감정 정보를 활용한 개인 사진 컬렉션의 이벤트 분류 및 대표사진 선정 모델

범정현, 염상길, 황지영, 추현승
성균관대학교 소프트웨어대학
Email: {bumjh, sanggil12, jjwhang, choo}@skku.edu

Event segmentation and key photo selection of personal photo collection with emotion information

Junghyun Bum, Sanggil Yeom, Joyce Jiyoung Whang, and Hyunseung Choo
College of Software, Sungkyunkwan University

요 약

스마트폰과 소셜미디어의 영향으로 개인이 일상적으로 소유하는 이미지 데이터의 규모는 날로 증가하고 있다. 이는 개인 사진 컬렉션을 자동으로 구성하고 검색하도록 돕는 요약 도구를 필요로 한다. 요약을 위한 일반적인 방법은 먼저 이벤트를 구분하고 각 이벤트내에서 대표 사진을 선택하는 것이다. 이벤트 경계를 구분하기 위한 속성의 하나로 이미지에서 얼굴을 인식하고 감정을 추출하여 사용한다. 콘텐츠-기반 세그멘테이션 기법에 인물의 감정요소를 포함시킨 확률론적 모델을 제안한다. 인물의 감정요소를 포함시킴으로써 이벤트 세그멘테이션 성능을 향상시키고 대표사진의 다양성을 높인다.

1. 서 론

스마트폰과 디지털 카메라의 일상적인 사용으로 이전보다 훨씬 많은 사진들이 손쉽게 찍히고 공유된다. 최근 개인의 라이프로그는 인스타그램, 페이스북 등 소셜미디어 상에 이미지로 기록된다[1]. 개인이 소유하고 보관하는 미디어 양이 급속히 증가하고 있다. 이러한 많은 양의 이미지 데이터는 수작업으로 분류, 검색 및 보관을 어렵게 만들고, 개인의 사진 컬렉션을 자동으로 구성하고 검색하도록 돕는 요약(Summarization) 도구를 필요로 하게 한다[2].

개인의 사진 및 비디오는 이벤트의 순차적인 활동의 기록물이다. 사진 및 비디오를 요약하는 효과적인 방법은 이를 이벤트로 분류한 다음 각 이벤트에서 주요 사진을 선택하는 것이다. 연속된 프레임으로 구성된 비디오 보다 개인의 사진 컬렉션에서 이벤트를 구분해 내는 것은 이벤트 경계를 불분명하게 만드는 노이즈 이미지들로 인해 더욱 어렵다. 기존 포토 컬렉션 요약 기법들은 주로 시간과 위치 정보를 활용하는 컨텍스트-기반 세그멘테이션과 이미지의 색상, 질감, 명도 등과 같은 콘텐츠-기반 세그멘테이션 방법을 사용하고 있다. 본 논문에서는 센티멘트 요소를 포함하는 클러스터링을 수행하는 이벤트 세그멘테이션 기법을 소개한다. 인물의 감정요소를 포함시킴으로써 이벤트 세그멘테이션 성능을 향상시킨다.

본 논문의 구성은 다음과 같다. 논문의 2장에서 관련 연구를 카테고리별로 소개하며 3장에서는 제안모델을 소개한다. 4장에서 실험 및 실험결과에 대해 설명하고 5장에서는 결론과 향후 연구에 대해서 서술한다.

2. 관련연구

2.1 이벤트 세그멘테이션

시간은 이벤트를 구분하는데 있어서 매우 유용한 정보이다. 특정 이벤트가 발생한 날에는 다량의 사진이 찍힌다. 일반적으로 동일한 하위이벤트의 사진은 비교적 짧은 시간안에 찍힌다. Cooper 등은 이벤트 감지를 위해, 사진의 타임스탬프간 유사성에 기반하여 이벤트를 구분한다[3]. 이벤트 경계는 높은 Intra-class 시간 유사성과 낮은 Inter-class 유사성으로 두개의 인접한 사진 그룹을 분리한다. Graham 등 은 시간에 따라 사진이 찍히는 패턴을 통해 클러스터를 생성하고 분할하여 이벤트를 구분하였다[4]. Naaman 등은 시간과 GPS 정보를 이용하여 이벤트 세그멘테이션을 수행한다[5]. 초기 세그먼트는 시간과 공간의 유사성에 기초하여 생성한 다음 유사한 공간에 속한 사진들끼리 클러스터링하여 최종 이벤트 구조를 만든다. Mei 등은 타임스탬프와 얼굴, 색상 등의 콘텐츠 그리고 카메라 설정을 포함하는 멀티모달 속성을 이용하여 확률론적 접근법을 제안한다[6]. Guo 등은 시간과 사물, 장면 등 Convolutional Neural Network (CNN)에서 훈련된 속성을 추출한 후 이벤트 분류자로 식별하는 계층적 모델을 제시한다[7].

2.2 대표 사진 선정

각각의 분할된 이벤트 내에서 대표성을 갖는 핵심사진을 선택하여 사진 컬렉션을 요약한다. Cooper 등은 대표 사진으로 각 이벤트에서 타임스탬프가 가장 빠른

사진을 추천한다[3]. Jun, L.등은 사진에서 검출된 얼굴의 면적과 시간 간격 기준을 적용하여 대표사진을 선정한다[8]. Mei 등은 현재 이벤트에서 최대 사전 확률(Maximum priori probability)을 가진 사진을 대표사진으로 선택한다[6]. Shen 등은 품질, 대표성 및 인기도 특징의 조합으로 순위 매기는 알고리즘을 제안한다. 명도, 색상, 대상영역의 크기, 대비 등 콘텐트 속성을 추출하여 AVA 데이터셋에서 훈련된 분류자로 미학적 품질을 평가하고, 대표성은 이벤트에 사진이 속할 확률 값을 사용하며, 인기도는 단일 이벤트내에 유사한 이미지 수가 많은 사진을 측정하여 평가한다[9].

3. 제안 Approach

시간 혹은 위치 정보의 누락, 연속된 사진 프레임 사이에 섞여있는 관련 없는 사진 등은 사진 컬렉션에서 이벤트(혹은 하위 이벤트)를 구분하는 것을 어렵게 만든다. 본 논문에서는 시간, 위치, 고수준 콘텐츠 속성 및 감정 속성을 추출하여 세그멘테이션하는 확률론적 모델을 제안한다.

3.1 사진 속성(features) 추출

시간, 위치, 색상 및 CNN과 감정의 멀티모달 속성이 사진을 표현하기 위해 사용된다. 사진을 촬영하는 순간 시간과 초점거리 등과 같은 메타데이터는 표준 이미지 헤더 EXIf(EXchangeable Image file) 포맷으로 저장된다. 메타데이터로부터 시간과 위치 정보를 사진의 콘텐트로부터 색상, 사물 및 감정 정보를 추출한다.

Time - EXIf 헤더로부터 타임스탬프를 추출한다.

Location - Exif 헤더의 GPS 정보로부터 위도와 경도를 추출한다.

Color - N 픽셀을 가진 이미지 x에 대해 RGB 색 공간에서 64차원의 컬러 히스토그램 $H(x) = [h_1, h_2, \ldots, h_{64}]$을 만든다.

CNN - 머신러닝 오픈소스 라이브러리인 TensorFlow를 활용하여 이미지에서 2048-차원의 속성을 추출하는 Convolutional Neural Network(CNN)를 구현한다[10]. Inception-v3 모듈은 120만개 이미지를 가지고 ILSVRC-2012 훈련데이타로 학습되었고 ImageNet 데이터베이스에서 1,000개 카테고리를 구별할 수 있다. Inception 모듈의 마지막 층인 Softmax layer를 제외하고 2048-차원의 속성을 구한다. 우리는 CNN 속성이 전체 이미지 특성을 결정하지 않도록 PCA를 적용하여 128-차원으로 축소한다.

Emotion - 마이크로소프트의 Congnitive 서비스는 안면 인식과 감정 인식을 위한 Face API, Emotion API 등을 제공한다[11]. Face API는 이미지에서 안면 사진을 감지하고 기계학습을 통해 동일인인지 여부를 신뢰도 점수로 반환한다. Emotion API는 사용자의 안면 사진을 분석하여 행복, 놀람, 두려움, 슬픔, 역겨움, 경멸, 화남, 무표정의 8개의 감정을 0~1 사이의 신뢰도 점수로 제공한다. 감정 속성을 추출하기에 앞서 사진 컬렉션에서 중요 인물의 얼굴을 등록하여 훈련시킨다. 4명의 인물에 대한 감정을 추출한다면 감정값은 32-차원(4명*8종)의 속성이 되고, 얼굴이 인식되지 않는 사진이라면 32-차원의 감정 속성은 모두 0이 된다.

3.2 이벤트 세그멘테이션

각 사진 $x_i \in X = \{ x_1, x_2, \ldots, x_N \}$은 하나의 알려지지 않은 잠재 클래스 - 이벤트 $e_j \in E = \{e_1, e_2, \ldots, e_K\}$ 에 속한다. N은 사진의 수이고 K는 이벤트의 수이다. 사진 x_i가 이벤트 e_j로부터 생성될 확률은 $p(x_i|e_j)$로 나타낼 수 있다.

사진을 시간(T), 위치(L), 색상(C), 사물(O), 감정(E)의 5개의 속성 벡터로 표현한다. 사진의 모든 속성은 서로 독립적이라고 가정하면 식(1)과 같이 가우시안 혼합 모델로 정의할 수 있다. 동일한 이벤트에 있는 사진은 멀티모달 속성 공간에서 동일한 분포를 공유한다.

$$p(x_i|e_j) = p\,(T_i|e_j)p(L_i|e_j)p(C_i|e_j)p(O_i|e_j)p(E_i|e_j) = \prod_{l=1}^{L} p(x_{i,l}|e_j) \quad (1)$$

여기서 $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,L})$ 이다. $x_{i,l}$ 은 사진 x_i의 ℓ번째 메타데이타이고 $\mathrm{L} = 5$인 속성벡터의 수이다. 각 $x_{i,l}$요소는 가우시안 분포에 의해 생성된다.

가우시안 분포의 파라미터는 최대우도법(Maximum Likelihood method)에 의해 추정될 수 있다. 잠재변수인 이벤트에 대해 로그가능도는 식(2)와 같이 공식화할 수 있고, 기댓값-최대화 (Expectation-Maximization) 알고리즘을 적용하여 훈련시킨다.

$$\ell(\mathrm{X};\,\theta) \triangleq \log\Big(\prod_{l=1}^{N} p(x_i|\theta)\Big) = \sum_{i=1}^{N} \log\Big(\sum_{j=1}^{K} p(e_j)p(x_i|e_j,\theta)\Big) \quad (2)$$

EM 알고리즘으로 파라미터를 최적화하기에 앞서 이벤트 수를 미리 정의하여야 한다. 우리는 문제를 단순화하기 위해 rule of thumb에 따라 식(3)과 같이 이벤트 수 k를 구한다[12]. 클러스터 개수는 입력된 오브젝트의 절반 값의 제곱근으로 구할 수 있다.

$$k = round\left(\sqrt{N/2}\right)\ (N = Number\ of\ photos) \quad (3)$$

파라미터 초기 값은 K-means 알고리즘을 사용하여 계산한다. 그런 다음 E-step에서 최대가능도가 계산된다. M-step 단계에서 이벤트 e_j의 사후확률을 계산하고 로그가능도 함수를 최대화하는 새로운 파라미터 값을 선정한다. E-step과 M-step을 수렴조건을 만족할 때까지 반복한다.

3.3 대표 사진 선정

K개 이벤트 클러스터 중 각 이벤트 e_j 에서 가장 높

은 최대 확률 $p(x_i|e_j)$를 갖는 사진 x_i는 해당 이벤트의 대표사진으로 간주될 수 있다. 즉, 가우시안 분포의 중심에 해당하는 사진을 대표사진으로 선정한다.

4. 실험 및 결과

4.1 실험 데이타셋

제안한 이벤트 세그멘테이션 기법의 유용성을 실험하기 위해 6명의 사용자로부터 11개의 데이터셋(3,680장의 사진)을 수집하였다.

표 1. 이벤트 세그멘테이션 실험 데이터셋

Person	Dataset	Photos	Key Photos
User1	1	498	16
User2	3	1541	48
User3	1	152	9
User4	2	754	26
User5	2	614	25
User6	1	121	8

4.2 사진 속성 추출

시간과 위치 값은 EXIf헤더로부터 추출된다. 저장된 메타데이타에 시간과 위치 값이 0으로 기록된 데이터가 상당 수 존재하여 보정이 필요하다. 시간 및 위치 값은 사진이 생성된 순서에 가장 가까운 데이터와 정렬하여 조정한다. 클러스터링을 위해 시간 및 위치 값은 0~1사이의 값으로 Normalize한다. CNN은 TensorFlow의 Image Recognition 오픈소스 라이브러리를 활용하고 및 Emotion 값은 마이크로소프트의 Emotion API를 활용하여 추출한다.

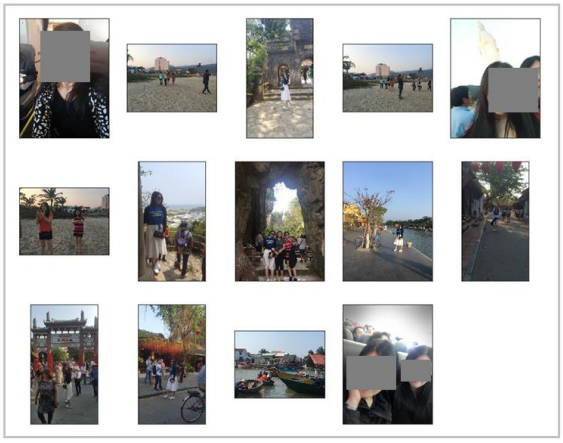

그림 1. User1의 대표 사진 선정 예

4.3 이벤트 세그멘테이션 및 대표사진 선정

이벤트 세그멘테이션을 위해 scikit-learn 오픈소스 라이브러리를 활용하여 가우시안 혼합 모델을 구현한다. 감정 속성이 이벤트 경계간 식별의 성능을 향상시키는지 확인하기 위해 ①시간과 위치, ②시간, 위치 및 색상, ③시간, 위치, 색상 및 사물, ④시간, 위치, 색상, 사물 및 감정의 4가지 경우에 대해 클러스터링하고 이벤트 세그멘테이션 결과를 비교 평가한다. 그림 1은 제안 모델로 이벤트 세그멘테이션 후 이벤트에서 최대확률을 갖는 사진을 대표사진으로 선정한 결과를 보여준다.

5. 결론 및 향후 연구

개인의 사진 컬렉션을 자동으로 구성하고 요약하는데 있어 감정 요소를 속성화하여 이벤트 세그멘테이션 성능을 향상시키는데 적용하였다. 향후 Key 이미지 선정을 위한 알고리즘 개발과 User study를 통해 보다 객관적인 평가를 진행할 것이다.

ACKNOWLEDGEMENT

본 논문은 기초연구사업 (NRF-2010-0020210)과 과학기술정보통신부 및 정보통신기술진흥센터의 Grand ICT연구센터지원사업 (IITP-2017-2015-0-00742), 과학기술정보통신부 및 정보통신기술진흥센터 (2014-0-00547, 자율 제어 네트워킹 및 자율 관리 핵심 기술 관리)의 연구결과로 수행되었음

참고문헌

[1] Lee, Eunji, et al. "Pictures speak louder than words: Motivations for using Instagram," Cyberpsychology, Behavior, and Social Networking, 18(9), 552-556, 2015.

[2] Ceroni, A., et al. "Investigating human behaviors in selecting personal photos to preserve memories," ICMEW, 1-6,2015.

[3] Cooper, M., et al. "Temporal event clustering for digital photo collections," ACM TOMM, 1(3), 269-288, 2005.

[4] Graham, A., et al. "Time as essence for photo browsing through personal digital libraries," Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, 2002.

[5] Naaman, Mor, et al. "Automatic organization for digital photographs with geographic coordinates," Proceedings of the 2004 Joint ACM/IEEE Conference on. IEEE, 2004.

[6] Mei, Tao, et al. "Probabilistic multimodality fusion for event based home photo clustering," Multimedia and Expo, 2006 IEEE International Conference on. 1757-1760, 2006.

[7] Guo, Cong, and Xinmei Tian. "Event recognition in personal photo collections using hierarchical model and multiple features." *MMSP*, 1-6, 2015.

[8] Jun, L., et al. " Automatic summarization for personal digital photos," IEEE Fourth International Conference on Information, Communications and Signal Processing, 3, 1536-1540, 2003.

[9] Shen, X. and X. Tian "Multi-modal and multi-scale photo collection summarization." Multimedia Tools and Applications 75(5): 2527-2541, 2016.

[10] "TensorFlow," https://www.tensorflow.org/

[11] "Microsoft Cognitive Services," https://www.microsoft.com/ cognitive-services/

[12] Mardia, K. V., J. T. Kent, and J. M. Bibby, "Multivariate Analysis," Academic Press Inc. London LTD, 1979