다단계 클러스터링 기법을 이용한 이미지 클러스터링 기법에 관한 연구*

한정규, 김석대, 황수찬
한국 항공대학교 컴퓨터 공학과
{jazzjay, bigrock94, schwang}@mail.hangkong.ac.kr

Similarity-based Image Clustering Method using Hierarchical Clustering Technique

요약

본 논문에서는 유사도(similarity) 기반 이미지 클러스터링 기법에 대하여 논하고자 한다. 비트맵 이미지의 특징을 추출하고 이러한 특징에 기반한 유사도 측정 함수들을 소개하고 이미지 클러스터링 알고리즘과 구현을 통한 실험 예제들에 대해서 설명한다. 이 실험에서 우리는 유사도에 따라 이미지들이 계층적(Hierarchical)으로 집단화 되는 계층적 클러스터링 알고리즘을 사용하였다. 이미지의 특징 표현을 위해서는 HSV 기반의 히스토그램을 이용하였다. 본 논문에서 제안한 기법의 실험 결과는 이미지 데이터베이스에서 유사한 이미지를 검색하는데 높은 효율성이 있는 것을 보여준다.

1. 서론

최근 대용량의 이미지 데이타베이스에 대한 내용기반 검색을 위한 기법들에 대해 많은 연구가 진행되어 왔다 [1,2,3,4]. 내용기반 정보검색(CBIR) 시스템에서 사용자들은 이미지의 특징정보에 기반하여 웹 페이지나 데이터베이스로부터 관련 있는 이미지들을 검색할 수 있다.

특징기반 질의는 색상(color), 무늬(texture), 모양(shape), 윤곽(contour) 등과 같은 이미지의 특징에 따라 질의 방법과 결과가 달라진다. 예를 들어서, 색상이나 무늬를 이용한 질의는 사용자가 색상과 무늬를 이용하여 질의를 할 수 있게 한다[3]. 특징 정보는 사용자에 의해 제공되거나 시스템에 의해 생성되고 그 정보들은 데이터베이스에 원래의 이미지와 함께 저장되고 검색의 지표로 사용된다.

기존 텍스트 기반의 데이터베이스에서와 달리 이미지에 대한 검색은 유사질의를 기반으로 수행된다. 유사질의는 이미지 간의 유사도(similarity)를 이용하여 연관 이미지를 검색하는 것을 말한다. 사용자는 예제 이미지를 제출하고 데이터베이스 내에서 이와 일치하거나 혹은 유사한 이미지를 검색하도록 한다. 일반적으로 유사질의의 결과는 하나의 이미지가 아닌 시스템에 의해 정의된 유사 거리 내의 이미지들의 집합이다.

전형적인 검색 알고리즘들은 사용자가 검색하고자 하는 이미지와 유사한 특징을 가진 질의 이미지를 제공할 것을 요구한다. 그러한 알고리즘들은 일반적으로 데이터베이스에서 최적의 검색 결과를 반환한다. 그러나 검색과정에서 질의 이미지는 대용량 이미지 데이터베이스나 웹 페이지에서 저장된 모든 이미지와 비교 계산을 하게 되어 많은 시간적인 비용이 발생한다. 이러한 문제를 해결하기 위해 많은 이미지 검색 시스템이나 인터넷 이미지 검색 엔진들은 비슷한 특징을 가진 이미지들을 클러스터로 집단화해서 데이터베이스에 저장하는 클러스터 기반 인덱싱 기술을 필요로 하게 된다. 즉, 검색 시에 주어진 질의 이미지와 관련 있는 클러스터에 있는 이미지들만이 비교대상이 된다 [1,6].

본 논문에서는 이미지 데이터베이스에 대한 **유사도 기반 다단계 클러스터링(Hierarchical Clustering) 기법**을 소개한다. 이를 위해 비트맵 이미지에 대한 특징 추출 기법과 색상정보에 기반한 유사도 측정 함수에 대해 기술한다. 또한 클러스터들에 대한 빠른 인덱스를 위해 다단계 기법을 이용한다[1]. 다양한 이미지 그룹에 대한 실험은 본 논문에서 제안한 기법이 비슷한 이미지들을 구성하는데 매우 유용함을 보여준다.

본 논문의 나머지 부분은 다음과 같이 구성되어 있다. 2장에서는 본 논문과 관련 되어 있는 앞 선 연구들에 대해 소개하고 3장에서는 이미지 클러스터링 과정과 구현에 대해서, 4장에서는 다양한 이미지 그룹에 대한 실험과 그 결과에 대해 기술하고 있다. 마지막에는 결론과 향후 연구 방향에 대해 기술한다.

2. 관련 연구

2.1 내용기반 이미지 검색 시스템들

많은 이미지 검색 시스템들이 이미지들의 내용에 기반한 검색을 위해 많은 연구를 진행 중이다. Serge의 Blobworld 시스템은 내용기반정보검색 시스템 중 대표적인 검색 시스템이다[4,5]. 이 시스템은 객체 기반 질의를 지원하고 이미지 내용에 대한 자동색인을 지원한다. 이 시스템에서 이미지들은 세크먼테이션(조각)을 통한 특징 정보로 색상, 무늬 등을 포함한다.

C-BIRD는 디지털 라이브러리를 구성하기 위한 내용기반 이미지 검색 시스템이다[6, 7, 8]. C-BIRD는 키워드, 색상 히스토그램, 무늬, 모양과 같은 다양한 이미지 특징을 사용한 내용 기반 이미지 검색을 허용한다. 또한 C-BIRD는 관련 있는 키워드를 자동적으로 생성한다.

QBIC [9], Chabot [10] 그리고 Photobook [11]과 같은 자료 중심 질의들에 초점을 둔 시스템들도 있다. 이런 시스템들은 내용 기반 질의를 구현하기 위해 색상과 무늬 같은 저 수준의 이미지 특징들을 사용한다. SKICAT(Sky Image Cataloging and Analysis Tool) [12]은 이미지 세그먼테이션(조각) 과정에서 얻어진 객체들을 분류하기 위해 트리나 통계적인 최적화 방법을 사용한다. SKICAT은 천체 이미지들에 대한 연구를 위해 개발된 시스템이다

2.2 색상 히스토그램(Color Histogram)

이미지 데이터의 특징 표현 방식의 하나로 히스토그램은 효율성과 정확성을 위해 내용 기반 이미지 검색 시스템에서 널리 사용되고 있다. 히스토그램의 계산은 쉽지 않지만 그 단순성으로 인해 내용 기반 이미지 검색 시스템에서 효과적인 요소로 적용된다. 색상에 따른 이미지 검색을 위한 색상 히스토그램 특징을 이용한 방법에는 두 가지가 있다. 첫 번째는 지역색상정보(local color feature)에 기반한 검색이다. 이 방법은 하나의 이미지를 직사각형 영역들로 나눈다. 그 후 나뉘어진 모든 이미지는 이런 직사각형 영역에 대응하는 정규화된 히스토그램의 집합으로 표현된다. 이 방법은 직사각형 영역의 크기를 선택하는 것이 중요하다. 반대로 하나의 이미지를 하나의 영역으로 처리하는 전역색상정보(global color feature) 방식이 있다.

* 본 논문은 과학기술부 한국화학재단 지정 경기도 지역협력연구센터(RRC)인 한국항공대학교 인터넷정보검색연구센터의 지원에 의한 것임

지역색상정보를 사용하는 경우, 두 이미지 사이의 유사도는 대응되는 직사각형 영역의 히스토그램 사이의 차이를 계산 함으로서 측정된다. 결과적으로 두 이미지 간의 유사 도는 대응되는 영역 간의 유사도들의 합으로 계산된다. 히스토그램들을 비교하기 위해 교차측정기법이 사용된다 [1]. 주어진 두 개의 정규화된 히스토그램, $p=\{p1,p2, ..., p_m\}$, $q=\{q1,q2, ..., q_m\}$ 의 유사도 $S_{p,q}$는 다음과 같이 정의된다.

$$S_{p,q} = \sum \min(p_i, q_i)$$

최근, 몇몇 연구자들에 의해 공간 정보를 조합시킴으로써 색상히스토그램을 개선하는 기법이 제안되었다. Hsu등은 이미지에서 구분되는 색상들간의 공간 배열 정보를 얻으려고 시도했다 [13]. 이 방법에서 이미지는 최대 엔트로피를 사용하는 직사각형 영역으로 나뉘어지고 각 영역은 우세한 하나의 색상 가진다. 두 이미지 간의 유사도는 같은 칼라를 가진 영역 사이의 겹치는 정도를 말한다. Hsu등이 제안한 기법은 실질적인 계산, 특히 분할 알고리즘을 요구한다.

Huang.등이 제안한 기법은 칼라들간의 공간 상관 관계를 이용한다[14]. 그들의 접근방법은 색상 상관도(correlograms)로 불리며 공간 데이터 분석에서의 상관도기술과 관련된다.

3. 이미지 클러스터링

이번 장에서는 전역색상정보를 이용한 유사도 기반 이미지 클러스터링 기법을 소개한다. 이미지를 클러스터링 하기 위해 본 논문에서는 Philips Lab의 계층적 클러스터링 알고리즘을 이용하였다[1]. 이 알고리즘의 주요 목적은 하나의 클러스터에 유사한 이미지들을 집단화하고 그 클러스터의 중심 이미지를 계산하는 것이다. 결과적으로 데이터베이스내의 모든 이미지와 질의 이미지를 모두 비교하지 않고 하나의 클러스터내의 이미지들에 대한 비교만을 수행 하도록 한다.

아래의 그림 3.1은 이미지 클러스터링 과정을 나타낸다.

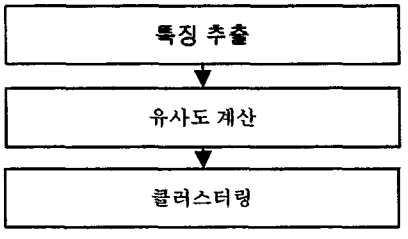

그림 3.1 이미지 클러스터링 과정

3.1 특징 추출

본 논문에서 제안한 이미지 클러스터링 기법의 첫 번째 단계는 모든 이미지의 칼라 특징 값들을 계산하는 것이다. 우리의 실험에서는 HSV 색상공간을 이용한 128 bin의 색상 히스토그램을 사용한다.

3.2 유사도 측정과 클러스터링

이미지 검색을 위한 유사도 계산을 위해 많은 히스토그램 기반 방법론들이 제안되어 왔다. 그러한 접근들은 유사도 측정을 위해 색상히스토그램을 사용한다. 유사도는 데이터베이스에 있는 이미지와 질의 이미지의 색상히스토그램을 비교함으로써 얻어진다. 이러한 접근들의 차이점은 주로 색상 공간의 선택과 히스토그램간의 유사도 계산 방법에 있다.

논문의 초반에 언급했던 것처럼 본 논문에서는 하나의 이미지에서 하나의 색상히스토그램을 생성하는 전역적히스토그램을 사용한다. 즉 전체 이미지는 전역적인 색상 정보를 반영하는 하나의 영역으로서 간주된다. 본 논문의 시스템은 이 히스토그램에 기반한 유사도를 생성한다.

제안하는 시스템은 반복되는 프로시저에 의해 이미지 클러스터를 생성한다. 최초로 모든 가능한 이미지 쌍들을 비교한다. 한 쌍의 이미지의 유사도가 주어진 초기 값보다 더 크다면 그 두 이미지는 같은 클러스터에 속한다. 다음 단계에서는 위에서 생성된 클러스터와 처음 단계에서 포함되지 않았던 나머지 이미지들이 고려 대상이 된다. 동일한 단계가 모든 이미지 쌍과 클러스터에 적용된다. 이러한 단계는 클러스터의 수가 미리 정의된 값에 다다르거나 모든 이미지들이 클러스터링 될 때까지 수행된다.

알고리즘 3.1은 본 논문의 클러스터링 프로시저를 보여준다. 이 알고리즘에서 *OBJECT* 구조는 이미지의 메모리상 데이타구조이다. 이 알고리즘에는 이미지 식별자, 색상 히스토그램과 경로 정보가 포함되어 있다. 본 논문에서는 이러한 정보가 클러스터링 단계 이전에 미리 계산된 것으로 가정한다.

```
---------------------------------------------------------------------
OBJECT array[MAXNUM_OF_IMAGES];
CLSTR cluster[MAXNUM_OF_CLUSTERS];
1. Set first =1, second =1
2. Select a next pair of cluster : cluster[first],cluster[second]
  3. Set index1=first, index2 = second. where index1 and index2 is a
 index of array[]
  4. IF(clutstered(cluster[first]) OR
   clutstered(cluster[second])) then goto step 2.
  5. IF(cluster[first] != LEAF) set index1 = center image of cluster[first];
   IF(cluster[second] != LEAF) set index2 = center image of
 cluster[second];
6. Compute similarity:
   img_ComputeSimilarity(array[index1].image,  array[index2].image,
 &ImageSim)
7. IF (ImageSim < Smallest_Similarity ) Set
   Smallest_Similarity = ImageSim,
   selected1 = first, selected2 = second
  8. IF ((first < Cluster_Counter) OR (second < Cluster_Counter r)) then
 goto step 2
9. IF (Smallest_Similarity < threshold)
 MakeNewCluster(selected1, selected2);
10. IF(Cluster_Counter <
 MAXNUM_OF_CLUSTERS) then goto step 1
11. Stop.
---------------------------------------------------------------------
```

알고리즘 3.1. 클러스터링 알고리즘

*CLSTR*은 클러스터를 표현하는 데이터 구조이다. *CLSTR*은 *OBJECT* 구조에 대한 인덱스로 사용된다.

본 논문에서 제안하는 알고리즘에서는 단계가 깊어질수록 탐색할 클러스터의 개수는 절반 이상으로 감소한다. 이것이 계층적 클러스터링 알고리즘을 사용하는 주된 장점중의 하나이다.

3.3 클러스터링 예제

그림 3.2는 8개의 예제 이미지들에 대한 계층적 클러스터링 과정을 보여준다. 예제의 경우 클러스터링 과정은 두 개의 클러스터가 남을 때까지 계속된다. 세 개의 클러스터 9,10,11은 첫 번째 클러스터링 단계에서 생성된다. 이미지 5와 8은 아직 어떤 클러스터에도 속하지 않는다. 다음 단계에서 이미지 5와 8은 클러스터 9,10,11을 클러스터링 타깃으로 고려한다. 두 번째 단계의 결과로 클러스터 12와 13이 생성된다. 세 번째 단계에서 클러스터 14가 생성되고 마지막 두 개의 클러스터가 남았기 때문에 클러스터링 과정이 끝나게 된다. 질의가 주어 졌을 때 검색 과정은 질의 이미지와 두 개의 클러스터 14와 12를 비교하면서 시작된다. 만일 질의 이미지가 3번 이미지와 유사하다면 클러스터 13과 9가 차례로 검색 될 것이다. 클러스터 9에 포함된 모든 이미지들이 질의 이미지와 비교될 것이고 질의 결과도 여기서 선택 될 것이다.

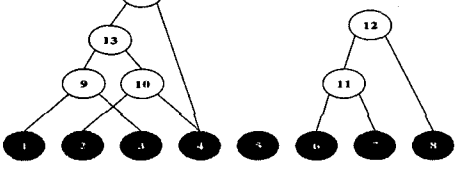

그림 3.2. 클러스터링 과정 예제

4. 실험

이번 장에서는 본 논문에서 제안한 이미지 클러스터링 시스템의 실험 결과를 논의 한다. 특징 추출과 유사도 계산 함수를 정의하고 색상 히스토그램에 기반한 계층적 클러스터링 알고리즘을 구현하였다.

실험은 인위적으로 생성된 간단한 10개의 이미지를 가지고 진행한다. 그림 4.1은 생성되는 클러스터의 수를 네 개로 했을 때의 클러스터링 결과를 보여준다. 이 실험에서 알고리즘3.1은 그림4.1과 같이 색상히스토그램에 기반한 클러스터를 생성하였다. 두 개의 이미지 2,3을 포함하는 클러스터 12는 대부분 빨강(red)과 작은 흰(white) 원으로 구성되어 있다. 두 번째 단계에서 클러스터 15와 14가 생성되고 클러스터의 수가 네 개가 된다. 이미지 4의 경우 그 칼라 특징은 이미지 3과 7로부터 거의 같은 거리에 위치한다. 그러나 클러스터 12의 중심 이미지의 색상히스토그램은 클러스터 13의 칼라 히스토그램보다 더 가까운 특징 값을 갖고 있다. 그 이유는 칼라 히스토그램이 계산 될 때 클러스터 13의 중심 이미지에 있는 파랑(blue) 의 값이 감소했기 때문이다.

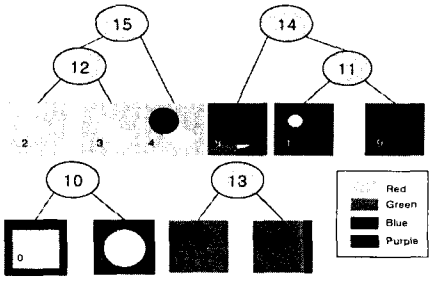

그림 4.1. 클러스터링 예제

극단적인 경우로 그림 4.2의 두 이미지를 고려할 수 있다. 이 두 이미지는 완전히 다른 그림이지만 동일한 칼라 분배와 방향성을 갖고 있다. 이것은 골퍼 셔츠의 빨강(red)과 꽃의 빨강(red)이 거의 같은 양의 값을 갖고 있기 때문이다.

그림 4.2. 동일한 칼라 분배를 갖는 다른 이미지의 예

5. 결론

본 논문에서는 이미지 데이터베이스에 대한 유사도 기반 다단계 클러스터링 기법과 그 실험을 소개했다. 비트맵 이미지에 대한 특징 추출 기법과 색상 정보에 기반한 유사도 함수를 제안했다. 본 논문에서는 클러스터에 이미지를 집단화 하기 위해 계층적 클러스터링 알고리즘을 사용하였다.

다양한 그룹의 이미지에 대한 실험은 본 논문에서 제안한 기법은 적은 이미지 데이터베이스에 대해 유용하다는 것을 보여준다.

향후 연구로서 특징 추출 함수를 확장하기 위해 키워드, 공간 정보, 객체의 모양 같은 부가적인 이미지 특징들을 고려하고자 한다. 그리고 좀더 효율적으로 클러스터를 생성하고 클러스터에 있는 노이즈 이미지를 줄이기 위해 계층적 클러스터링 알고리즘을 개선시키고자 한다.

{참조}

[1] S. Krishnamachari, M. Abdel-Mottaleb., "Hierarchical clustering algorithm for fast image retrieval", Part of the IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases VII, San Jose, California, January 1999. pp427-435

[2] J. Chen, C. Bouman and J. Allebach, School of Electrical and Computer Engineering Purdue University, "Fast Image Database Search using Tree-Structured VQ", Proceedings of ICIP97, vol. 2, pp. 827-830, SantaBarbara, CA, Oct. 26-29, 1997.

[3] C. Ordonez and E. Omiecinski, College of Computing Georgia Institute.,"Discovering Association Rules based on Image Content", Proceedings of the IEEE Advances in Digital Libraries Conference,ADL'99, 1999.

[4] C. Carson, S. Belongie, H. Greenspan, J. Malik., "Region-Based Image Querying", Proceedings CVPR '97 Workshop on Content-Based Access of Image and Video Libraries., 1997.

[5] S. Belongie, C. Carson, H. Greenspan, J. Malik, "Recognition of Images in Large Databases Using a Learning Framework", Technical Report TR 97-939,U.C. Berkeley, CS Division, 1997.

[6] O. Zaiane, J. Han, Z. Li, J. Hou, "Mining Multimedia Data", Proc. CASCON'98: Meeting of Minds, Toronto, Canada, November 1998.

[7] O. Zaiane, J. Han, Z. Li, J. Chiang, and S. Chee, "MultiMedia-Miner: A System Prototype for MultiMedia Data Mining", Proc. 1998 ACM-SIGMOD Conf. on Management of Data, (system demo), Seattle, Washington, June 1998.

[8] Z. Li, O. Zaiane, B. Yan , "C-BIRD: Content-Based Image Retrieval from Digital Libraries Using Illumination Invariance and Recognition Kernel", In International Workshop on Storage and Retrival Issues in Image and Multimedia Databases, in conjunction with the 9th International Conference on Database and Expert Systems (DEXA'98), Vienna, Austria, August 1998.

[9] C. Faloutsos, W. Equitz, M. Flickner, W. Niblack, D. Petkovic, R. Barber., "Efficient and Effective Querying by Image Content", Journal of Intelligent Information Systems, vol 3, pp. 231-262.1994.

[10] Virginia E. Ogle, M. Stonebraker, University of California, Berkeley., "Chabot : retrieval from a relational database of images", IEEE Computer, vol.28, no.9, pp. 40-48, 1995

[11] A. Pentland, W. Picard, S. Sclaroff., "Photobook:Content-based manipulation of image databases", International Journal on Computer Vision, Fall 1995.

[12] U. Fayyad, , D. Haussler, and P. Storoltz., "Mining scientific data". Communicationsof the ACM, 39(11):51–57,November 1996.

[13] W. Hsu, T. Chua, and H. Pung., "An Integrated color-spatial approach to content-based image retrieval", In ACM Multimedia Conference, pages 305-313, 1995.

[14] J. Huang, S. Kumar, M. Mitra, W. Zhu, R. Zabih., "Image Indexing Using Color Correlograms", In IEEE Conference on Computer Vision and Pattern Recognition, pages 762-768, 1997.