JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 20, No. 11, pp. 2141-2145, Nov. 2019

자립형 모바일 증강현실을 위한 효율적인 특정 물체 인식

이 수 원
경상대학교 컴퓨터과학과, 경상대학교 기초과학연구소

Efficient Specific Object Recognition for Standalone Mobile Augmented Reality

Suwon Lee
Department of Computer Science and The Research Institute of Natural Science, Gyeongsang National University

[요 약]

본 논문은 자립형 증강현실 시스템 구현을 위한 효율적인 특정 물체 인식 알고리즘을 제안한다. 이를 위해 순수 모바일 기기의 자원만을 활용하여 특정 물체를 정확하게 인식하는 기법을 제안한다. 특히 물체의 수가 1,000개에 가까울 때 인식 단계에서 발생하는 정확도 문제와 메모리 및 속도 문제에 대한 해결책을 제시한다. 실험을 통해 단일 물체를 대상으로 한 기존의 자립형 모바일 증강현실 시스템과 비교했을 때 8%의 인식률 저하로 인식 대상 물체의 수를 1,000개로 확장할 수 있었다. 또한 모바일 기기 상에서 18Mbytes의 메모리만으로 1,000개의 물체를 대상으로 실시간 연산이 가능함을 보인다.

[Abstract]

This paper proposes an efficient specific object recognition algorithm for standalone mobile augmented reality (AR) system. For this purpose, we propose a technique to accurately recognize specific objects using only the resources of pure mobile devices. Specifically, we propose a solution to the accuracy, memory, and speed problems that occur in the recognition process when the number of objects is close to 1,000. Compared to existing mobile AR systems that recognizes only one object, we could increase the number of recognition objects to 1,000 at the expense of only 8% reduction in the recognition rate. Further, the real-time recognition of 1,000 objects uses only 18 Mbytes of memory on mobile devices.

색인어 : 자립형 모바일 증강현실, 모바일 증강현실, 증강현실, 실시간 물체 인식, 특정 물체 인식

Key word : Augmented Reality, Mobile Augmented Reality, Real-time Object Recognition, Specific Object Recognition, Standalone Mobile Augmented Reality

http://dx.doi.org/10.9728/dcs.2019.20.11.2141

Received 26 September 2019; **Revised** 04 November 2019
Accepted 25 November 2019

***Corresponding Author; Suwon Lee**

Tel: +82-55-772-1394
E-mail: leesuwon@gnu.ac.kr

Ⅰ. 서 론

증강현실(augmented reality, AR)이란 사용자가 보고 있는 현실 세계에 가상의 콘텐츠를 정교하게 정합하여 사용자로 하여금 가상의 콘텐츠가 마치 현실 세계에 존재하는 것 같은 착각을 불러일으키게 하는 기술을 말한다. 가상의 콘텐츠는 증강현실의 사용 목적에 따라 이미지, 동영상, 3차원 물체 등으로 다양화되며, 현실 세계에 존재하는 매개체에 증강되어 사용자에게 새로운 경험과 정보를 제공한다.

최근 모바일 증강현실 시장은 스마트폰 이용자 증가와 더불어 빠르게 성장하고 있다[1]. 모바일 기기의 대중화는 증강현실에 대한 기대감을 증가시키는 동시에 제한된 자원 문제를 해결해야 하는 숙제도 안겼다. 제한된 자원은 증강현실의 대상이 되는 물체의 수를 한정하며, 이는 서비스의 단순화로 이어진다. 증강현실의 대상이 되는 물체의 수가 많은 상황에서 모바일 증강현실을 구현하기 위한 방법으로 서버-클라이언트 기반의 시스템이 제안되어 왔다[2,3]. 서버-클라이언트 기반의 시스템은 대부분의 연산과 메모리 활용을 서버에서 수행한 후, 그 결과를 모바일 기기에 전송한다. 그러나 이러한 방식은 네트워크의 연결 상태가 서비스의 품질을 좌우하며, 증강현실의 중요한 요소 중 하나인 몰입감(immersion)을 저해한다. 높은 몰입감을 위해서는 사용자의 시점이 물체를 향했을 때, 그 즉시 서비스가 이루어져야 하는데, 이는 매 프레임마다의 실시간 처리를 요하며 매 프레임을 서버로 전송하여 처리하기에는 현실적인 어려움이 따른다. 대게는 서비스 도중에 화면을 터치하는 등의 사용자의 협조를 요구하며, 이는 몰입감의 저하로 이어진다. 증강현실 서비스의 다양화를 위해서는 다수의 특정 물체를 대상으로 모바일 기기에서 모든 처리가 이루어지는 시스템의 개발이 필요한 실정이다.

본 논문은 다수의 특정 물체를 대상으로 한 자립형 모바일 증강현실을 구현하기 위한 물체 인식 알고리즘을 제안한다. 이는 기본 특징을 설계하고, 이를 학습 및 인식에서 활용하는 과정을 포함한다. 실험을 통해 모바일 기기에서 18MBytes의 메모리만을 사용하여 1,000개의 특정 물체의 실시간 인식이 가능함을 보인다.

본 논문은 다음과 같이 구성된다. 2장에서는 기본 이론이 되는 특정 물체 인식을 위한 지역 특징(local feature) 추출 과정에 대해 살펴보고, 단일 물체 대상의 자립형 모바일 증강현실에 대해 소개한다. 3장에서는 자립형 모바일 증강현실 서비스를 다수의 물체로 확장하기 위해 제안한 알고리즘에 대해 설명한다. 여기에서는 기본 특징을 설계하고, 이를 활용하여 학습 및 인식하는 과정을 포함한다. 4장에서는 제안한 기법에 대한 실험 결과를 제시하고, 마지막으로 5장에서 결론을 맺고 향후 연구 방향을 모색한다.

Ⅱ. 이 론

2-1 특정 물체 인식을 위한 지역 특징 추출

특정 물체 인식을 위한 지역 특징 추출은 키포인트(keypoint) 검출과 기술(description) 과정을 통해 이루어진다. 대표적인 방법으로는 SIFT[4]가 있다. SIFT는 두 가우시안 이미지의 차(Difference of Gaussian, DoG)를 이용해 국소적 극 값(local extrema)을 선별하여 크기 변화에 반복성(repeatability)이 높은 키포인트를 검출한 후, 키포인트를 둘러싼 국소 지역의 경사도(gradient) 이미지를 히스토그램(histogram)으로 표현하여 어파인(affine) 변화에 강인한 고차원의 벡터(high-dimensional vector)를 생성한다. 하지만 SIFT는 검출과 기술과정에서 모두 많은 연산 량을 요구하기 때문에 모바일 환경에서 활용하기 위해서는 속도 최적화가 필요하다.

2-2 단일 물체 대상 자립형 모바일 증강현실

본 논문에서는 서버나 다른 기기의 도움 없이 모바일 기기에서 모든 처리가 이루어지는 증강현실을 자립형(standalone) 모바일 증강현실이라 부른다. 최초의 자립형 모바일 증강현실 시스템은 Wagner[5]에 의해 제안되었다. Wagner는 SIFT를 모바일 환경에 알맞도록 설계한 PhonySIFT를 제안하였다. PhonySIFT는 DoG 대신에 FAST[6]를 사용하여 주변 픽셀들의 단순 밝기 값의 비교만으로 빠르게 키포인트를 검출한다. 기존 FAST에 크기 속성을 부여하기 위해 이미지 피라미드(image pyramid)의 각 이미지마다 독립적으로 키포인트를 검출하여 크기 공간(scale space)을 근사하였다. 기술과정에서는 SIFT의 차원(dimension)을 기존 128차원에서 36차원으로 줄여 모바일 기기에서 실시간 연산이 가능하도록 하였다.

Ⅲ. 제안한 방법

3-1 기본특징 설계

PhonySIFT는 단일 물체를 대상으로 설계되었으며, 이 또한 다수의 물체를 대상으로 할 때는 속도 및 메모리 문제가 발생한다. 본 연구에서는 다수의 물체를 대상으로 하기 위해 PhonySIFT의 속도 및 메모리 효율을 극대화한 양자화된(quantized) PhonySIFT(QPhonySIFT)를 제안하고 이를 기본 특징으로 사용한다. 기존의 PhonySIFT는 특징 벡터를 0과 1사이로 정규화한 4바이트(bytes)의 실수 값으로 표현한다. QPhonySIFT는 이를 0과 255 사이로 선형 변환하여 양자화한 1바이트의 정수 값으로 표현하여 메모리 효율을 극대화한다. QPhonySIFT는 특징 벡터 값이 256개로 유한하기 때문에 추후 인식 단계에서 특징들 사이의 매칭(matching)을 수행할 때 고속으로 특징들 간의 거리(distance)를 계산할 수 있다는 추가 장점

을 가진다. 아래는 D차원의 두 QPhonySIFT 특징 벡터 f^1과 f^2의 squared L2 거리의 고속 계산법이다.

$$dist(\mathrm{f}^1, \mathrm{f}^2) = \sum_{i=1}^{D} (f_i^1 - f_i^2)^2 = \sum_{i=1}^{D} T(f_i^1, f_i^2) \quad (1)$$

수식(1)의 $T(f_i^1, f_i^2)$는 $(f_i^1 - f_i^2)^2$의 값을 미리 계산하여 저장해 놓은 값이다. QPhonySIFT의 특징이 사용하는 값이 256개로 유한하기 때문에 모든 조합인 65,536개의 $(f_i^1 - f_i^2)^2$의 값을 미리 계산해놓는 것이 가능하다. 본 연구에서는 QPhonySIFT에 의해 추출된 $D = 36$인 특징을 기본 특징으로 사용하여 메모리와 속도의 효율성을 극대화한다.

3-2 학습 단계

학습할 물체의 이미지로부터 QPhonySIFT 특징을 추출한다. 추출된 특징은 N_p개의 키포인트 집합과 이에 대응되는 N_f개의 특징 벡터 집합을 포함한다. 본 연구에서는 실험적으로 선택된 $N_\mathrm{p} = N_\mathrm{f} = 500$을 이용해 특징을 추출하였다.

추출된 총 특징의 수는 학습 물체의 수에 비례하며, 인식 단계에서 특징들 사이의 매칭 속도를 좌우한다. 대규모의 물체를 대상으로 실시간 매칭이 가능하도록 고속 근사 최근 이웃 검색기(fast approximate nearest neighbor searcher, FANNS)를 학습한다. FANNS는 검색 공간을 축소하여 높은 확률로 최근 이웃을 매우 빠른 속도로 찾아 주는 역할을 한다. 검색 공간을 축소하는 대표적인 방법으로는 해싱(hashing)과 나무(tree) 구조가 있다. 본 연구에서는 k차원 나무(k-dimensional tree, k-d tree)를 이용해 공간을 축소하고, 여러 개의 서로 다른 k-d tree가 상호 보완하여 검색의 정확도를 높일 수 있도록 무작위(randomized) kd-tree[7]를 구축해 FANNS를 학습하였다.

3-3 인식 단계

질의(query) 이미지로부터 QPhonySIFT 특징을 추출한다. 추출된 특징을 FANNS를 이용해 근사 매칭을 수행한다. 이 과정에서 각각의 특징 f^q에 대해 근사 최근 이웃 특징 $\widehat{NN}(\mathrm{f}^q)$이 매칭되고, 둘 사이의 거리가 수식 (1)에 의해 고속으로 계산된다. 각각의 특징 f^q에 대해 매칭된 특징 $\widehat{NN}(\mathrm{f}^q)$의 출처를 이용해 f^q가 어떤 학습 물체의 이미지를 지지하는지 정도를 아래와 같이 계산한다.

$$\mathrm{Score}_{\mathrm{f}^q}(I_i) = \begin{cases} \exp - \dfrac{dist(\mathrm{f}^q, \widehat{NN}(\mathrm{f}^q))^2}{2\sigma^2} & \text{if } \widehat{NN}(\mathrm{f}^q) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

수식 (2)는 특징 f^q가 매칭된 특징 $\widehat{NN}(\mathrm{f}^q)$의 출처인 이미지 I_i를 지지하는 점수이며, 이때 매칭된 두 특징 사이의 거리가 활용된다. 모든 특징에 대해 수식 (2)를 계산하고, 가장 높은 점수를 받은 이미지를 인식 대상 후보로 추정한다. 질의 이미지의 특징과 인식 대상 후보 이미지의 특징 사이의 매칭을 최종적으로 수행하고, 매칭의 결과로 PROSAC[8]을 이용해 인식 대상을 검증한다. 이 과정에서 인식의 성공 및 실패 여부가 결정되며 인식이 성공한다면 추가적으로 초기 자세가 계산된다.

Ⅳ. 실험

4-1 특징 비교

첫 번째 실험으로 특징 자체의 고유한 성능 비교를 위해 Vienna 데이터베이스[5]를 활용하였다. Vienna 데이터베이스는 한 장의 위성사진을 출력해서 촬영한 연속적인 프레임들이 테스트 세트로 활용된다. 테스트 세트는 총 다섯 세트로 구성되는데, 각각의 세트는 길이가 서로 다르며, 서로 다른 외부 환경 변화를 포함하고 있다. 실험에는 통계적 분류기의 대표인 Random Forest[9], 2진(binary) 특징의 대표인 ORB[10], 모바일 기기에 최적화된 특징의 대표인 PhonySIFT[5], 본 연구에서 제안한 특징인 QPhonySIFT를 비교 대상에 포함하였다.

표 1은 테스트 세트별 인식률을 보여준다. Random Forest는 전체적으로 고른 성능을 보였으며, ORB는 2진 특징의 특성으로 알려진 대로 흐려짐에 강인하지만 시점 변화와 가려짐에 취약점을 보였다. 반면에 PhonySIFT와 QPhonySIFT는 흐려짐에는 약하지만 시점 변화와 가려짐에 좋은 성능을 보였다. 이는 본 연구의 기본 특징을 2진 특징이 아닌 PhonySIFT에 기초하여 설계한 이유이다. 증강현실의 시나리오에 있어서 영상의 흐려짐은 카메라의 움직임이 빠를 때인 추적 단계에서 더 빈번히 발생하고, 초기 인식 단계에서는 가려짐과 시점 변화가 더 빈번히 발생하기 때문이다. QPhonySIFT는 PhonySIFT에 비해 전체적으로 약 5% 떨어진 인식률을 보였다. 이는 양자화 과정에서 수반되는 정보의 손실로 인한 결과이며, 메모리 효율 및 속도의 최적화를 위해 손실된 정확도이다.

표 1. 특징들의 정확도 비교
Table 1. Accuracy comparison of the features

Test Sets	Features			
	Random Forest	ORB	PhonySIFT	QPhonySIFT
Simple	0.80	0.93	0.96	0.94
Occlusion	0.77	0.53	0.82	0.75
Tilt	0.45	0.37	0.46	0.44
Fast Movement	0.60	0.74	0.51	0.42
Loss of Target	0.55	0.54	0.53	0.52
Total	0.64	0.61	0.66	0.61

표 2. 특징들의 연산 시간과 메모리 사용량 비교
Table 2. Computation time and memory usage comparison of the features

Features	Computation time for each operation (ms)					Memory (KB)
	Detection	Description	Matching	PROSAC	Total	
Random Forest	3.9	-	8.4	2.2	14.5	10,240
ORB	3.9	4.2	3.1	2.2	13.4	16
PhonySIFT	3.9	5.0	4.2	2.2	15.3	72
QPhonySIFT	3.9	5.0	1.9	2.2	13.0	18

표 2는 구간별 연산속도 및 메모리 사용량을 보여준다. 키포인트 검출과 PROSAC은 모든 특징이 같은 방법을 공유하고 있으며 Random Forest의 경우 알고리즘의 특성상 키포인트 기술 과정이 생략된다. 전체 속도에서 보듯이 모든 특징이 실시간성을 보장하고 있지만 그 중에서도 ORB와 QPhonySIFT가 가장 좋은 성능을 보였다. QPhonySIFT의 경우 PhonySIFT에 비해 정확도 손실이 있었지만 매칭 과정에서 오는 속도 이득으로 인해 PhonySIFT보다 약 2.3ms 빠른 성능을 보였다. Random Forest는 하나의 이미지를 학습하기 위해 10Mbytes 이상의 메모리를 필요로 하여 메모리 효율성이 매우 좋지 않다. 그 외 세 가지 특징은 메모리 사용량 측면에서 모두 우수하지만 이 또한 ORB와 QPhonySIFT가 월등하다. QPhonySIFT는 PhonySIFT에 비해 4배의 메모리 효율을 가진다. 종합적인 효율성 측면에서는 ORB와 QPhonySIFT가 가장 좋은 선택이라고 볼 수 있다. 모든 속도 테스트에는 삼성 갤럭시 S9을 사용하였다.

4-2 특정 물체 인식 성능 실험

두 번째 실험으로 제안한 방법의 특정 물체 인식 성능 실험을 수행하였다. Vienna 데이터베이스는 인식하고자 하는 물체가 하나일 때만을 대상으로 하였기 때문에 다수의 물체를 고려하기 위해 켄터키 대학에서 배포한 UKBench 데이터베이스[11]를 사용하였다. UKBench 데이터베이스는 2,550개의 물체를 서로 다른 4개의 시점에서 촬영하여 총 10,200장의 이미지로 구성되어 있다. 본 연구에서는 목표 이미지인 Vienna 위성사진을 인식하는 데 있어서 오답(distractor) 역할을 하는 다수의 이미지가 필요하기 때문에 하나의 물체 당 하나의 이미지가 바람직하다. 하나의 물체를 다른 시점으로 촬영한 여러 개의 이미지가 포함되면 물체의 중복이 발생한 것이기 때문이다. 따라서 물체의 중복 없이 하나의 물체 당 하나의 이미지를 이용하여 총 2,550개의 오답을 구성하였다. 학습 단계에서 포함되는 오답의 수가 인식 단계에서 인식하고자 하는 물체의 수를 의미하며, 오답 사이에서 정답을 얼마나 잘 찾느냐로 성능을 평가하였다.

표 3. 인식 대상 물체의 수 증가에 따른 제안한 방법의 특정 물체 인식 성능
Table 3. Specific object recognition performance of the proposed method according to the number of distractors

	Number of distractors				
	50	100	200	500	1,000
Accuracy	0.60	0.59	0.59	0.59	0.58
Time (ms)	23.0	24.4	30.5	33.3	37.5
Memory (KB)	0.9	1.8	3.6	9	18

표 3은 제안한 방법이 특정 물체를 얼마나 정확하고 효율적으로 인식하는 지에 대한 실험 결과를 보여준다. 오답의 수, 즉 인식 대상 물체의 수가 증가할수록 정확도의 저하를 보이지만 하나의 물체를 대상으로 한 기존 방법과 비교하였을 때는 9%의 정확도 손실만으로 인식 대상 물체의 수를 1,000개의 물체로 확장할 수 있었다. 효율성 측면에서는 모바일 기기에서 18MBytes의 메모리만을 사용하여 1,000개의 특정 물체의 실시간 인식(초당 26.7 프레임 처리)이 가능함을 확인하였다. 해당 실험을 통해 제안한 방법은 1,000개의 특정 물체 인식을 대상으로 하는 자립형 모바일 증강현실 서비스를 구현하기 위한 특정 물체 인식 방법으로 활용이 가능함을 증명하였다.

Ⅴ. 결 론

본 연구에서는 모바일 증강현실 시스템에서 인식 대상 물체의 수가 많아질 때 발생하는 모바일 기기의 자원 제약 문제를 해결하기 위한 방법을 제안하였다. 특징 설계에서부터 학습 및 인식 단계에 이르기까지 정확도를 최대한 유지하면서 속도 및 메모리의 효율성을 극대화하는 새로운 방법들을 제시하였다.

제안한 물체 인식 알고리즘은 다수의 특정 물체를 대상으로 하는 자립형 모바일 증강현실 서비스에 대한 솔루션을 제공한다고 볼 수 있다. 이는 증강현실 응용에 있어서 시나리오의 다양화에 기여할 뿐만 아니라 다수의 특정 물체를 제한된 자원만으로 인식 및 자세 계산까지 해야 하는 컴퓨터 비전의 응용 등에도 간접적으로 활용될 것으로 기대된다.

1,000개의 물체를 대상으로 서비스가 가능한 수준의 성능을 보였지만 규모 불변의 문제를 완전히 해결한 것은 아니다. 기본 특징을 더 정교하게 설계하거나 FANNS의 효율성을 극대화하는 연구 등의 연구를 추가로 진행한다면 10,000개 단위의 물체로의 확장도 가능할 것으로 기대된다.

감사의 글

본 연구는 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2018R1C1B5046098)과 문화체육관광부 및 한국콘텐츠진흥원의 2019년도 문화기술 연구개발 지원사업[No. 1375026932, 사용자 참여형 문화공간 콘텐츠를 위한 AR 플랫폼 기술개발]으로 수행되었음.

참고문헌

[1] J. Hong, M. R. Yu, and B. Choi, “An Analysis of Mobile Augmented Reality App Reviews Using Topic Modeling,” *Journal of Digital Contents Society*, Vol. 20, No. 7, pp. 69-76, Feb 2014.

[2] S. Gammeter, A. Gassmann, and L. Bossard, “Server-Side Object Recognition and Client-Side Object Tracking for Mobile Augmented Reality,” in *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, San Francisco, CA, pp. 1-8, 2010.

[3] J. Jung, J. Ha, S.W. Lee, F.A. Rojas, and H.S. Yang, “Efficient Mobile AR Technology Using Scalable Recognition and Tracking Based on Server-Client Model,” *Computer & Grraphics*, Vol. 36, No. 3, pp. 131-139, May 2012.

[4] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91-110, November 2004.

[5] D. Wagner, A. Mulloni, and D. Schmalstieg, “Real-Time Detection and Tracking for Augmented Reality on Mobile Phones,” *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 3, pp. 355-368, August 2009.

[6] E. Rosten and T. Drummond, “Machine Learning for High-Speed Corner Detection,” in *Proceeding of European Conference on Computer Vision*, Graz, Austria, pp. 430-443, 2006.

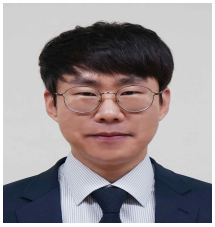
[7] C. Silpa-Anan and R. Hartley, “Optimised Kd-Trees for Fast Image Descriptor Matching,” in *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, pp. 1-8, 2008.

[8] O. Chum and J. Matas, “Matching with PROSAC-Progressive Sample Consensus,” in *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, pp. 220-226, 2005.

[9] V. Lepetit and P. Fua, “Keypoint Recognition Using Randomized Trees,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 9, pp. 1465-1479, July 2006.

[10] E. Rublee, V. Rabaud, K. Konolige, and G.R. Bradski, “ORB: an Efficient Alternative to SIFT or SURF,” in *Proceeding of International Conference on Computer Vision*, Barcelona, Spain, pp. 2564-2571, 2011.

[11] D. Nister and H. Stewenius, “Scalable Recognition with a Vocabulary Tree,” in *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, NY, pp. 2161-2168, 2006.



이수원(Suwon Lee)

2012년 : 한국과학기술원 (공학석사)
2017년 : 한국과학기술원 (공학박사)
2018년～현 재: 경상대학교 컴퓨터과학과 조교수
※관심분야 : 증강현실(Augmented Reality), 컴퓨터비전(Computer Vision) 등