반도체디스플레이기술학회지 제18권 제4호(2019년 12월)
Journal of the Semiconductor & Display Technology, Vol. 18, No. 4. December 2019.

시계열 스트리트뷰 데이터베이스를 이용한 시각적 위치 인식 알고리즘

박천수[*]·최준연[**†]
[*]성균관대학교 컴퓨터교육과, [**†]세종대학교 소프트웨어학과

Visual Location Recognition Using Time-Series Streetview Database

Chun-Su Park[*] and Joon-Yeon Choeh[†]**
[*]Computer Education, Sungkyunkwan University, [**†]Software, Sejong University

ABSTRACT

Nowadays, portable digital cameras such as smart phone cameras are being popularly used for entertainment and visual information recording. Given a database of geo-tagged images, a visual location recognition system can determine the place depicted in a query photo. One of the most common visual location recognition approaches is the bag-of-words method where local image features are clustered into visual words. In this paper, we propose a new bag-of-words-based visual location recognition algorithm using time-series streetview database. The proposed algorithm selects only a small subset of image features which will be used in image retrieval process. By reducing the number of features to be used, the proposed algorithm can reduce the memory requirement of the image database and accelerate the retrieval process.

Key Words : Visual location recognition, Bag-of-words, Geo-tag, Image retrieval, Common keypoints

1. 서 론

스마트 폰 카메라와 같은 디지털 영상 획득 장치의 대중적인 보급으로 일반 사용자가 촬영한 사진을 이용하는 멀티미디어 서비스가 각광을 받고 있다. 그 중 위치 정보를 포함한 영상(geo-tagged images)으로 구성된 데이터베이스를 이용하여 사용자 사진이 촬영 된 물리적 위치를 구하는 시각적 위치 인식(VLR:visual location recognition) 연구가 활발히 이루어지고 있다. VLR 기술은 사용자가 촬영한 사진이 건물이나 인공 모형 등 위치가 고정된 특정 구조물을 포함하는 경우, 데이터베이스 내부의 인근 위치에서 촬영된 이미지를 검색하여 사용자 사진이 촬영된 위치를 추정한다.

[†]E-mail: zoon@sejong.edu

일반적으로 시각 기반 위치 인식 문제를 해결하기 위해 두 가지 작업을 수행해야 한다 [1]. 첫 번째는 데이터베이스 영상 중에서 쿼리(query) 이미지와 시각적으로 유사한 이미지들을 검색하는 작업이고, 두 번째는 검색된 이미지 중에서 쿼리 이미지와 실제로 같은 장소에서 촬영된 이미지를 결정하는 것이다. 위와 같은 작업 과정에서는 데이터베이스의 모든 이미지와 쿼리 이미지 사이의 유사도 계산 과정이 수행되어야 한다. 일반 적으로 시각적 위치 인식에 사용되는 데이터베이스는 수천 장에서 수만 장의 이미지를 저장하고 있기 때문에 이미지의 화소 값(pixel value)을 이용하는 유사도 계산 방식은 상당한 계산 오버헤드(overhead)를 야기한다 [2,3].

이를 해결하기 위해 많은 연구자들은 Bag-of-Words(BoW) 기반의 이미지 검색 기술을 사용해 검색 시스템의 복잡도를 낮추는 연구를 진행해왔다 [4-6]. 기본적으로 BoW기반 기술은 전체 이미지 화소 정보를 이용해 검색 과정을

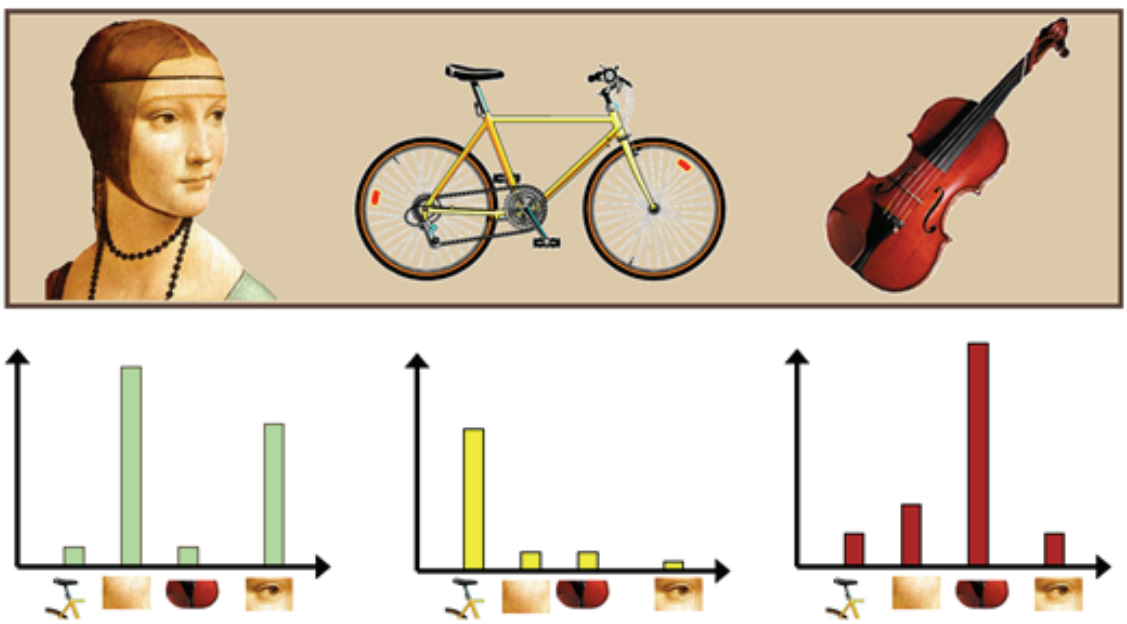

Fig. 1. Image representation using BoW.

수행하는 대신, 이미지 내부의 몇몇 특징점(keypoints)을 구하고 이를 이미지 간에 비교함으로써 유사 이미지를 검색한다. 현재까지 BoW 기반 이미지 검색 기술 이외에도 어휘 트리(vocabulary trees) [7], 빠른 공간 매칭(fast approximate spatial matching) [8,9], 해밍 삽입(Hamming embedding) [10]과 같은 기술들이 제안되었다.

본 논문에서는 시계열 스트리트뷰(time-series streetview) 데이터베이스를 이용하여 이미지 검색에 유용한 특징점은 유지하고, 나머지 특징점은 효과적으로 제거하는 기술을 제안한다. 제안하는 기술은 다른 시간에 인근 위치에서 촬영된 복수의 스트리트뷰 이미지를 이용하여, 가장 최근 이미지의 특징점 중에서 과거 사진에서도 존재하는 공통 특징점(common keypoints)를 선별한다. 이미지 검색 과정에서는 공통 특징점 정보만을 이용하여 쿼리 이미지와의 유사도를 계산한다. 따라서 제안하는 기술은 검색 속도를 향상시키고 동시에 데이터베이스 구성에 필요한 메모리 용량을 줄이는 효과가 있다.

2. BoW 히스토그램을 이용한 이미지 검색

BoW를 사용하는 이미지 검색 기술은 우선 영상 내부의 특징점을 찾고 해당 특징점에서 SIFT, SURF 등의 방법을 사용해 특성 벡터(feature vector)를 구한다. 구해진 특성 벡터는 정해진 코드북(code book)내의 유사 코드워드(codeword)로 양자화 하여 저장된다. 최종적으로 각 이미지는 코드워드 히스토그램(histogram)을 구성하여 해당 이미지의 특성을 표현한다. 이렇게 구한 BoW 히스토그램은 동일 종류의 물체를 포함한 이미지들 사이에서는 유사하고 다른 종류의 물체에 대해서는 서로 다를 것이라는 것이 BoW 히스토그램을 이용한 이미지 검색의 핵심 아이디어이다 (그림 1) [4,11].

쿼리 영상이 입력되면 쿼리 영상의 BoW 히스토그램을 구하고 데이터베이스에 저장된 각 이미지의 BoW 히스토그램과의 유사도를 조사하여 일정 값 이상의 유사도를 보이는 이미지들을 검색 결과로 출력한다. 일반적으로 tf-idf(term-frequency inverse document-frequency) 가중치 방식이 유사도 계산에 많이 사용된다.

BoW 히스토그램 기반의 이미지 검색 기술은 검색 복잡도를 크게 낮출 수 있으며, 데이터베이스 구성에 필요한 메모리 크기를 현격히 낮출 수 있는 것으로 알려져 있다.

3. 제안하는 이미지 기반 위치 인식 시스템

이미지에서 특징점을 추출하는 경우 일반적으로 움직이는 자동차나 사람, 지속적으로 모양이 변하는 나무와 같은 물체에서 많은 특징점이 추출된다. 이러한 가변 특징점은 쿼리 이미지에는 존재하지 않게 되기 때문에 검색 성능에 나쁜 영향을 미친다. 반대로 인근 위치에서 촬

영된 이미지에서 공통으로 존재하는 특징점은 쿼리 영상에서도 존재할 가능성이 높기 때문에 이를 이용하여 이미지 검색 성능을 향상 시킬 수 있다 [12,13]. 따라서 본 논문에서는 데이터베이스 이미지의 BoW 히스토그램을 구성하는 경우 부정확한 가변 특징점들을 사전에 제거하고 공통 특징점을 사용하는 기술을 제안한다. 하위 절에서는 본 논문에서 사용한 시계열 스트리트뷰 데이터베이스를 구성하는 방식과, 이를 이용하여 이미지간의 공통 특징점을 추출하는 기술을 구체적으로 설명한다.

3.1 시계열 스트리트뷰 데이터베이스 구성

Google Maps와 같은 지도서비스에서는 도로에서 360°파노라마로 촬영한 스트리트뷰 서비스를 제공하고 있다 [14]. 각각의 스트리트 이미지는 11대의 카메라를 이용해 촬영한 것이며, 위도, 경도 좌표와 촬영 방향도 수집 가능하다. 스트리트뷰 서비스를 시작한지 10년이 넘어가면서 동일 지역에서 여러 번에 걸쳐 촬영한 사진들이 누적되고 있고, 이들 과거 사진들도 촬영 시기 정보와 함께 제공하고 있다.

스트리트뷰는 이동하는 차량에서 촬영한 것이기 때문에 버스, 자동차, 보행자 등에 의해 건물이 보여지지 않고 가려지는 상황이 필연적으로 발생하는데, 이러한 과거 사진들을 이용하면 장애물로 인해 가려진 부분을 확인하고 식별 하는 것이 가능하다. 그림 2 는 트럭에 가려져서 건물의 일부를 식별할 수 없는 스트리트뷰 이미지인데, 이전에 촬영된 그림 3을 통해 건물 전체를 확인할 수 있다.

Fig. 2. Streetview with obstruction.

본 논문에서는 스트리트뷰 데이터 수집 대상 지역을 미국 샌프란시스코를 선정하였고, 수집 위치는 샌프란시스코 옐로우페이지에서 제공되는 주소지 5088 곳을 선별하였다 [15]. 각 주소 위치에서 특정 시점의 스트리트 이

Fig. 3. Streetview without obstruction.

미지는 촬영 시점에 주소지와 가장 인접한 위치에서 촬영된 이미지를 이용하여 구성하였다. 각 주소 위치에서는 동서남북 4방향으로 이미지를 수집하였으며, 3 시점의 시계열 이미지를 수집하여 총 61,056(=5088*4*3)개의 스트리트뷰 이미지를 확보하였다.

3.2 공통 특징점 추출 기술

제안하는 공통 특징점 추출 기술은 SIFT 방식을 이용하여 각 이미지에서 다수의 특징점을 찾고 해당 특성 벡터를 구하는 것으로 시작한다. 먼저 시계열 스트리트뷰 데이터베이스에 특정 위치에서 촬영된 $\mathbf{N}$ 개의 이미지 $\{\boldsymbol{f_1}, \boldsymbol{f_2}, \dots, \boldsymbol{f_N}\}$가 저장되어 있다고 가정하자. 여기서 $\boldsymbol{f_1}$ 은 시간적으로 가장 최근에 촬영된 영상이고 $\boldsymbol{f_N}$은 가장 오래된 영상이다. 또, $\boldsymbol{k_n^i}$를 $\boldsymbol{f_n}$, $\mathbf{n = 1, 2, \dots, N}$, 이미지의 $\mathbf{i}$ 번째 특징점, $\boldsymbol{d_n^i}$를 해당 특성 벡터라고 정의하자.

앞서 언급한 데로 스트리트뷰 데이터베이스는 정확히 동일 위치가 아닌 인접 위치에서 촬영된 유사한 이미지를 저장한다. 따라서 동일한 사물을 촬영한 이미지라도 이미지 내 사물의 화소 위치(pixel position)와 촬영 각이 다르기 때문에 단순히 화소 위치를 기준으로 공통 특징점을 추출하는 것은 불가능하다. 따라서 제안하는 기술은 화소 위치를 이용하지 않고, 물리적으로 같은 위치를 나타내는 공통 특징점의 경우 코드북 내의 동일한(또는 유사한) 코드워드로 양자화되는 점을 이용한다. 즉, 최근 영상과 과거 영상이 동일한 코드워드로 양자화 되는 특성 벡터를 가지는 경우 해당 특징점을 공통 특징점으로 간주한다.

과거 영상에서 나타난 코드워드를 기록하기 위해 제안하는 기술은 1차원 이진(binary) 마스크(mask) $\mathbf{U(x)}$, $\mathbf{x = 1, 2, \dots, X}$,를 이용한다. 여기서 $\mathbf{X}$ 는 전체 코드워드의 계수(코드북 크기)를 나타내며 이진마스크의 초기값은 0으

로 설정된다. 공통 특징점을 추출하기 위해서 제안하는 기술은 먼저 $\{\boldsymbol{f_2}, \dots, \boldsymbol{f_N}\}$ 과거 영상들의 특징점과 특성벡터를 구한다. 구해진 과거 이미지의 특성 벡터를 이용하여 각 특성벡터($\boldsymbol{d_n^i}$)와 유사한 $\mathbf{M}$ 개의 코드워드 $\{\boldsymbol{c_n^i(1)}, \boldsymbol{c_n^i(2)}, \dots, \boldsymbol{c_n^i(M)}\}$ 를 구하고 이를 1차원 이진 마스크 $\mathbf{U(x)}$에 다음과 같이 기록한다.

$$\mathbf{U}\left(\boldsymbol{c_n^i(m)}\right) = \mathbf{1}$$

여기서 $\mathbf{m = 1, 2, \dots, M}$이다. 동일한 마스크 기록과정을 모든 과거 영상을 대상으로 진행한다.

제안하는 기술은 이진 마스크를 구성한 후 $\boldsymbol{f_1}$에서 추출된 특성 벡터 중 과거 이미지에서 공통으로 존재했던 특성 벡터를 선별하는 과정을 거친다. 먼저 $\boldsymbol{f_1}$에서 추출된 각 특성 벡터 $\boldsymbol{k_1^i}$와 가장 유사한 코드워드 $\boldsymbol{c_i^i(1)}$를 구한다. 그 후 아래의 조건을 사용하여 전체 특징점 중 공통 특징점을 선별하여 집합 $\mathbf{K}$를 구성한다.

$$\mathbf{M} = \{\boldsymbol{k_n^i} | \mathbf{U}\left(\boldsymbol{c_1^i(1)}\right) = \mathbf{1}\}$$

최종적으로 구해진 공통 특징점 집합 $\mathbf{K}$를 이용하여 최근 영상의 BoW 히스토그램을 구하고 이를 이미지 검색 과정에 사용한다.

4. 시스템 비교 및 결과

본 논문은 제안하는 기술의 효과를 로드뷰 기반으로 구성된 시계열 스트리트뷰 데이터베이스를 이용하여 측정한다. 실험에 사용된 시계열 스트리트뷰 데이터베이스는 총 5058 위치에서 4개의 방향으로 3시계열 스트리트 이미지로 구성되었다. 특징 벡터 양자화에는 100,000개의 코드워드를 가지는 코드북을 이용하였다. 검색 성능을 측정하기 위해 총 334개의 쿼리 영상을 이용하였다. 본 논문에서는 제안하는 공통 특징점을 이용한 시각적 위치인식 알고리즘을 가장 최근 스트리트 이미지에서 추출된 모든 특징점을 사용하는 위치인식 알고리즘과 성능을 비교하였다 [16]. 이미지 검색은 tf-idf방식을 사용하여 수행하였다.

표 1은 실험 결과를 정리해서 보여준다. 제안하는 공통점 특징 방식은 이미지 검색에 사용되는 평균 특징점의 개수를 약 26.68% 정도 감소시키는 효과를 보였다. 또한 특징점 감소로 인해 평균 검색 시간이 약 15.65% 감소하는 것으로 조사 되었다.

위와 같은 속도 향상을 보임에도 제안하는 알고리즘은 기존 방식보다 약간 우수한 검색 성능을 보였다. 표 2는

Table 1. Number of Keypoints and Retrieval Complexity

	Number of keypoints per image	Time complexity (ms)
Conventional	2916	607
Proposed	2138	512

기존 방식과 제안하는 방식의 검색 성능을 보여준다. 표 2에서 $\mathbf{P(i), i = 10, 100, 1000}$은 334개의 쿼리 이미지를 데이터베이스에서 검색하여 상위 i개의 검색 결과에 실제 같은 위치에서 촬영된 스트리트뷰 이미지가 있을 확률을 나타낸다. 예를 들어 334개의 쿼리 이미지를 검색하여 상위 1000개의 검색 결과에 같은 위치에서 촬영된 스트리트뷰 이미지가 있을 확률$\mathbf{P(1000)}$은 기존 방식은 88.32%, 제안하는 방식은 88.92%로 조사되었다. 따라서 제안하는 기술은 검색 성능을 유지하면서 메모리 사용을 줄이고 검색 속도를 높일 수 있다.

Table 2. Performance Analysis

	$\mathbf{P(10)}$	$\mathbf{P(100)}$	$\mathbf{P(1000)}$
Conventional	4	90	295
Proposed	5	92	297

감사의 글

본 연구는 한국연구재단에서 지원하는 연구비를 지원받아 수행하였음(NRF-2019R1F1A1055593).

참고문헌

1. T. Sattler, M. Havlena, K. Schindler, M. Pollefeys, “Large-Scale Location Recognition and the Geometric Burstiness Problem”, IEEE conference on computer vision and pattern recognition, pp. 1582–1590, 2016.
2. G. Schindler, M. Brown, Szeliski, “City-scale location recognition”, IEEE conference on computer vision and pattern recognition, pp. 1–7, 2007.
3. Y. Li, D. J. Crandall, D. P. Huttenlocher, “Landmark classification in large-scale image collections”, Proceedings of the IEEE International Conference on Computer Vision, pp. 1957–1964, 2009.
4. Recognizing and Learning Object Categories, http://people.csail.mit.edu/torralba/shortCourseRLOC/index.html.
5. P. Bhattacharya, M. Gavrilova, “A survey of landmark recognition using the bag-of-words framework”, Intelligent Computer Graphics, pp. 243–263. 2012.
6. D. Nist´er, H. Stew, “Scalable recognition with a

vocabulary tree", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2161–2168, 2006.
7. D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree", IEEE conference on computer vision and pattern recognition, pp. 2161–2168, 2006.
8. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching", IEEE conference on computer vision and pattern recognition, pp. 1–8, 2007.
9. G. Tolias and Y. Avrithis, "Speeded-up Relaxed Spatial Matching", ICCV, pp. 1653–1660, 2011.
10. H. Jegou, M. Douze, and C. Schmid, "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search", ECCV, pp.304-317, 2008.
11. http://darkpgmr.tistory.com/125
12. J. Knopp, J. Sivic, T. Pajdla, "Avoiding confusing features in place recognition", LNCS, vol. 6311, pp. 748–761, 2010.
13. N. Naik, et al. "Computer vision uncovers predictors of physical urban change", Proceedings of the National Academy of Sciences, pp. 7571-7576, 2017.
14. https://www.google.com/streetview/
15. https://www.yellowpages.com/san-francisco-ca
16. J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, "Object Retrieval with Large Vocabularies and Fast Spatial Matching", Computer Vision and Pattern Recognition, 2007.

접수일: 2019년 11월 27일, 심사일: 2019년 12월 12일, 게재확정일: 2019년 12월 13일