시·공간특징에 대해 적응할 수 있는 ROI 탐지 시스템

An Adaptive ROI Detection System for Spatiotemporal Features

박민철*, 최경주**
한국과학기술연구원 시스템 연구부*, 충북대학교 전기전자컴퓨터공학부**
Min-Chul Park(minchul@kist.re.kr)*, Kyung-Joo Cheoi(kjcheoi@chungbuk.ac.kr)**

요약

본 논문에서는 동영상에서 시간과 공간특징을 선택적으로 사용한 ROI(Region of Interest) 탐지 시스템을 소개한다. 동영상에서 명암도, 색상 등과 같은 공간특징을 사용한 공간상의 현저도 뿐만 아니라 시간상의 현저도도 얻기 위하여 모션이라는 시간특징을 사용하였다. 본 시스템에서는 동영상이 입력되면 공간특징 및 시간특징을 탐지하고, 이 특징과 관련된 기존의 심리학적 연구결과를 바탕으로 이들을 분석한다. 이렇게 분석된 결과는 하나로 통합되어 이를 기반으로 ROI 영역을 탐지한다. 일반적으로 비디오 영상에서 움직이는 개체나 영역은 같은 영상 안의 다른 개체나 영역보다 먼저 주의가 가게 되므로, 본 시스템에서는 분석된 결과를 통합하는 데 있어 시간특징인 모션을 공간특징보다 우선하여 통합한다. 시스템의 성능 분석을 위하여 동일한 실험영상을 가지고 인간을 대상으로 한 심리학적 실험을 우선 수행하였으며, 이 결과를 기준으로 본 시스템에서 얻어진 결과를 비교하여 모형의 성능을 분석하였다. 실험결과 공간특징만을 사용했을 때 보다 시간특징을 같이 사용함으로써 시스템의 성능을 보다 향상시킬 수 있었다.

■ 중심어 : | 시각적 주의 | 사공간특징 | 통합 | ROI 탐지 |

Abstract

In this paper, an adaptive ROI(region of interest) detection system for spatialtemporal features is proposed. It utilizes spatiotemporal features for the purpose of detecting ROI. It is assumed that motion representing temporal visual conspicuity between adjacent frames takes higher priority over spatial visual conspicuity. Because objects or regions in motion usually draw stronger attention than others in motion pictures. In case of still images visual features that constitute topographic feature maps are used as spatial features. Comparative experiments with a human subjective evaluation show that correct detection rate of visual attention region is improved by exploiting both spatial and temporal features compared to the case of exploiting either feature.

■ keyword : | Visual Attention | Spatiotemporal Features | ROI Detection |

* 본 연구는 한국전산원의 광개토 컴소시엄 : HFC/FTTH 기반의 통신 방송융합서비스 연구과제로 수행되었습니다.

접수번호 : #051122-001
접수일자 : 2005년 11월 22일
심사완료일 : 2006년 01월 09일
교신저자 : 최경주, e-mail : kjcheoi@chungbuk.ac..kr

I. 서 론

생물학적인 시스템은 주어진 시각장면을 모두 병렬적으로 처리하기 보다는, 보다 복잡한 고차원적인 처리를 위해 주의가 집중되는 일정한 영역들을 순간적으로 포착하여 그 부분만을 순차적으로 처리해가는 전략을 사용한다. 이러한 생물학적 시스템에 있어서의 시각적 주의는 실시간으로 입력되는 시각정보처리에 있어서의 복잡도 문제 및 처리용량의 한계를 극복한다는 측면에서 매우 중요한 동물의 지적 능력으로, 현재 컴퓨터 비전(Computer Vision)에서 주로 나누어서 처리되고 있는 전경과 배경의 분리와 인식의 문제를 효율적으로 결합할 수 있으나 아직은 충분히 활용되지는 못하고 있다.

시각적 주의 시스템에 관한 연구는 심리학, 정신물리학, 생리학, 신경과학 등 다양한 학문을 기반으로 연구되어져 왔으며, 공학적인 측면에서 시각적 주의 시스템은 영상 압축, 컴퓨터 시각(Computer Vision)에서의 목표물 자동탐지, 영상데이터베이스 등과 같은 몇몇 애플리케이션에 적용되어져 왔다[2].

인간의 대상(object)에 대한 인지과정을 이해하고자 하기 위해서는, 우선 우리의 시각 신경 시스템이 어떻게 입력된 사물의 시각 특징을 뽑아내는지, 그리고 이들을 어떤 방법으로 통합하여 대상을 인지하는지에 대한 과정을 우선 이해해야 한다[3][4]. 초기의 해부학자와 신경학자들은 PP(posterior parietal) 피질 영역이 입력된 대상의 여러 시각 특징들을 조합하여 하나의 통합된 공간 표상을 만들어냄으로써 이 대상의 각 부분 요소들에 대한 공간적 배치를 분석한다고 하였다[5]. 또한 Triesman은 '특징 통합 이론(Feature Integration Theory)'[6]을 통하여 우리의 눈에 입력되는 영상의 각 영역은 색, 모양 등의 기본 특징들의 조합으로 표현될 수 있으며, 이를 이용하여 특정 영역으로의 억제 및 주의 집중이 발생한다[7]고 하였다. 이와 같은 가정은 생리학적으로는 각 특징을 처리하는 시각피질 부분이 다르며, 단계적으로 이루어져 있다는 여러 연구결과들을 바탕으로 하고 있으며, 이런 내용을 따른다면 대상을 인식하기 위해서는 주의가 필요하고, 특히 여러 대상이 섞여있는 경우에는 하나의 대상과 관련된 영역에 선택적으로 주의를 줄 수 있는 능력은 필수적이라는 것이다. 이러한 이유로 인해 시각적 주의에 관련된 많은 연구들이 이 이론에 기초를 두고 있으며, 상당한 영향을 받고 있다.

이 이론을 바탕으로 연구된 기존의 시각적 주의 시스템에서는, 시스템에 영상이 주어지면 주어진 영상으로부터 색상, 명암도 등의 기본 특징들에 대한 특징 맵을 추출하고, 이들의 가중치 합으로 현저도 맵을 생성한 후, 이를 기반으로 가장 현저한 위치로 주의를 이동한다. 현저도 맵은 시각적인 환경에 놓여있는 특이할만한 대상들에 대한 정보를 가지고 있는 2차원 맵[1]으로, 시각장의 각 장소마다 현저한 대상을 표시함으로써, 이러한 공간적인 현저함을 기반으로 하여 주의가 가해진 장소를 찾을 수 있도록 가이드 한다. 여기서 현저도 맵의 구성은 이웃하는 비슷한 특징은 억제하고, 그렇지 않은 특징은 활성화시키는 경쟁 메커니즘에 바탕을 두고 있다[7-9].

본 논문에서 제안하는 시스템은 동영상을 처리할 수 있는 ROI 탐지 시스템으로서, 인간의 시각적 주의 능력을 모방하여 입력되는 동영상에서의 ROI를 탐지하도록 한다. 본 연구에서는 다른 자극들보다 두드러진 특정 부분에 대하여 집중하는 저 차원적인 시각적 주의 기능인 상향식 방식을 채택하고 있으며 Triesman의 '특징 통합 이론'을 바탕으로 한다. 제안하는 시스템은 동영상을 처리하기 때문에 움직임이 탐지되면 시간 특징으로써 모션정보를 사용하고 탐지되지 않으면 하나하나 프레임별 정지영상에서의 공간특징을 이용한다.

[그림 1]은 제안하는 시스템의 전반적인 동작과정을 보여준다. 그림에서도 볼 수 있듯이 시스템은 '분석'과 '통합'이라는 2개의 주요 모듈로 구성되어 있다. 제안하는 시스템에 RGB 컬러의 동영상이 입력되면, '분석' 모듈에서 밝기, 색상 등과 같은 공간특징과, 모션과 같은 시간특징이 추출되어 특징 맵을 구성하고, 이후 '통합' 모듈에서 이들이 각각 공간 현저도 맵과 시간 현저도 맵으로 통합된 후, 이 두 개의 현저도 맵이 다시 하나의 시·공간 현저도 맵으로 재구성되어진다. 각각의 초기 시각 특징에 대한 분석 및 통합은 각 특징과 관련된 기존의 심리학적인 연구결과를 바탕으로 이루어진다. 분석과정 중 추출된 특징에 대하여 인간의 동심원 형태의 "ON-중심, OFF-주변" 수용야(Receptive Field)에서

보이는 세포 반응도를 모방한 중심-주변 연산을 수행하는데, 이는 추출되어진 특징들 중 다른 주변의 특징들과 비교했을 때 현저히 차이가 나는 특징을 부각시키고 그렇지 않은 특징들은 억제시키는 역할을 한다. [그림 2]는 이러한 수용야의 예를 보여준다. 그림에서 나타나는 바와 같이 수용야의 세포의 반응도는 자극이 나타나는 수용야 영역의 위치나 크기에 따라 다르다. [그림 2]에서 점선으로 된 박스는 자극이 나타나는 영역을, 회색빛의 막대바는 자극이 나타나는 수용야의 영역 별 세포의 반응의 정도를 나타낸다. 자극 변화가 수용야의 "ON" 영역에 일어나면 이 세포의 반응도는 매우 커지는 반면([그림 2(b)] 참조), "OFF" 영역에 일어나면 반응은 일어나지 않으며([그림 2(c)] 참조), 수용야 전반에 걸쳐 자극이 나타나게 되면 세포는 이에 반응하지 않고 오히려 억제된다([그림 2(d)] 참조).

다음 2장과 3장에서 제안하는 시스템을 구성하는 주요 2개의 모듈인 '분석'과 '통합' 모듈에 대해 기술한다. 이어 4장에서 제안된 시스템의 실험결과를 보이고, 5장에서 결론을 맺는다.

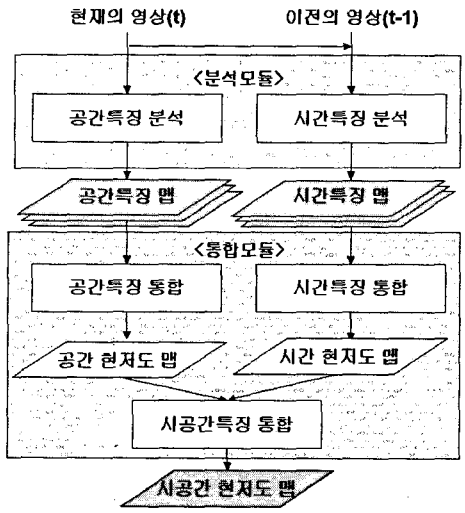


그림 1. 제안하는 시스템의 전반적인 동작과정

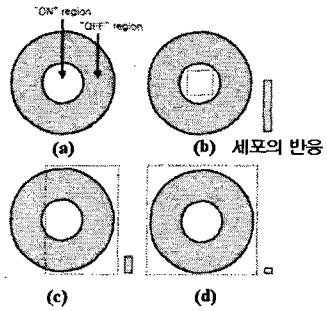


그림 2. 동심원 형태의 ON-중심, OFF-주변 모형

II. 사공간특징 분석

1. 공간특징 분석

[그림 3]은 본 시스템에서 사용된 공간특징 분석 방법을 보여준다. 위에서 이미 언급한 바와 같이 입력된 영상에서 몇 가지 특징을 추출하여 초기 특징 맵을 구성한 후, 구성된 특징 맵에 대하여 방위에 조율된 중심-주변 연산이 수행되어진다.

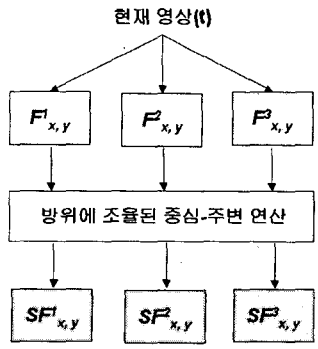


그림 3. 공간특징 분석 모듈의 전반적인 흐름도

본 시스템에서는 공간특징으로서 밝기와 색상 특징을 채택하였다. 색상은 인간의 시각이 물체를 구분할 수 있는 가장 큰 특징 중의 하나로 주의를 생성하는데 큰 역할을 하며, 밝기 역시 시각장면의 현저한 영역을 선택하는데 유용한 특징이 될 수 있고, 색상정보가 사용될 수 없는 경우에도 유용하게 사용될 수 있다는 사실에 바탕하여 채택하였다.

인간의 망막에는 약 1억 2,700만개의 빛 자극에 대해 반응하는 시세포인 간상체(rod)와 추상체(cone)가 존재

하며, 이들은 빛의 강도와 형태, 그리고 색깔을 신경정보로 변환한다. 이 중 94%의 1억 2000만개에 해당하는 간상체는 빛의 강도에 매우 민감하여 색상에 대한 정보 없이 아주 약한 빛만으로도 물체를 감지할 수 있게 한다. 간상체들이 색상을 구분하는 데는 도움이 되지 않는 반면, 추상체들은 색상을 감지하며 또한 매우 정교하여 물체의 세부적인 면까지 보게 해 준다[11][12]. 인간의 색상 지각 능력은 망막에 있는 이러한 3가지 종류의 추상체(cone)들의 반응에 의해 발휘된다. 인간의 눈에는 여러 가지 파장의 빛을 3가지 종류의 색상으로 지각할 수 있는 3가지 종류의 추상체가 있으며, 이것은 각기 적색, 녹색, 청색의 빛 중 어느 하나에 반응하는 광색소를 가지고 있다. 이러한 추상체의 반응들은 신경절 세포에 넘겨지고 바로 LGN을 거쳐 색상정보를 뇌로 전달한다. 색상정보를 뇌로 전달해 주는 신경회로는 3가지 추상체들에 의한 정보를 '적/녹', '황/청'의 반대쌍의 색상정보로 바꾸어 전달한다[13].

본 시스템에서는 밝기정보를 이용한 무색상 관련 공간특징 맵 $F^1_{x,y}$과. 이러한 인간의 색상처리에 관련된 정보를 바탕으로 하여, '적/녹', '청/황'의 2가지 색상대립(color opponency)을 모델링 한 색상 콘트라스트에 대한 2개의 색상관련 공간특징 맵($F^2_{x,y}$, $F^3_{x,y}$)을 생성하였다.

무색상 공간특징 맵은 밝기정보를 이용하는데, 밝기정보는 컬러영상을 입력받았을 경우 색상정보를 사용하여 식(1)과 같은 형식으로 추출할 수 있다. 여기서, R, G, B는 각각 적색, 녹색, 청색을 나타내는 추상체 반응이라 가정한다.

$$F^1_{x,y} = (R+G+B)/3 \quad (1)$$

색상에 관련된 공간특징 맵은 만드는 순서는 다음과 같다. 먼저, 3가지 추상체 R, G, B로부터 $r=R-(G+B)/2$, $g=G-(R+B)/2$, $b=B-(R+G)/2$, $y=R+G-2(|R-G|+2)$와 같이 다른 색상과 조금도 희석되지 않은 순순한 색상에 조율된 채널 r, g, b, y를 생성한다. 이렇게 생성된 채널 r, g, b, y를 사용하여 '적/녹' 대립 세포에 따른 $F^2_{x,y}$을, '청/황' 대립 세포에 따른 $F^3_{x,y}$를 식(2)와 식(3)에 의해 만든다[11-12][14-15].

$$F^2_{x,y} = r-g \quad (2)$$

$$F^3_{x,y} = b-y \quad (3)$$

이렇게 만들어진 3개의 공간특징 맵 $F^k_{x,y}$(k=1,…, 3)은 각각의 특징 맵이 가지고 있는 특징값들의 국부적 경쟁력에 의해 각각의 특징 맵에 대응되는 공간특징 맵인 $SF^k_{x,y}$ (k=1,…, 3)으로 업데이트되어진다. 업데이트 방법은 이 장의 3절을 참고하길 바란다.

2. 시간특징 분석

현재 프레임(t)과 이전 프레임(t-1)으로 이루어져 있는 연속적인 명암도 영상 시퀀스를 입력으로 받아 시간특징인 모션정보를 추출한다[13]. 이 영상 시퀀스에서 블록매칭기법을 사용하여 모션벡터를 추출한 후 모션맵 $M_{x,y}$을 구성한다.

제안하는 시스템은 이 모션맵에 대하여 3가지 모듈을 통해 모션분석을 수행한다([그림 4] 참조). 본 절에서는 '이중대립수용야'와 '노이즈여과'에 관련된 모듈만을 설명하고, 방위에 조율된 중심-주변 연산은 3절에서 설명한다.

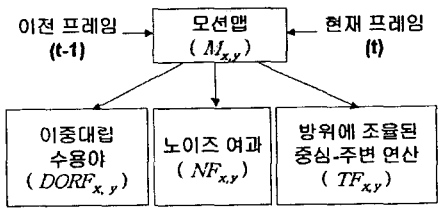


그림 4. 시간특징 분석 모듈의 전반적인 흐름도

2.1 DORF 모듈 : 이중대립수용야 모델링

많은 해부학적, 생리학적 연구들에 의하면 영장류의 대뇌 시각피질 영역 중 MT(middle temporal area) 영역이 모션분석에 있어 주요한 역할을 하고 있다고 한다[16]. 이 MT 세포의 반응도는 모션자극이 MT 세포의 수용야 내부 영역과 주변 영역에 어떻게 주어졌는가에

따라 다르다([그림 5] 참조). 수용야 내부 영역에 수용야의 최적방향으로 모션자극이 주어졌을 때 MT 세포는 최대로 반응한다. 그런데 위와 같은 모션자극이 있는 상태에서 수용야의 주변영역에 모션자극이 최적방향의 반대방향으로 주어지면 MT 세포의 반응은 더욱 증대된다. 그러나 만일 이때 수용야 내부영역에 모션자극이 주어지지 않는다면 다시 말해 수용야 주변영역에만 최적방향의 반대방향으로 모션자극이 주어지고 수용야 내부영역에는 아무런 자극도 주어지지 않는다면 세포의 반응은 거의 없게 된다. 이러한 "이중대립수용야"의 특성은 모션의 경계를 탐지하고, 탐지된 모션의 경계를 기반으로 하여 모션의 범위를 정의하는데 유용하게 사용할 수 있다[17]. DORF(Double Opponent Receptive Field) 모듈은 이러한 MT 세포의 "이중대립수용야" 성질을 모방하여 설계되었으며, [그림 5]와 같이 주변의 모션자극쌍의 방향이 최적방향과 반대방향이고, 중심의 모션자극이 최적방향으로 주어졌을 때 최대로 반응하도록 설계되었다[18]. [표 1]은 8개 방향에 따른 최적방향과 관련된 중심과 주변영역의 모션벡터 위치를 보여주며, 각 픽셀의 반응값은 8개의 방향 ($\theta \in (0, \pi/4, , ..., 7\pi/4)$마다 식(4)에 의해 계산된다. 식(4)에서 $V_c(\theta)$ 는 관찰자 시점에서 본 중심 모션벡터를 나타내고, $V_{s1}(\theta+\pi)$ 와 $V_{s2}(\theta+\pi)$는 주변 모션벡터를 나타낸다.

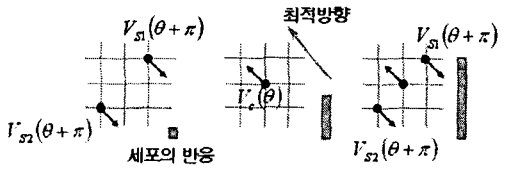


그림 5. "DORF"를 기반으로 설계된 DORF 모듈

표 1. 방향과 위치에 대한 DORF 모듈의 룩업테이블

방위	주변1 (S1)	중심 (C)	주변2 (S2)
0, 4π /4	(x+0, y−1)	(x, y)	(x+0, y+1)
π /4, 5π /4	(x−1, y+1)	(x, y)	(x+1, y−1)
2π /4, 6π /4	(x−1, y+0)	(x, y)	(x+1, y+0)
3π /4, 7π /4	(x−1, y−1)	(x, y)	(x+1, y+1)

$$DORF_{x,y} = \sum_{\theta}\left(\sum_{m,n} |R_{m,n}(\theta)| \right)$$

만일 $|V_c(\theta)| \neq 0$이면,

$$|R(\theta)| = \left| -\frac{1}{2} V_{s1}(\theta+\pi) + V_c(\theta) - \frac{1}{2} V_{s2}(\theta+\pi) \right| \quad (4)$$

그렇지 않을 경우,

$$|R(\theta)| = \left| -\frac{1}{4} V_{s1}(\theta+\pi) + V_{s2}(\theta+\pi) \right|$$

2.2 NF 모듈 : 노이즈 여과 모델링

MT 세포는 수용야에 하나는 세포의 최적방향으로 움직이고 다른 하나는 최적방향의 반대방향으로 움직이는 2개의 점을 제시하면 세포의 반응도가 매우 약해지는 성질을 가지고 있다([그림 6] 참조).

NF(Noise Filtration) 모듈은 이러한 MT 세포의 특성을 모방하여 설계되었으며, [그림 6]과 같이 4개의 방위(2π/4, .., 5π/4)를 가질 수 있는 주변의 모션벡터 쌍의 방향이 서로 반대방향이 되었을 때 반응하도록 설계하였다[18]. [표 2]는 4개 방향에 따른 최적방향과 관련된 중심과 주변의 모션벡터의 위치를 보여주며, 픽셀의 반응값은 식(5)에 의해 계산된다. 식(5)에서 $V_c(\theta)$는 식(4)에서와 마찬가지로 관찰자 시점에서 본 중심 모션벡터를 나타내고, $V_{s1}(\theta+\pi)$ 와 $V_{s2}(\theta+\pi)$는 주변 모션벡터를 나타낸다.

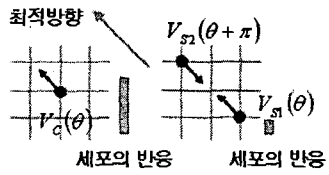


그림 6. "NF"를 기반으로 설계된 NF 모듈

표 2. 방향과 위치에 대한 NF 모듈의 룩업테이블

방위	주변1 (S1)	중심 (C)	주변2 (S2)
2π /4	(x+0, y−1)	(x, y)	(x+0, y+1)
3π /4	(x+1, y−1)	(x, y)	(x−1, y−1)
4π /4	(x+1, y+0)	(x, y)	(x−1, y+0)
5π /4	(x+1, y+1)	(x, y)	(x−1, y+1)

$$NF_{x,y} = \sum_{\theta}\left(\sum_{m,n} |R_{m,n}(\theta)| \right)$$

만일 $|V_c(\theta)| \neq 0$이면, $|R(\theta)| = |V_c(\theta)|$ (5)

그렇지 않을 경우,

$$|R(\theta)| = |V_{s1}(\theta) - V_{s2}(\theta+\pi)|$$

3. 방위에 조율된 중심-주변 연산

방위에 조율된 ON-중심, OFF-주변 연산[10][14][15][19]이 2.1절에서 구성된 공간특징 맵 $F^k_{x,y}(k=1,2,3)$과, 2.2절에서 구성된 모션맵 $M_{x,y}$에 수행되어진다. 이 연산을 통하여 국부적인 영역에서 다른 주변의 특징들과 비교하여 현저히 차이가 나는 특징을 부각시키고 그렇지 않은 특징들은 억제시킴으로서 각 특징 맵 별 중요도 측정치를 부여할 수 있다. 본 시스템에서는 이러한 중심-주변 연산을 위해 방위를 가진 다중 스케일의 필터 뱅크인 DOOrG 필터를 사용하였다[15]. DOOrG 필터는 양의 값을 가지는 가우시안이 음의 값을 가지는 가우시안보다 더 좁은 폭을 가지는 2개의 가우시안의 차(Differece of Gaussians)로 구성되어 있으며, 이 때 2개의 가우시안은 원형의 좌우 대칭 형태를 띨 수도 있고, 타원형 형태를 가질 수도 있다. 이 형태는 두 가우시안의 폭 σ_x, σ_y에 의해 결정되어진다. 각 DOOrG 필터는 식 6과 같이 8개의 방위를 가지도록 계산되는데, 이를 통해 전 단계의 특징 지도가 가지지 못했던 방위특징을 함유할 수 있도록 한다. 특정 방위를 가진 DOOrG 필터 $DOOrG_{x,y}(r_1, r_2, \theta, \sigma)$는 $D^1 \cdot G^1 - D^2 \cdot G^2$로 정의될 수 있으며, $G_{x,y}(\sigma_x, \sigma_y, \theta)$는 식 (7)과 같은 2-D 가우시안이고, 상수 D^1과 D^2의 값은 각각의 요소들의 계수의 합이 1로 정규화되도록 설정된다. 일반적으로 2개의 가우시안의 이심률(eccentricity)은 $r_2 = \sigma_x / \sigma_y$ 이며, ON과 OFF의 폭간 비율은 $r_1 = \sigma_{off} / \sigma_{on}$ 에 의해 정의되며 또한 고정된다. DoorG 필터에 대한 자세한 설명은 [15]와 [21]을 참조하라.

$$SF_{x,y} = \sum_{\theta}\left(\sum_{m,n} F_{m,n} \cdot |D^1 \cdot G^1 - D^2 \cdot G^2| \right)^2$$
$$TF_{x,y} = \sum_{\theta}\left(\sum_{m,n} M_{m,n} \cdot |D^1 \cdot G^1 - D^2 \cdot G^2| \right)^2 \quad (6)$$

단, $G^1 = G_{x-m,y-n}(\sigma, \gamma_1 \cdot \sigma, \theta)$,

$G^2 = G_{x-m,y-n}(\gamma_2 \cdot \sigma, \gamma_1 \cdot \gamma_2 \cdot \sigma, \theta)$

$$G_{x,y}(\sigma_x, \sigma_y, \theta) = e^{-\frac{(x\cos\theta + y\sin\theta)^2}{2\sigma_x^2}} \cdot e^{-\frac{(-x\sin\theta + y\cos\theta)^2}{2\sigma_y^2}} \quad (7)$$

III. 사공간특징 통합

기존 연구들에 의해 인간의 선택적 주의에 영향을 준다고 알려진 많은 특징들이 확인되어져 왔으나 이러한 특징들에 대해서 어떠한 특징이 다른 특징에 비해 더 중요한지, 이들 특징간의 관련성은 어떠한지에 대해서는 거의 알려진 바가 없으며, 또한 이를 알아내기도 매우 어렵다. 모션과도 같은 꽤 중요도가 높은 몇몇 특징이 있다는 것은 알 수 있으나, 이것도 다른 특징에 비해 얼마만큼 더 중요한 것인지 덜 중요한 것인지 정확히 알아낼 수가 없다. 예를 들어 어떤 특징의 경우 특정 영상에서는 다른 특징에 비해 중요도가 높은 수 있을지 모르지만 다른 영상에서는 서로 반대 입장이 될 수도 있는 것이기 때문이다[2]. 본 논문에서는 공간특징 통합과 시간특징 통합에 서로 다른 방법을 사용하였다.

1. 공간특징 통합

본 논문에서 사용하는 특징 통합 방법은 공간특징 맵을 구성하고 있는 픽셀들의 통계적 정보와 국부적인 경쟁력 특성을 이용한 아주 간단한 방법으로 그 수는 적지만 의미있는 활동성을 보이는 공간특징이 함유된 맵이 강조되고, 그렇지 않은 맵은 억제한다. 여러 공간특징 맵을 통합하여 하나의 현저도 맵을 생성하는 방법은 다음과 같다.

가장 먼저, 2.3절에서 계산된 각각의 공간특징 맵 $SF^k_{x,y}(k=1,2,3)$를 입력받아 이를 LoG 함수로 생성된 넓은 크기의 2차원 필터로 회선(convolution)한 후, 입력받은 원래의 영상과 더한다. 이러한 처리과정은 입력되는 맵을 구성하고 있는 픽셀들에 대해 좁은 범위에 있어서는 협동작용이, 넓은 범위에 있어서는 경쟁작용이 일어나게 하는 효과를 내게 되는데, 이는 특징 맵의 국부

적인 영역에서의 픽셀들과 다른 주변영역의 픽셀들을 비교하여 차이가 크면 해당 픽셀들을 활성화시키고 그렇지 않으면 억제시키는 경쟁 메커니즘이라 할 수 있다. 이러한 과정이 일정횟수 반복되고, 결과적으로 계산된 맵 $SC^k_{x,y}$는 식 (8)과 같이 계산되어 3개의 공간 현저도 맵을 만든다. 반복횟수는 실험적으로 3번 반복하였다.

$$SC^k_{x,y} = \frac{SF^k_{x,y} - MinSF}{MaxSF - MinSF}$$
$$SF^k_{x,y} = SF^k_{x,y} \times (MaxSF^k_{x,y} - AveSF^k_{x,y})^2 \quad (8)$$

여기서, $MaxSF^k_{x,y}$는 $SF^k_{x,y}$를 구성하는 특징값 가장 큰 값을, $AveSF^k_{x,y}$는 $SF^k_{x,y}$에서 $SF^k_{x,y}$를 구성하는 특징값의 최대값인 $MaxSF^k_{x,y}$을 제외한 값들의 평균값을 뜻하며, $MaxSF$와 $MinSF$는 각각 $SF^1_{x,y}$, $SF^2_{x,y}$, $SF^3_{x,y}$를 구성하는 모든 특징값 중 가장 큰 값과 가장 작은 값을 뜻한다. 활동량이 있는 모든 지점에 대해서 전체 맵 안에서 가장 큰 활동량과 평균 활동량을 비교하면 현재 지점의 활동량이 평균 활동량에 비해 얼마나 다른지 알 수 있게 된다. 이런 차이가 크면 클수록 해당 맵내의 특정 지점에서의 활동량이 다른 지점에 비해 두드러지다는 말이 되고, 그렇지 않으면 해당 맵은 별로 독특하지 활동양상을 보이는 값만을 가지고 있다는 말이 된다. 따라서 식 (8)에 의해 계산되어진 $SC^k_{x,y}$맵은 전체적으로 평이한 특징값을 가지고 있던 공간특징 맵들은 전체적으로 그 특징값이 저하되고, 그렇지 않은 맵은 예전에 비해 높은 값을 가지게 된다. 이를 통해 최종 목표 영역을 탐지하기 위해 필요없는 정보를 가지고 있는 특징맵을 걸러낼 수 있다. 이에 제안하는 시스템에서는 3개의 공간 현저도 맵 $SC^k_{x,y}$를 단순히 모두 합하여 하나의 공간 현저도 맵 $SC_{x,y}$을 만든다. [그림 7]은 본 논문에서 제안된 공간특징의 통합과정을 보여준다.

2. 시간특징 통합

시간특징 통합에 대해서는 통합에 관한 어떠한 심리학적 연구결과를 찾아볼 수 없었기에 현 상태에서는 서로 다른 3개의 모션분석 모듈로부터 얻어진 3개의 시간 특징 맵 $DORF_{x,y}$, $NF_{x,y}$, $TF_{x,y}$을 식(9)에서 보여진 바와 같이 단순히 합하여 하나의 시간 현저도 맵 $TC_{x,y}$를 생성하였다. [그림 8]은 본 시스템에서의 시간특징의 통합과정을 보여준다.

$$TC_{x,y} = \frac{1}{3}\sum_{x,y}(DORF_{x,y} + NF_{x,y} + TF_{x,y}) \quad (9)$$

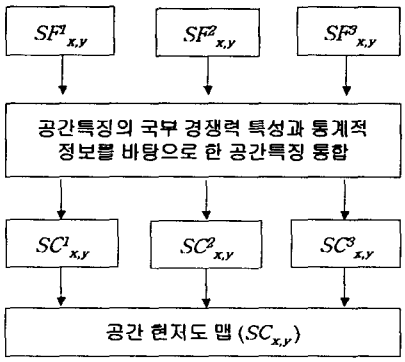


그림 7. 공간특징 통합 과정

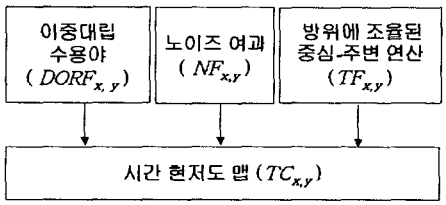


그림 8. 시간특징 통합 과정

3. 시간 현저도 맵과 공간 현저도 맵의 통합

공간 현저도 맵과 시간 현저도 맵은 [그림 9]에서 보이는 바와 같이 하나의 시·공간 현저도 맵으로 통합되어진다.

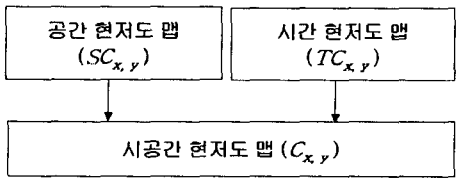


그림 9. 시공간 현저도 맵의 생성 과정

일반적으로 사람들은 공간변화보다는 시간변화에 먼저 주의를 주게 되므로 시간 현저도가 공간 현저도 보다 우선순위가 높다고도 볼 수 있겠다. 또한 생물학적 증거에 따르면 모션정보가 시각적 주의를 일으키는데 큰 역할을 한다고 알려져 있다[17].

이에 본 연구에서는 시스템에 입력된 영상에서 움직임이 감지되었을 경우 시간 현저도 맵에 가중치를 더 주어 통합하는 방법을 사용하였다. 즉, 2개의 시간 현저도 맵과 공간 현저도 맵을 통합하기 위하여 본 시스템에서는 식(10)과 같이 각 맵을 구성하고 있는 픽셀에 가중치를 두어 합하였는데 공간 특징 가중치인 α를 1, 시간 특징 가중치인 β를 2로 세팅하였다.

$$C_{x,y} = \alpha SC_{x,y} + \beta TC_{x,y} \quad (10)$$

IV. 실험결과

실험은 2가지로 나뉘어 수행하였는데, 먼저 첫 번째 실험은 시스템이 공간특징과 시간특징을 각기 제대로 사용하여 ROI를 탐지하는지의 여부를 알기 위해 심리학에서 시각적 주의 모형에 대한 실험을 할 때 일반적으로 쓰이는 실험 데이터를 기반으로 인위적으로 만든 간단한 영상([그림 10]과 [그림 11] 참고)을 사용하여 실험하였으며, 다른 한 가지 실험으로는 시스템이 시간특징과 공간특징을 모두 사용하여 제대로 ROI를 탐지하는지에 대한 여부를 알기 위해 우리 주변에서 흔히 볼 수 있는 실영상을 대상으로 실험하였다. 이때 실영상으로는 비디오 압축에서 표준으로 사용되는 영상 시퀀스(sequence)를 사용하였다.

1. 공간특징과 시간특징을 각기 사용하여 ROI 탐지

간단한 인공영상에 대한 실험을 위해 모양 및 크기([그림 10(a)], [그림 10(c)]), 색상([그림 10(b)]), 방위([그림 10(d)] 등의 한 가지 특징을 변화시킨, 다양한 종류의 'Pop-Out'에 관련된 영상을 사용하였다.

[그림 10]은 영상에 움직임이 없는 경우로 정지영상 하나하나에 대해 공간특징을 사용하여 ROI를 추출한 결과를 보여주며, [그림 11]은 움직임이 있는 영상으로 감지된 움직임에 대한 시간특징을 사용하여 ROI를 탐지한 결과를 보여준다. [그림 10(a)]~[그림 10(d)]은 테스트 영상을, [그림 10(e)]~[그림10(h)]은 제안하는 시

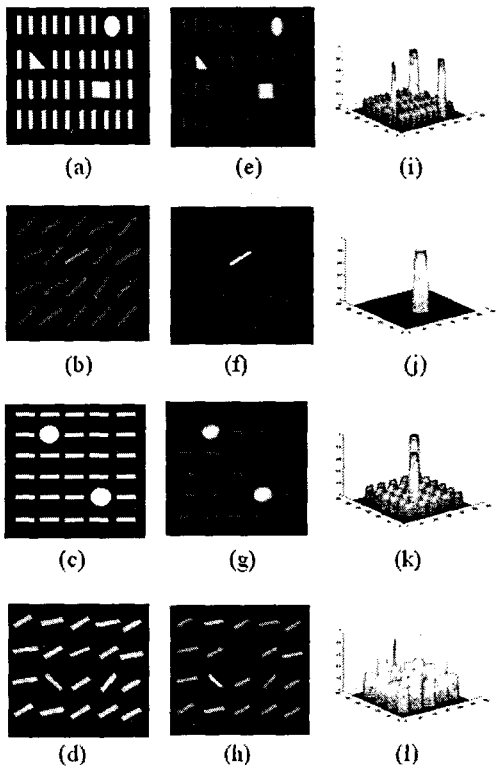


그림 10. 공간특징을 사용한 ROI 탐지

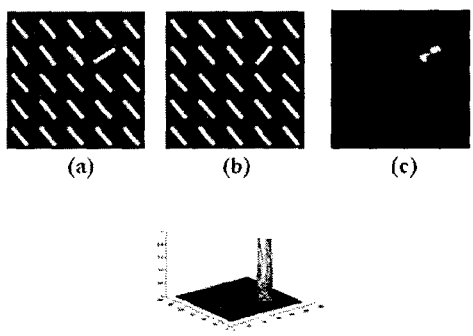


그림 11. 시간특징을 사용한 ROI 탐지

스템을 통해 얻어진 공간 현저도 맵을 나타내며, [그림 10(i)]~[그림10(l)]은 각각의 현저도 맵의 특징값 분포도를 보여준다. 또한 [그림 11(a)]~[그림11(b)]는 움직임이 있는 이전 프레임과 현재 프레임의 영상을 각각 나타내며, [그림 11(c)]는 제안하는 시스템을 통해 얻어진 시간 현저도 맵을, 그리고 [그림 11(d)]는 시간 현저도 맵의 특징값 분포도를 보여준다. [그림 10]과 [그림 11]의 결과에서도 보이는 바와 같이 제안하는 시스템은 정지영상과 동영상에 대해 각기 공간특징과 시간특징을 사용하여 입력된 영상에서 Pop-Out 되는 부분(ROI)을 제대로 탐지하고 있음을 알 수 있다.

2. 시공간특징을 모두 사용하여 ROI 탐지

실영상에 대한 실험영상으로 CIF(Common Intemediate Format) 포맷의 352X288 해상도를 가진 "클레어(Claire)", "파리(Paris)", "도로(Highway)"의 3가지 종류의 영상 시퀀스를 사용하였다. 각 영상 시퀀스는 150 프레임으로 구성되어 있고, 5초간 지속된다(1초당 30 프레임).

실영상은 윗 절에서 사용한 간단한 안공영상과 달리 시스템의 성능을 평가하기 위한 객관적인 평가기준이 없는 이유로 인해 동일한 테스트 영상에 대해 인간을 대상으로 한 심리학적 실험을 우선 수행하여 수행한 결과와 제안된 시스템을 통해 얻어진 결과와 비교 분석하였다. 인간을 대상으로 한 심리학적 실험에는 남녀40명이 참여하였다. 임의적으로 선택된 3개의 영상 시퀀스를 20~30세 사이의 20명의 남성과 20명의 여성에게 보여준 후, 제시된 영상 시퀀스에서 가장 주의가 가는 영역이나 물체에 마크 표시를 하라 지시하였다. 만일 주의가 가는 곳이 여러 곳이라면 주의가 가는 우선순위별 체크하도록 하였다.

제안하는 시스템으로부터 탐지된 ROI는 그레이 스케일 영상으로 표현된 현저도 맵으로 최종 출력된다. 최종 출력된 현저도 맵안의 명암도는 현저함 정도를 나타내는데, 이 때 더 밝은 명암도를 가진 부분일수록 더욱 현저하다는 것을 나타낸다. [표 3][표 4][표 5]는 인간을 대상으로 한 심리학적 실험결과를 나타내고, [그림 12][그림 13][그림 14]는 이에 대응되는 본 논문에서 제안하는 시스템을 통해 얻은 ROI를 보여준다.

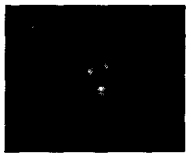
"클레어(Claire)" 영상 시퀀스의 경우, 인간을 대상으로 한 심리학적 실험에서는 입과 눈 영역이 현저한 영역으로 마크되었다(62.5%, [표 3] 참조). 만일 입과 눈 영역을 얼굴 영역으로 포함한다면 이의 비율은 70%까지 올라간다. 제안된 시스템에서는 단순히 공간특징만을 사용하였을 때에는 눈을 제외한 얼굴 영역을 ROI로 탐지하였으나([그림 12(c)] 참조), 시간특징이 같이 사용되면 입과 눈 영역, 그리고 희미하게나마 얼굴영역을 ROI로 식별한다([그림 12(e)] 참조). 이는 시간특징을 포함시킴으로써 ROI의 정탐지율이 개선되어질 수 있음을 보여준다.

표 3. "클레어" 영상 시퀀스에 대한 인간의 심리학적 실험 결과

성별	1순위	2순위
전체:40	입:15	입:12
	눈:10	헤어:10
	옷:8	눈:6
	얼굴:3	옷:3
	헤어:2	귀걸이:2
	귀걸이:2	볼터치:1
		고개:1

(a) 이전 프레임 (b) 현재 프레임

(c) 공간 현저도 맵 (d) 시간 현저도 맵



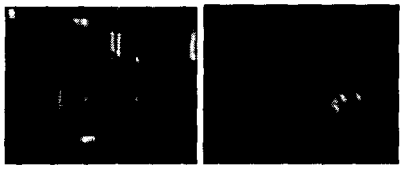
(e) 시·공간 현저도 맵

그림 12. 제안된 시스템에 입력된 "클레어" 영상 시퀀스와 이에 대한 시스템의 결과 예

표 4. "파리" 영상 시퀀스에 대한 인간의 심리학적 실험결과

성별	1순위	2순위
전체:40	공:24	펜:15
	책:4	공:11
	남자얼굴:2	여자얼굴:3
	여자얼굴:2	넥타이:2
	여자손:2	종이컵:2
	팬:2	책:2
	서류:1	서류:2
	남자헤어:1	팔찌:1
	테이블:1	인형:1
	남자입:1	

(a) 이전 프레임 (b) 현재 프레임



(c) 공간 현저도 맵 (d) 시간 현저도 맵

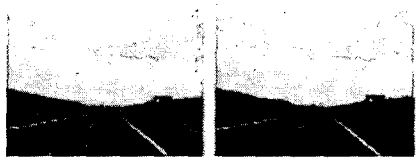
(e) 시·공간 현저도 맵

그림 13. 제안된 시스템에 입력된 "파리" 영상 시퀀스와 이에 대한 시스템의 결과 예

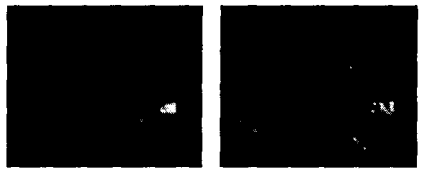
"파리(Paris)" 영상 시퀀스의 경우, 인간을 대상으로 한 심리학적 실험에서는 공이 현저한 영역으로 마크되었다(60%, [표 4] 참조). 제안된 시스템에서는 단순히 공간특징만이 사용되었을 때에는 핑크색의 컬러 노트와 황색의 문서를 가장 우선순위가 높은 ROI로 탐지하였고 공은 가장 낮은 우선순위로써 탐지하였으나([그림 13(c)] 참조), 시간특징이 함께 적용되었을 때에는 공을

표 5. "도로" 영상 시퀀스에 대한 인간의 심리학적 실험결과

성별	1순위	2순위
전체:40	도로선:21	표지판:14
	표지판:14	도로선:12
	구름:4	구름:8
	갈:1	검은점:1
		아스팔트:1
		왼쪽막대:1



(a) 이전 프레임 (b) 현재 프레임



(c) 공간 현저도 맵 (d) 시간 현저도 맵

(e) 시·공간 현저도 맵

그림 14. 제안된 시스템에 입력된 "도로" 영상 시퀀스와 이에 대한 시스템의 결과 예

1순위의 ROI로 탐지함을 볼 수 있다([그림 13(e)] 참조). 즉 시 · 공간특징을 선택적으로 적용하여 보다 일반적인 결과를 얻을 수 있음을 알 수 있다.

"도로(Highway)" 영상 시퀀스의 경우, 인간을 대상으로 한 심리학적 실험에서는 도로선과 표지판이 현저한 영역으로 마크되었다(87.5%, [표 5] 참조). 제안된 시스템에서는 공간특징만을 사용하였을 때 표지판이 가장 눈에 띄는 ROI로 탐지되고 도로선은 탐지되지 못하였으나([그림 14(c)] 참조). 시간특징이 같이 사용되었을 때에는 도로가 1순위의 ROI는 아니었지만 표지판과 함께 모두 탐지되었다([그림 14(e)] 참조).

비록 우리가 인간을 대상으로 한 심리학적 실험에 사용한 3가지 영상 시퀀스가 모두 색상이 함유된 컬러 영상이었지만, 사람들은 일반적으로 움직이는 객체 또는 영역에 보다 더 주의를 기울이고 있다. 이와 같이 모션은 시각적 주의를 일으키는데 가장 강한 영향을 주는 특징 중 하나라 할 수 있다[2][16].

V. 결 론

본 논문에서는 동영상에서 시간특징과 공간특징을 선택적으로 사용한 ROI 탐지 시스템을 제안하였다. 제안하는 시스템은 인간의 시각적 주의에 관련된 기존의 심리학적 연구결과에 기반을 둔 새로운 분석방법과 통합방법을 사용하여 ROI를 탐지해낸다.

기존의 시스템이 시각적 주의의 상향식 단서로써 공간특징을 사용하고, 시간특징은 하향식 단서로써 시스템의 피드백을 위해 사용하는 반면, 제안하는 시스템은 시간과 공간특징을 모두 상향식 단서로 사용하여 ROI를 탐지한다. 비록 제안된 시스템이 인간의 시각적 주의의 특정 부분을 모방한 것이기는 하지만, 몇 가지 이유로 인해 아직은 보완해야 할 점이 많다. 그 첫번째 이유로는 우리에게 알려진 인간의 인지 메커니즘에 대한 지식의 양이 너무나도 적다는 것이고, 또 다른 이유로는 입력된 정보를 처리하는 과정에서 추출된 특징 맵들을 하나의 현저도 맵으로 통합되는 과정이 결코 수월한 작업이 아니라는 것이다. 본 연구에서는 시간특징을 통합하여 시간 현저도 맵을 만들 때와 시간 현저도 맵과 공간 현저도 맵을 통합하여 하나의 시·공간 현저도 맵을 만들 때, 단순한 가중치를 준 합연산을 사용하였다. 통합에 있어 단순히 모든 특징 맵을 합하는 것은 신뢰성 있는 ROI를 탐지하지 못하는 경우가 많다고 이미 많은 연구들로부터 보고가 되고 있기는 하지만 이러한 연구들은 단순히 공간특징만을 통합하는 것을 목적으로 한 연구들이다.

비록 제안하는 시스템의 처리과정이 인간의 실제 시각적 주의가 처리되는 과정과 다른 면이 있다고 하더라도 제안된 시스템을 통해 얻은 실험결과는 인간을 대상으로 한 심리학적 실험을 통해 나온 결과와 대부분 일치함을 볼 수 있다. 시각적 주의에 관련된 기존의 심리학적 연구결과를 좀 더 보충하여 제안하는 시스템에 적용한다면 시스템의 성능이 개선될 수 있을 것이다. 향후 시간 현저도 맵과 공간 현저도 맵의 통합에 초점을 맞추어 더욱 다양한 영상 시퀀스를 대상으로 실험을 수행하여 제안된 모형의 성능을 평가할 것이다.

참 고 문 헌

[1] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," Vision Research, Vol.40, pp.1489-1506, 2000.

[2] W. Osberger and A. J. Maeder, "Automatic identification of Perceptually important regions in an image," Proceedings of Fourteenth Intl. Conf. On Pattern Recognition, pp.701-704, 1998.

[3] A. Clark, "Some Logical Features of Feature Integration," Inaugural lecture for the Italian Institute for Philosophical Studies, International School of Biophysics Study Program "From Neuronal Coding to Consciousness," Ischia(Naples), pp.12-17, 1998.

[4] A. Clark, "Neuronal Coding of Perceptual Systems," New Jersey: World Scientific, Series on Biophysics and Biocybernetics, Vol.9, pp.3-20, 2001.

[5] R. Anderson, L. Snyder, D. Bradley, and J. Xing, "Multimodal Representation of Space in the Posterior Parietal Cortex and Its Use in Planning Movements," Annual Review Neuroscience, Vol.20, pp.303-330, 1997.

[6] A. Treisman and G. Gelade, "A Feature -integration Theory of Attention," Cognitive Psychology, Vol.12, No.1, pp.97-136, 1980.

[7] K. Cave, "The FeatureGate Model of Visual Selection," Psychological Research, pp.182-194, 1999.

[8] N. Cepeda, K. Cave, N. Bichot and M. Kim, "Spatial Selection via Feature-driven Inhibition of Distractor Locations," Perception and Psychophysics, Vol.60, No.5, pp.727-746, 1998.

[9] M. Kim and K. Cave, "Top-down and Bottom-up Attentional Control: on the Nature of Interference from a Salient Distractor," Perception and Psychophysics, Vol.61, No.5, pp.1009-1023, 1999.

[10] A. Hanazawa, "Visual Psychophysics (2) : Neural Mechanisms of Visual Information Processing," Journal of Image Information and Television Engineers, Vol.58, No.2, pp.199-204, 2004.

[11] R. Boynton, *Human Color Vision*, New York: Holt, Rinehart and Winston, 1979.

[12] Hecht, and Eugene, *Optics*, 2nd Ed, Addison Wesley, 1987.

[13] T. Lee, T. Wachtler, and T. Sejnowski, "Color Opponency is an Efficient Representation of Spectral Properties in Natural Scenes," Vision Research, Vol.42, pp.2095-2103, 2002.

[14] R. Milanese, H. Wechsler, and S. Gil, "Integration of Bottom-up and Top-down Cues for Visual Attention Using Non-linear Relaxation," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.781-785, 1994.

[15] K. Cheoi and Y. Lee, "A Feature-driven Attention Module for an Active Vision System," Lecture Notes in Computer Science, LNCS2449, pp.583-590, 2002.

[16]A. Hiroshi, "Visual Psychophysics (8) : Visual Motion Perception and Motion Pictures," Journal of Image Information and Television Engineers, Vol.58, No.8, pp.1151-1156, 2004.

[17]http://web.psych.ualberta.ca/~iwinship/vision/visualpathways.html

[18] M. Park, S. Park, and T. Hamamoto, "A Selective Visual Attention Module based on Motion Stimuli," The Journal of the Institute of Image Information and Television Engineers, undergoing review process (Submitted for Special Issue "Human Information")

[19] M. Park, K. Cheoi, and T. Hamamoto, "A Smart Image Sensor with Attention Modules," IEEE Proc. of Computer Architecture for Machine Perception, pp.46-51, 2005.

[20] V. Navalpakkam, M. Arbib, and L. Itti, "Attention and Scene Understanding," Neurobiology of Attention, pp.197-203, 2005.

[21] 최경주, 이일병, "인간의 상향식 시각적 주의 기능을 바탕으로 한 영상의 현저한 영역 탐지", 정보과학회 논문지:소프트웨어 및 응용, 제31권, 제2호, 2004(2).

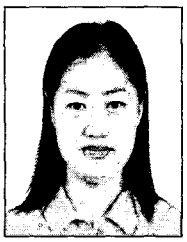
저 자 소 개

박 민 철(Min-Chul Park) 정회원

- 1993년 2월 : 홍익대학교 전자공학과(공학사)
- 1997년 3월 : 일본 동경대학 전자정보공학과(공학석사)
- 2000년 4월 : 일본 동경대학 전자정보공학과(공학박사)
- 2001년 6월~현재 : 한국과학기술연구원 시스템 연구부 선임연구원

<관심분야> : 내용기반 영상검색, 멀티미디어, 3차원 영상 디스플레이, 컴퓨터비젼

최 경 주(Kyung-Joo Cheoi) 정회원



- 1996년 2월 : 충북대학교 컴퓨터과학과 (공학사)
- 1999년 2월 : 연세대학교 컴퓨터과학과(공학석사)
- 2002년 8월 : 연세대학교 컴퓨터과학·산업시스템공학과(공학박사)
- 2002년 7월~2005년 2월 : LG CNS 연구개발센터
- 2005년 3월~현재 : 충북대학교 전기전자컴퓨터공학부 전임강사

<관심분야> : 컴퓨터비젼, 영상처리, 바이오컴퓨팅, 유비쿼터스컴퓨팅