초음파 영상을 위한 CADx 시스템을 설계하기 위하여 여러 가지 유사도 측정 방법을 적용하고 그 결과를 비교하였다.

거리 기반 유사도 측정 방법을 적용하기 위하여 취득한 데이터를 임의로 두 집합 S_1, S_2로 분류하고 두 집합에 대하여 교차검증 방법을 사용하여 훈련 집합과 실험 집합으로 분류하여 유사도 측정 방법을 적용하였다.

본 연구에서 질의 종괴와 참조 라이브러리 종괴 간의 유사도를 측정하기 위하여 적용한 유사도 측정 방법은 민코스키 부류(유클리드 거리, 체비쇼프 거리), F_2 부류(캔버라, 로렌시안), 교차점 부류(웨이브 헤지, 모티카), 내적 부류(코사인, 다이스)로 총 8가지이다. 각 유사도 측정 방법을 적용하여 ROC(Receiver Operating Characteristic) 분석을 수행한 결과는 각각 그림 5와 같다.



(a) (b)

그림 5 각 유사도 측정 방법의 성능: (a) 훈련 집합 S_1, 실험 집합 S_2 (b) 훈련 집합 S_2, 실험 집합 S_1

Fig. 5 Performance of similarity measures: (a) Training set S_1, Test set S_2 (b) Training set S_2, Test set S_1

Family	$k=3$	$k=5$	$k=10$	$k=25$	*Average*
Minkowski family	0.803	0.828	0.855	0.883	0.872
F_2 family	0.800	0.839	0.868	0.886	0.858
Intersection family	0.799	0.841	0.869	0.888	0.855
Inner Product family	0.795	0.835	0.855	0.889	0.869

(a)

Family	$k=3$	$k=5$	$k=10$	$k=25$	*Average*
Minkowski family	0.832	0.867	0.888	0.895	0.888
F_2 family	0.831	0.851	0.864	0.894	0.891
Intersection family	0.834	0.859	0.872	0.900	0.889
Inner Product family	0.842	0.870	0.891	0.900	0.894

(b)

Family	$k=3$	$k=5$	$k=10$	$k=25$	*Average*
Minkowski family	0.818	0.848	0.872	0.889	0.880
F_2 family	0.816	0.845	0.866	0.890	0.875
Intersection family	0.817	0.850	0.871	0.894	0.872
Inner Product family	0.819	0.853	0.873	0.895	0.882

(c)

표 1 유사도 측정 방법의 부류별 성능: (a) 훈련 집합 S_1, 실험 집합 S_2 (b) 훈련 집합 S_2, 실험 집합 S_1 (c) (a), (b)의 평균

Table 1 Performance of similarity measures by families: (a) Training set S_1, Test set S_2 (b) Training set S_2, Test set S_1 (c) Average of (a) and (b)

유방 초음파 영상의 CADx 시스템 설계를 위하여 적용한 8가지 유사도 측정 방법을 부류 별로 성능을 비교해 보았다. 그 결과가 표 1과 그림 6에 나타나 있다. 표 1은 검색된 종괴의 수 ($k=3,5,10,25$)에 따른 각 부류의 성능과 각 부류별 평균 성능을 나타낸 것이다. 그림 6은 검색된 종괴의 수($k=1,2,\cdots,50$)에 따른 각 부류별 성능을 그래프로 나타낸 것이다. 표 1과 그림 6의 (a)는 S_1을 훈련 집합으로 S_2를 실험 집합으로 하였을 때 검색된 종괴의 수(k)에 따른 각 부류의 성능을 나타낸 것이며, (b)는 S_2를 훈련 집합으로 S_1을 실험 집합으로 하였을 때 검색된 종괴의 수(k)에 따른 각 부류의 성능을 나타낸 것이다. 표 1의 (c)는 (a)와 (b)의 평균 성능을 나타낸다. 그 결과 4가지 부류의 성능이 비슷하지만 그 중에서도 내적 부류에 속한 유사도 측정 방법의 성능이 0.882로 약간 더 우수함을 보인다.

4. 결 론

영상의학과 전문의들의 유방암 진단에 대한 정확도를 향상시키기 위하여 CADx 시스템에 대한 연구가 진행되고 있다. CADx 시스템을 통하여 영상의학과 전문의에게 객관적인 정보를 제공하여 암 진단에 대한 정확도를 향상시키는 것이 본 연구의 주요 목표이다.

CADx 시스템을 설계하기 위하여 취득한 데이터를 임의로 두 집합으로 분류하였다. 분류된 두 집합에 대하여 교차 검증 방법을 사용하여 훈련 집합과 실험 집합으로 분류하여 유사도 측정 방법을 적용하고 ROC 분석을 통하여 각각의 성능을 비교하였다.

질의 종괴와 참조 라이브러리의 종괴 간의 유사도를 측정하기 위하여 민코스키 부류, 맨해튼 부류, 교차점 부류, 내적 부류의 4가지 부류에 대하여 각각 2가지 방법으로 총 8가지 거리 기반 유사도 측정 방법을 적용하였다. 그 결과를 ROC 분석을 통하여 비교 및 분석한 결과 8가지 방법 모두 거리에 기반한 방법이기 때문에 큰 차이를 보이지 않는다. 하지만 내적 부류에 속한 유사

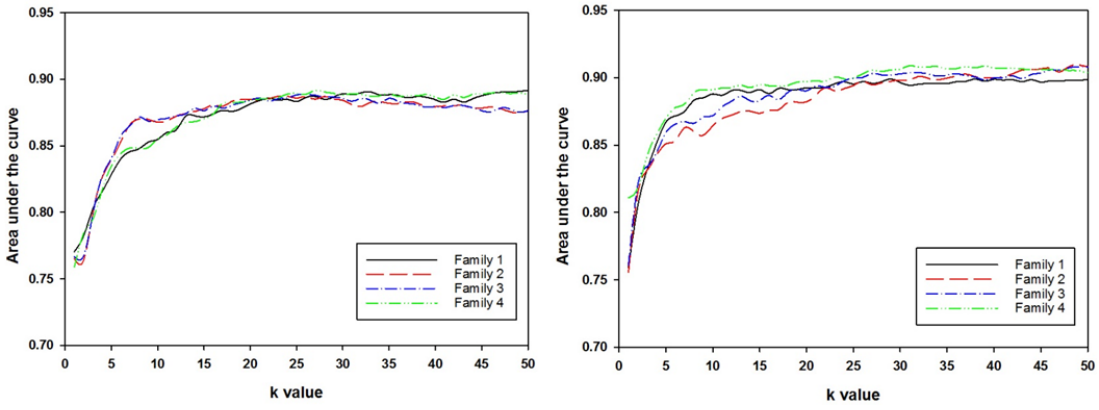


그림 6 유사도 측정 방법의 부류별 성능: (a) 훈련 집합 S_1, 실험 집합 S_2 (b) 훈련 집합 S_2, 실험 집합 S_1

Fig. 6 Performance of similarity measures by families: (a) Training set S_1, Test set S_2 (b) Training set S_2, Test set S_1

도 측정 방법의 성능이 나머지 부류에 속한 유사도 측정 방법들의 성능에 비하여 약간 더 좋음을 알 수 있다.

이후 연구에서는 유방 초음파 영상의 CADx 시스템의 성능을 향상시키기 위하여 더 다양한 거리 기반 유사도 측정 방법을 적용하여 그 결과를 비교 및 분석할 예정이다. 또한, 거리 기반 유사도 측정 방법보다 더 좋은 성능을 나타낼 것이라고 기대되는 서포트 벡터 머신(Support Vector Machine, SVM) 등의 분류기를 유사도 측정 방법으로 적용하고 그 결과를 비교 및 분석할 예정이다.

감사의 글

본 연구는 2016년도 강원대학교 대학회계 학술연구조성비로 연구하였음(관리번호-520160482).

References

[1] H. Cho, L. Hadjiiski, B. Sahiner, H. P. Chan, M. Helvie, C. Paramagul, *et al.*, "Similarity evaluation in a content-based image retrieval (CBIR) CADx system for characterization of breast masses on ultrasound images," *Medical Physics*, vol. 38, pp. 1820-1831, Apr 2011.

[2] P. Espín-López, A. Martellosio, M. Pasian, M. Bozzi, L. Perregrini, A. Mazzanti, *et al.*, "Breast cancer imaging at mm-Waves: Feasibility study on the safety exposure limits," in *Microwave Conference (EuMC), 2016 46th European*, 2016, pp. 667-670.

[3] H. Cho, L. Hadjiiski, B. Sahiner, H. P. Chan, M. Helvie, C. Paramagul, *et al.*, "A similarity study of content - based image retrieval system for breast cancer using decision tree," *Medical physics*, vol. 40, 2013.

[4] W. Yang, S. Zhang, Y. Chen, Y. Chen, W. Li, and H. Lu, "Effective shape measures in malignant risk assessment for breast tumor on sonography," in *Computer and Computational Sciences, 2008. IMSCCS'08. International Multisymposiums on*, 2008, pp. 51-56.

[5] H. Cho, L. Hadjiiski, B. Sahiner, H. P. Chan, C. Paramagul, M. Helvie, *et al.*, "Interactive content-based image retrieval (CBIR) computezr-aided diagnosis (CADx) system for ultrasound breast masses using relevance feedback," in *SPIE, Medical Imaging 2012*, 2012, pp. 831509-831509-7.

[6] J. Cui, B. Sahiner, H. P. Chan, A. Nees, C. Paramagul, L. M. Hadjiiski, *et al.*, "A new automated method for the segmentation and characterization of breast masses on ultrasound images," *Medical Physics*, vol. 36, pp. 1553-1565, May 2009.

[7] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, pp. 610-621, 1973.

[8] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City*, vol. 1, p. 1, 2007.

[9] K. Belattar and S. Mostefai, "Similarity measures for Content-Based Dermoscopic Image Retrieval: A comparative study," in *2015 First International Conference on New Technologies of Information and Communication (NTIC)*, 2015, pp. 1-6.

[10] N. Bouhmala, "How Good is the Euclidean Distance Metric for the Clustering Problem," in *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2016, pp. 312-315.

[11] M. D. Malkauthekar, "Analysis of euclidean distance and Manhattan Distance measure in face recognition," in *Computational Intelligence and Information Technology,*

2013. CIIT 2013. Third International Conference on, 2013, pp. 503-507.

[12] S. Viriyavisuthisakul, P. Sanguansat, P. Charnkeitkong, and C. Haruechaiyasak, "A comparison of similarity measures for online social media Thai text classification," in *2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2015, pp. 1-6.

저 자 소 개

김 민 정 (Min-jeong Kim)

1993년 03월 16일생. 2016년도 강원대학교 전자공학전공 졸업. 2016년~현재 강원대학교 BIT 의료융합학 석사과정

E-mail : mjeong9316@kangwon.ac.kr

조 현 종 (Hyun-chong Cho)

2009년 미국 플로리다 대학교 전기컴퓨터공학과 졸업 (석사, 박사). 2013년~현재 강원대학교 전기전자공학부 및 BIT 의료융합학 조교수