시리즈 방송 콘텐츠에 대한 장소 기반의 씬 그룹화 방법

김혜린 ○ 낭종호
서강대학교 컴퓨터공학과
hyerin16@gmail.com, jhnang@sogang.ac.kr

Background Based Scene Grouping in Video

Hye Rin Kim ○ Jongho Nang
Department of Computer Science and Engineering, Sogang University

요 약

최근 동영상의 내용 검색에 대한 요구가 증가하면서, 동영상 태깅이 필요하게 되었다. 시리즈 방송 콘텐츠는 한정된 세트장 개수로 인하여 장소의 중복성이 발생하는데, 이를 자동으로 검출함으로써 태깅 비용을 줄일 수 있다. 이러한 태깅을 위해서, 장소의 중복성을 자동으로 검출하고 그룹화하는 방법을 제안한다. 시리즈 방송 콘텐츠의 장소는 부분적인 중복과 뷰포인트, 조명의 변화가 발생하므로 SIFT피쳐를 이용하여 매칭한다. 이는 연산량이 매우 많은 피쳐이므로 색상 정보를 이용하여 후보군을 줄여주었고 Bag-Of-Words모델을 이용하여 유사도가 낮은 대상을 검색대상에서 제외시켰다. 마지막으로 후-검증을 통해 오-검출을 감소시켰다. 이러한 방법을 통해서 정확률은 약 0.8, 재현률은 약 0.6을 얻었고 가속화 방법을 통해 SIFT대비 연산량이 약 0.4배 감소하였다.

1. 서 론

시리즈 방송 콘텐츠는 일정 수의 세트장에서 촬영되어 같은 장소에서 촬영된 구간이 빈번하다. 본 논문에서는 시리즈 콘텐츠에서 장소적 중복성을 자동으로 검출하는 방법을 연구한다. 이를 통해 장소 기반의 검색이나 동영상 태깅 비용의 절감 등의 응용이 가능하다.

일반적으로 영상의 장소 중복은 <그림1-(a),(c)>와 같이 뷰포인트나 조명의 변화가 적용된다. 그리고 중복 영역은 <그림1-(b),(d)>와 같이 부분적으로 나타난다. 따라서 장소의 중복성을 검출하기 위해서는 영상을 SIFT와 같은 로컬특징의 집합으로 표현하는 것이 적합하다.


(a) 조명, 뷰포인트 변화


(b) 부분적인 중복 영역


(c) 조명, 뷰포인트 변화


(d) 부분적인 중복 영역

<그림 1> 장소의 중복성 예시

로컬 특징을 이용하여 장소의 중복성을 검출하기 위해서는 크게 세 가지의 문제를 해결해야 한다. 첫째로 콘텐츠 전체 프레임들에 대한 연산은 검색 공간이 매우 크므로 가속화를 위한 적합한 색인 방법이 필요하다. 둘째로 등장인물과 같은 전경이 매칭되는 경우를 구별할 수 있어야 한다. 마지막으로 로컬특징 간의 오-매칭(False matching)으로 인한 에러를 줄여야 한다.

본 논문에서는 동영상을 의미 단위인 씬으로 구성되어 있음을 가정한다. 두 씬이 동일한 장소인지의 여부는 각 씬에 포함된 프레임들간의 장소적 중복성 검출을 통하여 판단된다. 프레임 간의 장소 중복 영역은 SIFT매칭을 통하여 이루어진다. 이 때 연산량이 매우 크므로 영상의 전역적인 색상 정보의 유사성을 이용하여 SIFT매칭 대상을 빠르게 줄이는 방법을 적용하였다. 그리고 SIFT매칭 시 128차원 공간에서의 최근접 검색 비용을 줄이기 위하여, Bag-of-Words[2,3]의 방법에 기반하여 128차원을 학습된 코드북을 통한 1차원 공간으로 매핑하는 방식을 적용하였다. 마지막으로 양자화 벡터는 양자화 에러로 인하여 오-매칭이 증가하므로 [4,5]에서 제안한 후-검증 방법을 적용하였다.

각 8회로 구성된 시리즈 드라마 네 편에 대하여 실험한 결과, 정확률(Precision)은 약 0.8, 재현률(Recall)은 약 0.6을 얻었다. 색상 정보와 벡터 양자화를 이용한 가속화를 통해 씬 간의 비교 연산량이 약 0.4%정도 감소하는 효과를 얻었다.

2. 프레임간의 장소 중복 영역 검출

· 본 연구는 미래창조과학부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10044615, 클라우드 기반 개방형 소셜 방송미디어 콘텐츠 융합 생성, 편집 및 재생을 위한 미디어 제작 및 전송 시스템 개발]

프레임 간의 장소 중복 영역 검출 단계는 색상 정보를 이용하여 유사성이 낮은 프레임을 빠르게 거르는 과정과 프레임간의 SIFT 매칭 과정 및 SIFT 의 오리엔트를 이용한 후-검증 단계로 구성된다.

A. 색상 정보를 이용한 빠른 후보군 선별

두 영상 I_i, I_j의 SIFT 특징 벡터 수를 각각 n_i, n_j라 하면 두 영상간의 SIFT 매칭은 $n_i \times n_j$ 만큼의 $\|\mathrm{x}-\mathrm{y}\|_2,\ \mathrm{x},\mathrm{y} \in \mathbb{R}^{128}$ 연산이 필요하다. 즉 다수의 프레임들에 대한 SIFT 매칭은 매우 많은 연산량을 요구한다. 본 논문에서는 영상의 전역적인 색상 정보를 이용하여 SIFT 매칭 전 색상 유사도가 낮은 영상을 미리 제외하는 방법을 적용하였다. 먼저 영상 I_i는 HSV 색상 공간으로 변환된 후 HMMD 양자화[6]를 통하여 색상 히스토그램 $H_i \in \mathbb{R}^d$로 추상화된다. d는 HMMD 양자화 시 빈(bin)의 수로써 히스토그램의 차원을 결정한다. 색상 히스토그램간의 유사도는 장소 중복영역이 부분적으로 발생함을 고려하여 식(1)과 같이 히스토그램 교집합을 사용하였고 그 결과는 <그림 2>와 같다.

$$ColorSimilarity(I_i, I_j) = \sum_{k\in[1,d)} \min(H_i(k), H_j(k)) \quad (1)$$

<그림 2> 색상 정보를 이용한 SIFT 매칭 후보군 예시

B. SIFT 매칭

색상 히스토그램을 이용하여 선별된 <그림 2>와 같은 후보 영상에 대하여 SIFT 매칭을 수행한다. 색상 히스토그램은 영상에 대한 전역적인 특징이므로 영상의 기하학적인 변화에 민감하고 SIFT 매칭에 비하여 분별력이 낮다. 즉 색상 유사도의 임계치를 높게 설정할 경우 SIFT 매칭 대상이 적으므로 연산량이 줄어들지만, 재현률이 떨어지는 문제점이 있다. 따라서 본 논문에서는 색상 정보를 통한 SIFT 매칭 대상을 줄이는 과정과 함께, 벡터 양자화를 통한 SIFT 매칭 비용을 감소 시키는 방법을 적용하였다. SIFT 특징 벡터 $\mathrm{x} \in \mathbb{R}^{128}$에 대한 양자화 함수는 식(2)와 같이 정의 된다.

$$\mathcal{Q} : \mathrm{x} \mapsto [1, K] \quad (2)$$

일반적으로 Bag-of-Words[2,3]모델에서 매핑함수는 코드북 $\mathrm{Z} \in \mathbb{R}^{128\times K}$를 이용하여 식(3)과 같이 x와 가장 가까운 z_k 의 인덱스로 매핑하며 Z 는 식(4)와 같이 K 평균 클러스터링을 통해 총 왜곡 값이 최소화되도록 선별된 코드벡터들로 구성된다.

$$\mathcal{Q}(\mathrm{x}) = \mathrm{argmin}_{k\in[1,K]} \|\mathrm{z}_k - \mathrm{x}\|_p \quad (3)$$

$$\underset{\mathbb{G}}{\mathrm{argmin}} \sum_{i=1}^{K} \sum_{\mathrm{x}_j \in G_i} \|\mathrm{x}_j - \mathrm{z}_i\|_p \quad ,$$

$$\mathbb{G} = \{G_1, G_2, .., G_K\}, \quad \mathrm{z}_i = \bar{\mathrm{x}}_j \text{ for } \forall \mathrm{x}_j \in G_i \quad (4)$$

따라서 두 영상간의 SIFT 매칭 시 128 차원 공간에서의 최근접 검색을 위한 $\|\mathrm{x}-\mathrm{y}\|_2,\ \mathrm{x},\mathrm{y} \in \mathbb{R}^{128}$ 연산은 식(5)와 같이 자연수 비교를 통해서 근사화 된다.

$$\delta(\mathrm{x},\mathrm{y}) = \begin{cases} 1, & \mathcal{Q}(\mathrm{x}) = \mathcal{Q}(\mathrm{y}) \\ 0, & \mathcal{Q}(\mathrm{x}) \neq \mathcal{Q}(\mathrm{y}) \end{cases} \quad (5)$$

두 SIFT 특징 벡터 x,y간의 매칭 함수는 식(6)과 같이 x가 y의 최근접이고 x와 y의 거리가 임계치보다 작은 경우 1 을 반환하도록 정의된다.

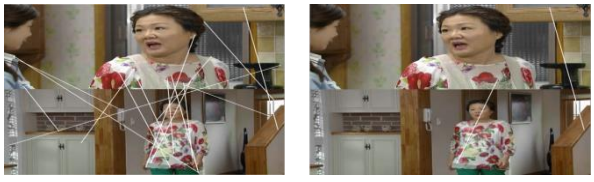
$$f_{NN}(\mathrm{x},\mathrm{y}) = \begin{cases} 1, & \text{if x is a NN of y and } \|\mathrm{x}-\mathrm{y}\|_2 < \varepsilon \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

따라서 프레임간의 장소 중복 여부를 검출하는 함수 PMF 는 식(7)과 같이 정의된다. 최근접 검색 계산시 $\delta(\mathrm{x}_i, \mathrm{y}_j)$가 0 인 경우에는 $f_{NN}(\mathrm{x}_i, \mathrm{y}_j)$를 계산할 필요가 없기 때문에 유사도가 낮은 SIFT 가 검색 대상에서 빠르게 제외되는 효과가 있다고 볼 수 있다.

$$\mathrm{PMF}(I_i, I_j) = \begin{cases} \text{true}, & \text{if } \sum_{\forall \mathrm{x}_i, \forall \mathrm{y}_j} \delta(\mathrm{x}_i, \mathrm{y}_j) \times f_{NN}(\mathrm{x}_i, \mathrm{y}_j) > \gamma \\ \text{false}, & \text{if } ColorSimilarity(I_i, I_j) < \theta \end{cases} \quad (7)$$

C. 후-검증을 통한 오-검출 감소 방법

$f_{NN}(\mathrm{x}_i, \mathrm{y}_j)$의 최근접 검색은 다른 배경에 대하여 유사한 로컬 시각 특징을 갖는 경우나 전경에 유사한 객체가 있을 경우 오-매칭이 빈번히 발생한다. 본 논문에서는 오-매칭을 줄이기 위한 후-검증 방법으로[4]에서 제안한 방법인 WGC(Weak Geometrical Consistency)를 적용하였다. WGC 는 SIFT 의 오리엔트의 차이 값 히스토그램을 구하고, 매칭된 두 SIFT 간의 오리엔트의 차이가 히스토그램의 피크 값에 해당하는 경우만 매칭된 것으로 판단한다. <그림 3>은 WGC 를 통해 오-매칭을 줄이는 예시이다.



(a) WGC 적용 전 (b) WGC 적용 후

<그림 3> WGC 를 통한 오-매칭 감소 예시

3. 장소 기반의 씬 그룹화

동영상은 의미 단위인 씬으로 나뉘며, 씬 $S = \{I_0, I_1, I_2, \dots\}$ 와 같이 키 프레임들의 집합으로

표현된다. 두 씬의 장소 중복성을 검출하는 함수는 식(8)과 같다.

$$\mathrm{PMS}(S_i, S_j) = \begin{cases} \text{true, if } \exists I_m, I_n,\ I_m \in S_i, I_n \in S_j, \mathrm{PMF}(I_m, I_n) = true \\ false,\ otherwise \end{cases} \quad (8)$$

같은 장소에서 촬영된 씬 들은 완전한 그래프를 구성하지 않을 수 있다. 즉, 동일한 장소의 씬 S_i, S_j, S_k 에 대하여 $\mathrm{PMS}(S_i, S_j) = true$, $\mathrm{PMS}(S_j, S_k) = true$ 이더라도 $\mathrm{PMS}(S_i, S_k) = false$ 인 경우가 발생할 수 있다. 같은 장소에서 촬영된 씬들의 그룹 $\mathbb{P}$는 식(9)와 같이 정의된다.

$$\mathbb{P} = \{S_0, S_1, S_2, \dots\}$$
$$\text{for } \forall S_i \in \mathbb{P},\ \exists S_j \in \mathbb{P} \text{ where } \mathrm{PMS}(S_i, S_j) = \text{true} \quad (9)$$

4. 실험 및 분석

A. 실험 환경

실험용 데이터 셋은 약 1시간 분량의 8회로 구성된 네 편의 드라마를 사용하였다. 각 동영상은 씬 단위로 분할하였고, 수동으로 장소 태깅을 하였다. [표1]은 시리즈 콘텐츠에서 장소의 중복성이 존재함을 보여준다.

[표1] 드라마의 장소 개수 및 장소 당 평균 씬 수

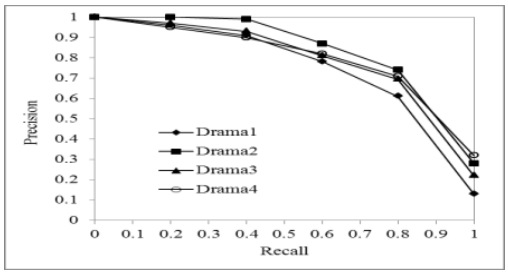
	드라마 1	드라마 2	드라마 3	드라마 4
장소 수	17	13	13	13
장소 당 평균 씬 수	5.35	6.76	9.00	8.38

B. 정확도

정확도는 식(10)과 같이 함수 PMS를 통하여 그룹화 된 $\forall\mathbb{P}$에 대하여 정확률과 재현률을 통해 측정하였다.

$$Precision = \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN} \quad (10)$$

TP(true positive)는 $\mathbb{P}$에 속하는 씬들 중 옳게 그룹화된 씬의 개수를 뜻하며, FN(false negative)는 같은 장소이면서 그룹화 되지 못한 씬의 개수를 뜻하고, FP(false positive)는 다른 장소인데도 $\mathbb{P}$에 그룹화 된 씬의 수를 뜻한다. 측정 결과는 <그림 4>와 같다.



<그림 4> 드라마 별 장소 그룹화

<그림 4>에서 실내 촬영이 많은 드라마 2 의 경우 높은 정확도가 나왔음을 확인할 수 있었다.

C. 연산 속도

[표 2]는 각 씬들 간의 PMS 연산에 대한 평균 시간으로, 색상 정보를 이용한 후보군 선별 과정과 벡터 양자화를 통한 최근접 검색 공간 감소 과정의 효과를 나타낸다. 두 가속화 방법을 적용하여 약 0.4 배 연산량이 감소한 것을 보이고 있다.

[표 2] PMS 연산에 대한 평균 처리시간(sec)

SIFT 매칭	벡터 양자화 + SIFT 매칭	색상 정보 + 벡터 양자화 + SIFT 매칭
44.876	3.376	0.181

5. 결론

본 논문에서는 중복되는 장소를 그룹화 하는 방법에 대해 연구하였다. 촬영된 장소는 부분적인 중복이 많아 SIFT 매칭을 사용하였다. 이는 많은 연산이 필요하므로 매칭의 후보군과 최근접 계산량을 줄이는 방법을 진행하였다. 이로 인해 생기는 에러를 해결하기 위해 후-검증방법을 사용하였다. 실험 결과 0.8 의 정확률과 0.6 의 재현률이 도출되었고, SIFT 대비 0.4 배 연산량을 줄일 수 있었다. 향후 연구로, 색상정보 양자화를 통해 연산량은 줄어들었으나, 조명변화에 강건하지 못한 연구가 필요하다.

참고문헌

[1] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision,* Vol.60, pp 91-110, 2004.

[2] J. Sivic, and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," *in Proc. of the IEEE International Conference on Computer Vision,* Vol.2, pp 1470-1477, 2003.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object Retrieval with Large Vocabularies and Fast Spatial Matching," *in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition,* pp 1-8, 2007.

[4] H. Jégou, M. Douze, and C. Schmid, "Improving Bag-of-Features for Large Scale Image Search," *International Journal of Computer Vision,* Vol. 87, pp 316-336, 2010.

[5] W. Zhao, X. Wu, and C. Ngo, "On the Annotation of Web Videos by Efficient Near Duplicate Search," *IEEE Transactions on Multimedia,* Vol.12, pp 448-461, 2010.

[6] T. Ojala, M. Aittola, and E. matinmikko, "Empirical Evaluation of MPEG-7 XM Color Descriptors in Content-Based Retrieval of Semantic Image Categories," *in Proc. of the International Conference on Pattern Recognition,* Vol. 2, pp 1021-1024, 2002.