한국군사과학기술학회지 제15권 제3호, pp. 298~305, 2012년 6월

학술논문 전자 · 통신 부문

지역적, 전역적 특징을 이용한 환경 인식

Scene Recognition Using Local and Global Features

강 산 들* Sandeul Kang
황 중 원* Joongwon Hwang
정 희 철* Heechul Jung
한 동 윤* Dongyoon Han
심 성 대** Sungdae Sim
김 준 모* Junmo Kim

Abstract

In this paper, we propose an integrated algorithm for scene recognition, which has been a challenging computer vision problem, with application to mobile robot localization. The proposed scene recognition method utilizes SIFT and visual words as local-level features and GIST as a global-level feature. As local-level and global-level features complement each other, it results in improved performance for scene recognition. This improved algorithm is of low computational complexity and robust to image distortions.

Keywords : Scene Recognition(장면 인식), Scene Classification(장면 분류), Global-local Features(전역적-지역적 특징점), Spatial Structure(공간적 구조), Gist(전역 정보 시스템 기술)

1. 서 론

영상을 기반으로 한 장면 인식[1,2], 혹은 분류는 최근 몇 년간 많은 주목을 받고 있는 분야이다. 특히, 로봇의 환경 인지 및 분석에 매우 유용하기 때문에 소형 로봇틱스(robotics)를 연구하는 연구자들은 이 분야에 주목하고 있다. 로봇의 장면 인지 및 분석 능력의 향상은, 위치추정(localization), 네비게이션(navigation), 및 매핑(mapping) 등에 사용될 수 있다. 이러한 환경 인지 및 분석을 위해서 기존에는 음파, 레이저 등과 같은 센서들을 사용하여 왔는데, 이러한 방식에는 한계점이 존재한다. 실내 환경에서는 이러한 방식이 평면의 벽, 좁은 복도나 가구 등과 같이 센서가 신호를 검출하기에 좋은 환경적 특성을 가지고 있기 때문에 상대적으로 강인한 위치 추정결과를 낼 수 있다. 그러나 실외환경에서는, 불규칙한 표면과 특징적인 물체들이 부족하고 넓은 공간적 구성으로 인하여 센서를 제대로 활용하기 힘들다고 할 수 있다. 이러한 문제를 해결하기 위해서 많은 접근법들이 존재하는데, 그 중에서 직관적이면서도 강력한 영상 기반의 방식이 가장 선호되고 널리 이용되고 있다.

장면 인식을 위한 영상을 이용한 접근방식은, 어떠한 특성에 중점을 두느냐에 따라 아래와 같이 분류될

† 2012년 4월 5일 접수~2012년 5월 25일 게재승인
* 한국과학기술원(KAIST)
** 국방과학연구소(ADD)
책임저자 : 강산들(mtplain@kaist.ac.kr)

수 있다.

가. 하위수준(low-level) 특징 기반의 환경인식

카메라로 찍힌 영상을 표현함에 있어서 하위수준 영상 기술자(image descriptor, e.g. SIFT[3], SURF[4], Color Descriptor[5])는 효과적인 방법으로 입증되어 왔다. 이러한 기술자들은 영상으로부터 각 픽셀(pixel) 단위로 주변의 픽셀들과는 구별되는 점들을 찾아내는데, 이러한 점들은 영상의 어파인(affine : 크기 변화와 평행이동, 회전 등으로 이루어진 선형기하학적변환) 왜곡이나 잡음 추가, 명암 변화 등에 강인한 특성을 갖는다. 즉, 이러한 불변적, 지역적 특성을 가진 특징점들을 이용하여 영상을 분류하고 특징지을 수 있다. 이를 이용하여 대상 영상(target image)을 기존에 확인된 영상과 비교하는 데에 사용할 수 있으며, 특히 로봇의 자율 복귀 문제에서는 로봇의 현재의 상황에 대한 인식 및 위치 파악을 위한 유용한 정보로서 활용이 가능하다고 할 수 있다.

하지만 이러한 방식은 흰색의 평평한 벽만 있는 공간이나 시야가 넓게 트인 외부환경과 같은 곳에서는 입력된 영상으로부터 특징점들을 많이 얻어 내는 것이 어려워, 이 방식만을 활용하여 주어진 영상을 특성화시키기는 어렵다는 단점이 있다. 또한 3차원 시점변화, 차폐나 배경의 무늬등과 같은 텍스쳐의 변화에 민감하다는 단점이 있다.

나. 영역(region) 기반 환경 인식

영역 기반 환경 인식 방식은 분할된 영상의 각 분할된 영역과, 그들 간의 배치 관계를 이용하여 주어진 영상을 특징짓는 방법이다. 이러한 접근법의 가장 주요한 문제 중 하나는, 각각의 분할된 영역이 영상을 구별할 만한 중요한 사물이나 영역을 훼손시키지 않으면서 분할되어야 하며, 또한 각 영역이 다른 영역과 구별될 만한 개별적 특성을 가질 필요가 있다는 것이다. 즉, 어떠한 방식으로 가장 효율적으로 주어진 영상을 분할할 것인지의 문제가 존재한다고 할 수 있다. 이 외에도 이러한 방식은 자세, 조명, 시점변화에도 영향을 잘 받는다고 할 수 있다.

다. 정황(context) 기반의 환경 인식

정황 기반의 접근법은 입력된 영상을 전체적으로 분석하고 거기에서 전역적인 시각적 특성을 추출하는 방식이다(e.g. GIST[1,2]). 이는 영상의 전체적인 인상을 정의한다고 볼 수 있다. 즉, 영상의 세부적인 특징 보다는 영상의 전반적인 특징에 주목하여, 주어진 영상이 어느 장면을 의미하는지, 예를 들어서 이 영상이 바다인지 산인지를 구별할 수 있는 정보를 제공한다. 이 방법의 장점은 영상의 분할과 각각의 개별 구역에 대한 분석을 요구하지 않는다는 점과 무작위적인 잡음, 주어진 영상의 부분적인 변화 및 조명 변화에 둔감하다는 점에 있다.

한편 이 방법은, 위치상 다른 위치에 있더라도(e.g. 해변, 숲 등) 공간적으로 비슷한 구조를 가진 영상들을 같은 카테고리 분류하기 때문에, 위치추정에 있어서는 불충분하다.

본 논문은 위에서 언급한 기존의 제한사항을 극복하기 위하여 발전되고 통합된 환경 인식 시스템을 제안한다. 이 시스템은 하위수준(low level)과 상위수준(high level)의 특징들을 합쳐서 서로의 단점을 보완하는 방법을 제시한다. 이 과정에서, 두 가지 방식으로부터 도출된 특징점들의 집합(feature sets)을 단순히 합치는 것이 아닌 각 방식이 가지는 단점을 개선시킨 후 이 둘을 결합시키고자 한다.

먼저 하위수준의 특징점을 도출하는 알고리즘에서, 지역 기반(local) 방식의 개념을 적용하는 과정에서 영상을 동일한 구간으로 나눠서 각 구역(region)에서 특징점을 추출하고, 공간적 피라미드(spatial pyramid)[7]와 유사한 방식을 적용한다. 이 알고리즘은 하위수준의 특징점을 도출하는 알고리즘에 구역 기반 알고리즘을 결합한 방식으로서, 영상의 세부적인 특징과 각 구역간의 배치 관계를 모두 이용할 수 있다는 장점이 있다. 그런 후에 GIST[2]에서 사용되는 에너지 스펙트럼 기반의 전역적인 특징점을 추출하여, 이 특징점들을 공간적 피라미드[7]로 나눠서 분석한다. 주어진 영상을 점진적으로 세밀한 영역으로 나누는 공간적 피라미드 방식은 서로 다른 두 영상으로부터 추출된 특징점들 간의 전반적인 구조적 비교가 가능하다.

본 논문은 다음과 같이 구성된다. 제 2장에서는 구현 방법에 대하여 자세히 기술하며, 그 소단락에서는 두 특징점들을 추출하는 방식의 개선 방법과 두 기술자(descriptor)를 결합하는 방법에 대하여 설명한다. 제 3장에서는 장면 위치 추정의 성능을 데이터 집단(dataset)을 통해 측정 및 분석한다. 그리고 마지막 장에서는 위에서 얻은 실험 결과를 통한 정리 및 결론

을 서술한다.

2. 이론해석 및 구현방법

가. 지역적 수준(local level)에서의 특징점 추출

주어진 영상을 이용한 위치추정(localization)에서 가장 어려운 점은, 차폐나 시점변화 등으로 인하여 테스트 영상과 키프레임(key frame, 학습된 지역적인 정보를 담은 분류된 영상)이 정확하게 같지 않다는 것이다. 이러한 문제를 해결하기 위하여, 두 영상의 중요한 부분만 비교하는 것이 유용할 수 있다. 이를 위하여 본 논문에서는, 서로 다른 두 영상을 비교할 때 각 영상으로부터 얻은 특징점만을 비교하여 두 영상이 얼마나 유사한지를 판단하도록 한다.

하지만 어떠한 방식으로 특징점을 뽑을 것인지는 생각해 볼 문제이다. 단순히 전체 영상에 대해서 세부적인 하위수준의 특징점들을 도출해 내는 것은 영상에 따라서 충분히 많은 특징점을 도출하지 못하기 때문에 상대적인 비교가 어려운 문제가 발생할 수 있다. 본 논문에서 사용되는 알고리즘[6]은, 영상을 조밀하게

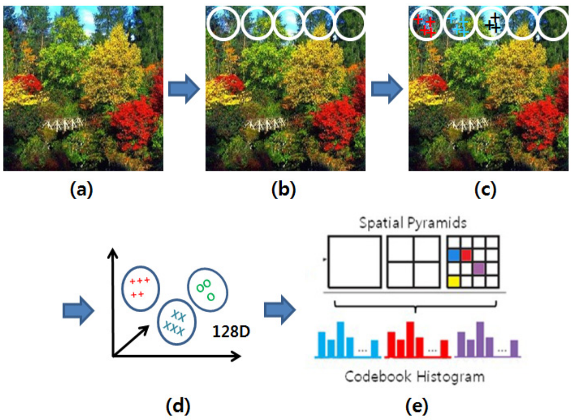

Fig. 1. Feature extraction on the local level

Above figure shows the process for getting local level features from a given image. (a) Input image, (b) Given image is segmented by same step size. (c) Get features from each part of the image by (b). Detected features are clustered by K-menas clustering algorithm and they construct visual words (d). We can obtain visual word histogram from each block. These histogram are used for comparing similarity of two images.

일정한 간격(step)으로 구역을 나눠서(예 : step = 10 pixel) 각 구역에서 특징점들을 추출하는 방식을 사용하였다. 즉, 각 구역이 결정되면 그들의 지역적 특성은 SIFT[3](혹은 SURF[4], Color Descriptor[5])를 통하여 추출된다. 만약 특징점들이 많고 상응하는 영상 기술자 또한 계산 속도가 오래걸린다면 영상 인식률을 유지하면서 step을 적절히 조절하여 그 계산 속도를 줄일 수도 있겠다.

먼저, 시각적 단어(visual words)에 대하여 히스토그램(histogram)을 얻기 위하여, 훈련 데이터(training dataset)로부터 추출한 특징점들을 k-means 클러스터링(clustering)[8]을 사용하여 분류함으로써, 시각적 단어를 생성한다. 공간적 정보를 활용하기 위해 각각의 공간 블록(block)에서 시각적 단어의 히스토그램을 생성한다. Fig. 1은 하위수준 특성의 히스토그램을 만드는 과정을 보여준다.

나. 전역적 수준(global level)에서의 특징점 추출

영상의 이산 퓨리에 변환(DFT : Discrete Fourier Transform)은 다음과 같이 정의된다.

$$\begin{aligned} I(f_x,f_y) &= \sum_{x,y=0}^{N-1} i(x,y)h(x,y)e^{-j2\pi(f_x x+f_y y)} \\ &= A(f_x,f_y)e^{j\Phi(f_x,f_y)} \end{aligned}$$

$i(x,y)$는 공간변수 (x,y)에 따른 영상의 강도(intensity)의 분포를 나타내고, f_x와 f_y는 공간의 주파수 변수이다. $h(x,y)$는 경계효과를 줄이기 위한 해닝 윈도우(hanning window) 함수이다. 이와 같은 방식은 공간좌표축(spatial domain)에서의 영상 정보를 추출하지 않고, 공간주파수좌표축(frequency domain)에서 영상의 정보를 추출하는 하나의 방식이다. 이를 통해서 얻은 결과 값의 실수부 $A(f_x,f_y)=|I(f_x,f_y)|$는 영상의 진폭 스펙트럼(spectrum)으로서, 영상의 전반적인 구조와 공간적인 주파수에 대한 정보를 제공한다. 그러나 이는 장면을 정확하게 구분하는 데에는 불충분하므로, 영상의 주된 구조와의 공간적 관계를 나타내는 정보를 가지고 있는 국지화된 에너지 스펙트럼인 스펙트로그램(spectrogram, localized energy spectrum)을 사용하였다.

스펙트로그램은 LabelMe Matlab Toolbox[9]에서 제공된 연산 알고리즘을 통해 얻을 수 있다. 이 Toolbox에 제공된 알고리즘을 따르면 스펙트로그램은 다음과

같이 계산된다. 먼저, 각기 다른 주파수와 방향성을 가진 Gabor 필터(filter)를 사용하여 필터링된 영상을 생성한다. 각 필터링된 영상은 공간적으로 여러 개의 블록(block)으로 나누어지고, 각 블록은 그 블록 내 픽셀(pixel) 값의 평균값으로 표현된다. 즉, 스펙트로그램은 결과적으로 앞서 기술한 평균값들의 벡터로 나타내진다. Fig. 2는 스펙트로그램을 만드는 과정을 보여준다.

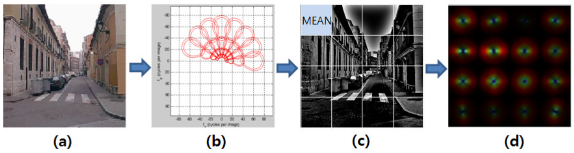

Fig. 2. Feature extraction on the global level
Given image (a) is filtered by Gabor filters (b) with different frequencies and orientations. Each filtered image is divided into several blocks and the mean value of pixels is computed from each block. A set of mean value is the spectrogram. The spectrogram is shown as above.

다. 두 개의 특성 결합

앞의 과정에서 구한 두 기술자, 즉 하위수준 특성의 히스토그램(D_1)과 상위수준 특성의 스펙트로그램(D_2)을 결합하는 과정은 다음과 같다.

먼저 이 두 기술자가 가지고 있는 값의 범위가 각각 다르기 때문에, 이들의 절대적인 수치가 아닌 각 기술자를 이루는 값들의 상대적인 수치가 의미를 가진다. 따라서, 값의 범위의 차이로 인해서 미치는 영향을 제거하기 위하여 정규화(normalization, $N(\bullet)$) 시키는 과정이 요구된다.

정규화된 기술자에는 실험을 통해서 최적화된 가중치(w_1, w_2, $w_1 + w_2 = 1$)가 각각 부여된다. 가중치가 부여된 두 기술자를 아래 (1)과 같이 서로 연결함으로써 주어진 영상을 대표하는 새로운 기술자(D_N)를 만들어 낸다.

$$D_N = [w_1 N(D_1)\ ;\ w_2 N(D_2)] \tag{1}$$

이러한 결합 방식을 통해 두 가지 방식으로 얻어낸 기술자를 최대한 활용하면서 둘의 시너지 효과를 얻을 수 있다.

라. SVM 학습과 영상의 위치 매칭(localization)

장면 위치추정을 위해 학습된 영상들(training image) 중 다른 영상들에 비해 구분되는 특성을 가진 영상들을 키 프레임으로 선택한다. 즉, 키프레임들을 학습 영상들을 구분하는 기준으로 삼는다.

이러한 키 프레임을 얻기 위해서 일정한 임계치(threshold)를 설정하여, 순차적으로 학습 영상을 입력받아(가장 처음에 입력된 영상은 기본적으로 키 프레임으로 지정된다) 현재 키프레임으로 지정된 영상과 입력된 영상의 유사도(similarity)가 임계치보다 작으면 새로운 키 프레임으로 지정하는 방식으로 키 프레임을 선정한다.(임의로 사용자에 의해 미리 구분되어져 있는 데이터 집단(dataset) 역시 사용될 수 있다).

장면 구분을 위해서 *Vedaldi*와 *Zisserman*[10]이 제안한 X^2-SVM을 사용하였다. 이 모델은 높은 정확도와 빠른 계산 속도를 보이는 모델로서, 실제로 이 방법은 거리 측정 방법(euclidean distance measure)과 거의 동일한 계산 시간을 가지면서도 더 높은 성공률을 보였다. 제안된 모델은 기존의 SVM의 커널함수(kernel function, Ψ)을 X^2으로 사용하였는데, 이는 다음과 같이 정의 된다,

$$\Psi(x) = \sqrt{x sech(\pi\omega)}\, e^{jw \log x} \tag{2}$$

where

x : kernel argument(component of feature vector)

ω : $\omega = \log\frac{y}{x}(x, y$ - kernel argument(components of feature vector)

우리는 각 키프레임을 하나의 클래스라고 정의하여, 각 클래스를 학습 영상들 중 키프레임과 유사한 몇 장의 영상들로 구성하였으며, 이렇게 선정된 학습 영상들에 대해서 우리가 새로 제안한 기술자(D_N)에 대해서 SVM 모델을 학습시켰다.

3. 실험결과

가. 실험용 데이터

캡처된 비디오 클립은 이동형 로봇이 캠퍼스 내 건물의 로비 및 주변을 돌면서 촬영한 영상으로부터 얻어진 것이다. 본 논문에서 사용한 영상데이터는 크게 실내 영상과 실외 영상, 두 분류로 구분할 수 있다.

실내 영상의 경우, 학습용 영상은 이동형 로봇을 로비를 한 쪽 방향으로 이동 함으로서, 시험용 영상은 로봇을 그 반대 방향으로 이동시켜 수집하였다. 실외 영상의 경우는, 낮과 밤에 건물 주변 잔디밭과 도로 영상을 촬영하였는데, 이는 조명 변화에 본 논문에서 제안한 알고리즘이 얼마나 강인한지를 측정함은 물론, 실내와 실외 환경에서 영상만 가지고 로봇의 위치를 추정할 때 발생할 수 있는 문제점들을 얼마나 잘 극복할 수 있는지를 확인하고자 함이었다. 실외환경에서도 학습용 및 시험용 영상은 실내 영상과 같은 방식으로 촬영하였다.

시험용 영상의 구분(classification)을 위하여 우리는 학습용 비디오 클립을 키 프레임을 기준으로 그 유사도에 따라서 몇 개의 클래스(class)로 구분하였다. 즉, 키프레임에 따라 로봇의 경로를 분획하였다. 또한 영상이 연속적으로 촬영되었기 때문에, 이로 인한 모호성을 감소시키기 위해서 모든 학습용 영상을 사용하지 않고, 각 키프레임을 기준으로 일정량의 영상만을 사용하여 SVM 모델에 적용시켰다.

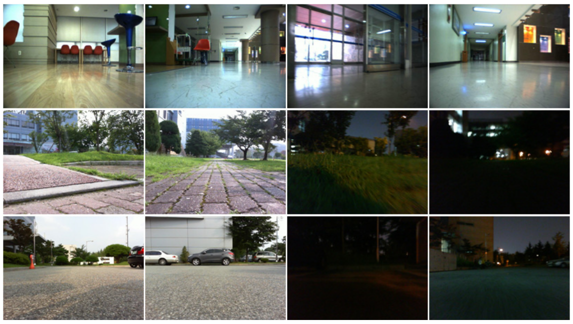

Fig. 3. Experiment image dataset

나. 매개변수(parameter) 조절 및 Preset 이용

제시된 알고리즘은 이동형 로봇의 위치 추정(혹은 환경 인식)을 위한 알고리즘으로서 정확성과 더불어 빠른 처리 속도를 요구한다. 이를 위해서, 입력된 영상(640×480)의 크기를 1/2 수준으로 줄인 영상(320×240)을 실제 시험 시에 사용하였으며, 또한 모든 영상을 이용하지 않고 전체 영상에서 10 frame 당 1 frame을 선택하여(즉, 전체 영상의 1/10만 가지고 측정함) 사용하였다. 이러한 속도 향상을 위한 제약 조건에도 불구하고 높은 정확도를 보이는 것을 확인할 수 있었다.

또한, 미리 구축해 놓은 시각적 단어(visual words), 즉, 프리셋(Preset)을 이용함으로써, 영상 학습 시에 실시간으로 시각적 단어를 생성하기 위해서 필요한 시간을 단축시킬 수 있었다. 실제로 이는 많은 CPU 점유율과 수행시간을 필요로 하는 작업이므로, 로봇이 주행하면서 구축하기에는 무리가 있기 때문이다. 우리는 인터넷에서 무작위로 얻은 실내 및 실외 영상 3000장을 이용하여 하위수준 영상 기술자를 위한 총 300개의 단어를 만들어, 이를 히스토그램을 구성할 때 사용하였다.

Fig. 4. Images for constructing visual words

다. 실내 영상 테스트

위에서 설명한 방법으로 얻은 실내 영상을 가지고 실험을 진행하였다. 우리가 수집한 학습용과 시험용 영상의 개수는 각각 5897개와 5085개이었다.

1) 지역적(local) 특성만을 이용했을 때와의 비교

Fig. 5는 하위 계층 특성, 즉 지역적 특성만을 이용하였을 경우의 실험 결과를 보여준다.(x축 : 테스트 영상, y축 : 키프레임(구분된 클래스 의미)) 유클리드 거리(euclidean distance) 차이를 이용한 구분(청색)은 79.97 % 정확도를 보인다. 이와 달리, X^2-SVM을 이용하여 분류했을 경우에는(녹색) 94.10 %의 정확도를 보이는 것을 확인할 수 있다. 이를 통해 X^2-SVM을 통

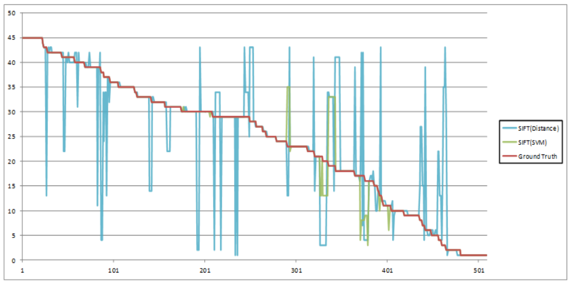

Fig. 5. Experiment results using local level features without using global level features

해서 약 14 % 정도의 성능이 향상 되는 것을 확인할 수 있었다. 그림에서 적색은 실제 매칭되어야 하는 키 프레임을 나타낸 것이다.

2) 전역적(global) 특성만을 이용했을 때와의 비교

Fig. 6은 상위수준 특성, 즉 전역적 특성만을 이용하였을 경우의 결과이다.(x축 : 테스트 영상, y축 : 키 프레임(구분된 클래스 의미)) 유클리드 거리(euclidean distance) 차이를 이용한 구분(청색)은 87.03 % 정확도를 보인다. 이와 달리, X^2-SVM을 이용하여 분류했을 경우에는(녹색) 88.98 %의 정확도를 보이는 것을 확인할 수 있다. 이를 통해 X^2-SVM을 통해서 약 2 % 정도의 성능이 향상 되는 것을 확인할 수 있었다. 위 두 실험을 통해서 단순히 유클리드 거리 차이를 구하는 것보다 X^2-SVM에 의존하여 입력된 테스트 영상을 분류하는 것이 훨씬 더 높은 정확도를 보이는 것을 알 수 있다.

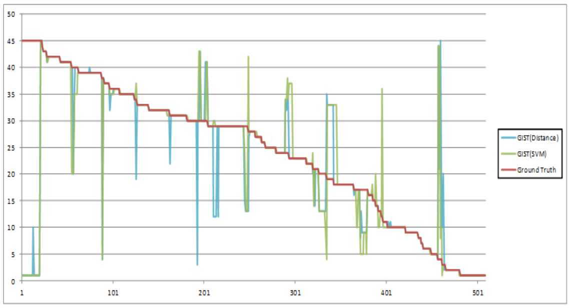

Fig. 6. Experiment results using global level features without using local level features

3) 제안된 알고리즘

Fig. 7은 본 논문에서 제안한 알고리즘을 사용하였을 경우의 결과다.(x축 : 테스트 영상, y축 : 키프레임(구분된 클래스 의미)) 어떠한 부분은 여전히 좋지 않지만, 하위수준 혹은 상위수준 중 한 가지의 특성만을 사용했을 경우와 비교하였을 경우, 확연히 향상된 결과(95.8 %)를 보임을 알 수 있다.

위 결과로 알 수 있듯이. SVM 구분을 통하여 유클리드 거리를 이용한 구분보다 더 높은 성공률을 얻을 수 있다. 또한 하위수준 특성은 전역 계층 특성을 이용한 경우보다 더 높은 정확도를 보인다. 비록 한 가지 계층(하위수준)의 특성이 다른 계층(전역 계층)의 특성보다 더 나은 정확도를 보이지만, 이 둘을 융합하여 사용함으로서 결과를 확연하게 향상시킬 수 있었다.(Table 1)

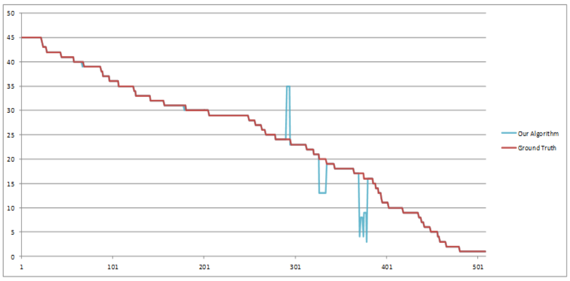

Fig. 7. Experiment result using our suggested algorithm

Table 1. Experiment results of indoor test

알고리즘		정확도(Ground Truth 대비)
지역적(local)	Euclidean	79.97 %
	SVM	94.10 %
전역적(global)	Euclidean	87.03 %
	SVM	88.98 %
제안된 알고리즘		95.8 %

라. 실외 영상 테스트

낮과 밤 시간에 건물 주변의 도로와 잔디밭에서 이미 설명한 바와 같은 방식으로 촬영하여 테스트 영상 데이터를 얻었다. 실외 영상에 대해서는 우리가 제안한 알고리즘이 실외에서 얼마나 높은 성능을 나타내는지를 실험해 보았으며, 결과는 아래 Table 2와 같다.

Table 2. Experiment results of outdoor test

장소	시간	정확도(Ground Truth 대비)
도로	낮	99.29 %
도로	밤	98.62 %
잔디밭	낮	99.89 %
잔디밭	밤	95.24 %

각 실험에 대한 ground truth 대비 테스트 결과에 대한 그래프는 아래 Fig. 8 ~ Fig. 11에서 보이는 바와 같다.(x축 : 테스트 영상, y축 : 키프레임(구분된 클래스 의미))

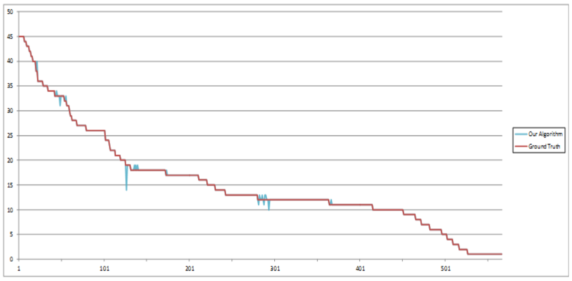

Fig. 8. Experiment results of outdoor road test at daytime

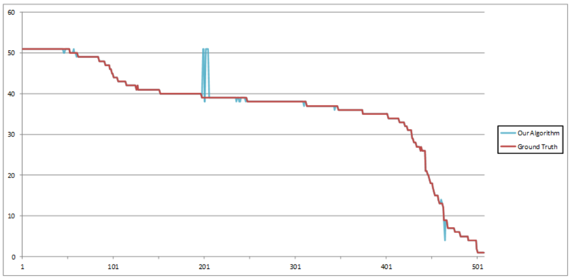

Fig. 9. Experiment results of outdoor road test at nighttime

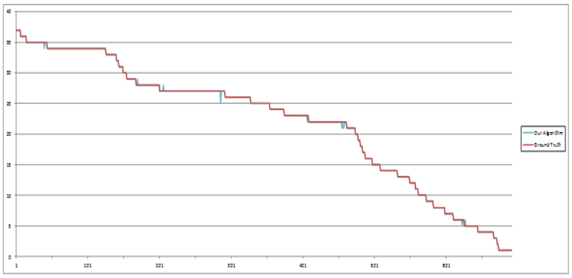

Fig. 10. Experiment results of outdoor lawn test at daytime

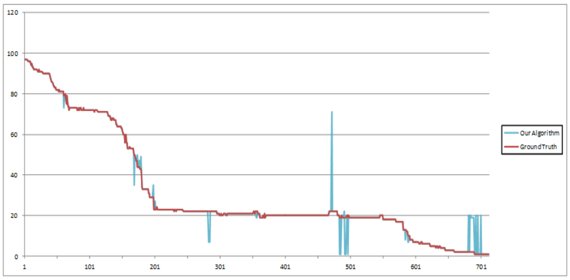

Fig. 11. Experiment results of outdoor road test at nighttime

실외 영상에서는 실내 영상 데이터보다 훨씬 높은 정확도를 보였다. 하지만 밤 영상에서는 낮에 비해서 낮은 정확도를 나타내었는데, 이는 밤에 보이는 조명에 의한 블러링(blurring)현상이 심하거나 혹은 조명이 거의 없어서 입력되는 영상이 눈으로도 구별하기 어려운 영역에서 발생하였다. 하지만 이러한 영역을 제외하고는 전반적으로 높은 정확도를 보이는 것을 확인할 수 있었다.

4. 결 론

본 논문에서는 영상을 기반으로 하는 위치 추정 또는 환경인식에서 지역적 특징과 전역적 특징을 결합하여 매칭하는 새로운 모델을 제안하였다. 이 모델은 최적화된 하위수준과 상위수준 특성들을 합침으로서, 하위수준이나 상위수준 개별보다 더 나은 성능을 보일 수 있었다. 또한, 지표(landmark)와 같은 주요 특징점 및 각 영역간의 배치관계에 의존하기 때문에 차폐나 밀도변화, 시점변화 등에 더욱 강인한 성능을 보인다. 본 논문에서 제안한 모델을 이용하여 영상의 매칭을 이용한 위치 추정 및 환경인식 결과를 개선 할 수 있으며, 로봇의 위치추정 기술 등에 활용할 수 있다. 끝으로 학습 시에 모든 영상들이 대한 지역적 특징을 나타내는 특징점들에 대한 영상 기술자를 추출할 때 계산 소요시간이 매칭 속도에 비해 클 수 있기 때문에 높은 인식률을 유지하면서 낮은 영상 기술자 계산 속도를 얻어내는 효과적인 방법을 연구하는 것이 향후 연구 과제로 제시될 수 있다.

후　　기

본 논문은 국방과학연구소의 “휴대용 소형로봇 자율복귀 기술 개발” 사업의 일환으로 수행되었음.

References

[1] Oliva, A. and Torralba, A., “Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope”, International Journal of Computer Vision 42 (3), pp. 145～175, 2001.

[2] C. Siagian and L. Itti, “Rapid Biologically-inspired Scene Classification using Features Shared with Visual Attention”, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2), pp. 300～312, 2007.

[3] Lowe, D. G., “Distinctive Image Features from Scale-invariant Keypoints”, International Journal of Computer Vision 60 (2), pp. 91～110, 2004.

[4] Bay, H., Ess, A. and Tuytelaars, T., Van Gool, L., “Speeded-Up Robust Features(SURF)”, Computer Vision and Image Understanding 110 (3), pp. 346～359. 2006.

[5] Van De Sande, K., Gevers and T., Snoek, C., “Evaluating Color Descriptors for Object and Scene Recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9), Art. No. 5204091, pp. 1582～1596, Sept. 2010.

[6] Liu, Ce and Yuen, Jenny and Torralba, Antonio and Sivic, Josef and Freeman, William T., “SIFT Flow : Dense Correspondence across Different Scenes”, ECCV, 2008.

[7] Lazebnik, S., Schmid, C. and Ponce, J., “Beyond Bags of Features : Spatial Pyramid Matching for Recognizing Natural Scene Categories”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2, Art. No. 1641019, pp. 2169～2178, pp. 1～10. 2006.

[8] S. Lloyd, “Least Squares Quantization in PCM”, IEEE Transactions on Information Theory, Vol. 28, pp. 129～137, 1982.

[9] B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, “LabelMe : A Database and Web-based Tool for Image Annotation”, International Journal of Computer Vision, pages 157～173, Volume 77, Numbers 1-3, May, 2008.

[10] A. Vedaldi, A. Zisserman, “Efficient Additive Kernels via Explicit Feature Maps”, IEEE Conference on Computer Vision and Pattern Recognition, 2010.