동영상 요약 시퀀스 생성을 위한 하이브리드 유사 프레임 비교 기법

옥창석, 권대건, 조환규
부산대학교 컴퓨터공학과
e-mail:csock@pusan.ac.kr, duskan@pusan.ac.kr, hgcho@pusan.ac.kr

A Hybrid Comparing Method of a Similar Frame for Generating Video Summarization Sequences

Chang-Seok Ock, Dae-Gun Kwon, Hwan-Gue Cho
Dept of Computer Science & Engineering, Pusan National University

요 약

멀티미디어의 규모가 급격하게 늘어나고 있는 현재, 영화와 같은 동영상은 용량에 있어 사진과 비교했을 때 상당한 크기를 가지고 있고 그만큼 많은 정보를 담고 있다. 이렇게 많은 정보를 얻기 위해 사용자들은 많은 시간을 소비해야 한다. 이러한 비효율적인 측면의 보완을 위해 동영상의 각 프레임의 유사도를 판단하여 유사한 프레임들은 하나로 모으고, 유사하지 않은 프레임들은 구분하여 요약된 시퀀스로 보여줄 수 있는 방법이 필요하다. 이러한 관점에서 봤을 때 동영상은 시간적 순서에 따라 프레임이 배열되어 있고 인근 프레임 간에는 Coherence가 존재한다는 장점이 있다. 따라서 우리는 이러한 장점을 최대한 이용하여 동영상의 요약 시퀀스를 생성하기 위해 일차적으로 필요한 유사 프레임을 비교할 수 있는 기법을 제안한다. 제안하는 기법은 각 프레임의 공간적인 정보를 활용 할 수 있는 특징점 기반의 기법과, 각 프레임의 색 분포 정보를 활용 할 수 있는 히스토그램 기반의 기법을 Hybrid하게 적용하여 유사 프레임을 판단한다. 제안한 기법을 통해 도출한 결과를 통계학적으로 검증을 위해 널리 사용되는 Precision과 Recall을 이용하여 검증한다.

1. 서론

요즘 많은 사람들이 영화에 대해 관심을 가지고, 영화를 하나의 여가활동으로 여기고 있다. 여기서 시중에 나와 있는 수많은 영화에 대해 얻을 수 정보는 인터넷 검색을 통한 평점과 요약된 줄거리 정도이다. 이러한 한정된 정보는 관람객으로 하여금 영화의 본질에 대한 충분한 정보가 되지 못한다. 단순히 텍스트로 된 줄거리보다는 이미지로 된 일련의 사진들이 조금 더 관람객에게 쉽게 다가 갈 수 있다. 시각적인 정보를 최대한 활용할 수 있도록 영화의 주된 내용을 이미지화 하여 시간적순서대로 배열하면 관람객의 입장에서는 충분한 정보제공이 될 것이다. 이러한 접근법은 영화의 포스터와는 또 다른 느낌으로, 하이라이트 부분만을 이어붙인 포스터는 그 내용전달에 있어서는 의미가 있으나 사건의 시간적 순서와 같은 유기적인 정보는 전달하지 못한다. 이러한 성질은 영화 뿐 아니라 드라마, 뮤직비디오 등의 일반적인 동영상에도 마찬가지로 적용된다.

무엇보다도 동영상은 내부적으로 시간적인 순서에 맞춰 프레임이 구성되어 있으며 각 프레임들은 카메라의 전환에 의한 화면의 변화가 있기 전 까지는 유사한 영상으로 이루어져 있다는 장점을 가지고 있다. 그리고 각 프레임간에는 공간적 응집도(Spatial-Coherency)가 있다는 장점이 있다. 이러한 장점들은 영화의 요약 이미지를 자동으로 생성하기에 적합한 특징이라고 할 수 있다. 따라서 우리는 동영상에서 이러한 카메라의 전환점을 추정하기 위해 이러한 특징들을 이용할 수 있는 유사 프레임 판정기법을 제안한다. 이 기법을 통해 프레임을 추출, 시퀀스화 하여 영화의 요약된 정보를 아날로그 필름형식으로 만드는 것을 최종 목표로 한다.

이 기법의 기본적인 틀은 SURF(Speeded Up Robust Features)[7]와 RANSAC(RANdom SAmple Consensus)[8]의 조합으로 이루어진다. 이는 특징점 기반의 유사도 비교 기법과 히스토그램을 이용한 유사도 비교 기법의 조합으로 두 기법이 Hybrid 형식으로 적용되어 유사 프레임을 측정한다.

본 실험에서는 동영상의 요약시퀀스를 생성하기 전 핵심부분인 유사 프레임을 비교하고 유사도를 판단하는 기법을 제안하고 그 결과에 대해 Precision과 Recall을 이용하여 검증한다.

2. 관련 연구

유사한 이미지를 비교하는 분야에서는 이미 아주 많은 연구들이 있었고, 현재에도 상당히 많은 연구들이 활발히 진행중에 있다. 세부 분야별로는 각 이미지 내부의 콘텐츠(Object)기반의 유사도를 비교하는 분야와, 시간적인 순서로 유사도를 비교하는 분야, 각각의 색 분포를 비교하는 분

야 등이 있다. 물론 이러한 각 분야를 아우르는 Hybrid한 기법들도 많이 나와 있다. 이러한 연구의 핵심은 얼마나 유연하게, 얼마나 정확하게 두 이미지의 유사도를 판정하는 가에 있다. 여기서 이미지의 유사도 판정을 위해 사용한 각 연구의 알고리즘이 그 결과를 도출하는데 큰 몫을 한다.

이미지 유사도 판정과 관련된 연구는 아주 많기 때문에 그 중 대표적인 연구들만 표 1에 정리하였다.

<표 1> 이미지 유사도 판정 알고리즘 정리.

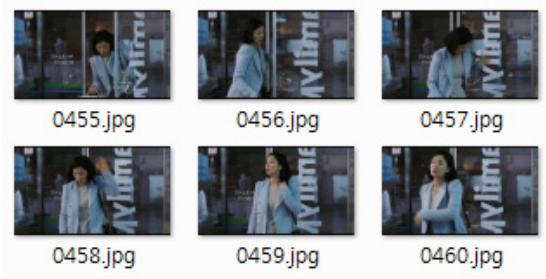
저자	방법
M. Cooper외[1]	Temporal event 기반
J. Goldberger외[2]	MoG간의 KL-Divergence 이용
R. Baeza-Yates외[3]	Graph Matching 이용
Z. Wang외[4]	Wavelet 이용
K.H. Kim외[5]	Human-Recognition Color 이용

Cooper[1]의 방법은 시간적인 이벤트에 기반하여 사진들을 클러스터링 하는 방법을 소개하고 있다. 클러스터링을 하기 위해서는 사진의 유사도를 측정해야 하는데 이 때 Temporal Event를 활용한다. Goldberger[2]는 각 이미지를 MoG (Mixture of Gaussians)를 이용하여 모델링 한 후 두 이미지의 KL (Kullback-Liebler)-Divergence를 Approximation하여 유사도를 측정한다. Baeza-Yates[3]의 방법은 이미지를 평면 그래프로 바꾸어 그래프기반의 유사도를 측정한다. Wang[4]의 방법은 이미지에 복잡한 Wavelet을 적용하여 왜곡을 준 후 SSIM(Spatial domain structural similarity) index알고리즘을 이용하여 유사도를 측정한다. Kim[5]의 방법은 사람의 색인지 능력에 준하는 색상모델을 이용하여 이미지의 히스토그램을 비교하는 방법으로 유사도를 측정한다.

3. 동영상 프레임 추출

큰 용량의 동영상을 매 프레임마다 샘플링하여 유사도를 측정하는 것은 아주 비효율적이다. 대체로 동영상의 FPS(Frames Per Second)는 30내외로 1초에 30프레임정도가 재생된다. 하지만 이 1초내에 실제 영상에 담기는 Spatial 정보의 변화는 아주 작다. 따라서 우리는 실험에 사용하는 동영상의 Sampling Rate를 1초 간격으로 하여 매 1초마다의 프레임을 추출하여 그 유사도를 측정한다.

그림 1은 6초간의 실제 동영상에서의 프레임변화를 나타낸 것이다.



(그림 1) 6초간의 프레임 변화

그림 1과 같이 일반적인 동영상에서 아주 빠른 내용전개를 위한 몇몇 부분을 제외하면 1초라는 시간 안에 많은 내용이 바뀌지 않는다. 이러한 성질을 기반으로 실험용 영화에서 프레임을 추출한다.

실험에 사용하는 영화는 정상적인 루트로 구입하여 보관중인 이정향감독, 송혜교 주연의 영화 '오늘'이다[6]. 연구를 위해 영화의 일부분을 사용하는데 있어 여기에 출처를 밝힌다.

4. 유사도 측정 방법

우리가 제안하는 두 프레임의 유사도를 측정하는 방법은 특징점 기반의 유사도 판정기법과 Gray-Level 히스토그램기반의 유사도 판정기법을 결합한 Hybrid 모델을 사용한다. 최종적인 유사도는 각 방법의 유사도에 비중을 조절하고 합하여 결정한다.

유사도를 측정하기 위한 방법은 먼저 SURF와 RANSAC의 조합을 이용하는 것이다. 이 두 알고리즘은 영상처리와 이미지 프로세싱분야에서 널리 쓰이고 있다.

SURF는 잘 알려진 알고리즘인 SIFT(Lowe, 1999)[9]와 같은 맥락이지만 내부 알고리즘의 간략화와 적분이미지사용을 통해 성능을 크게 향상시킨 것으로 Robust한 특징점을 찾는 알고리즘이다. Herbert Bay에 의해 처음 제안되어 Object recognition이나 3D Reconstruction분야에서 널리 사용되고 있다. 그리고 RANSAC은 1981년도에 Fischler와 Bolles에 의해 처음 퍼블리싱 되었으며 Line fitting, Parameter Estimation분야에서 많이 쓰이고 있다.

본 실험에서는 SURF를 이용하여 비교대상이 되는 두 프레임의 특징점을 빠르게 찾아낸 후 매치하여 Pair로 만든다. 이렇게 생성된 결과를 RANSAC을 이용하여 Outlier를 제거하는 과정을 거쳐 최종적으로 남아있는 Pair의 비율을 계산하여 두 프레임의 유사도 측정기준으로 사용한다.

수식 (1)에서 $Pair_o$는 SURF를 이용하여 찾아낸 초기의 Pair의 수를 나타내고, $Pair_r$은 RANSAC을 이용하여 $Pair_o$에서 Outlier를 제거하고 남은 Inlier Pair의 수를 나타낸다. 이 때 계산된 유사도를 R_f라고 한다.

$$R_f = \frac{Pair_r}{Pair_o} \qquad (1)$$

그 다음 방법은 두 프레임의 Gray-Level Histogram을 비교하는 것이다. 각 프레임을 255-Gray Level Histogram으로 나타내고 그 유사도를 판단한다. 이렇게 측정된 Histogram 유사도를 R_h라고 한다.

이렇게 정의된 두 R_f와 R_h는 최종 유사도 R을 판정하기 위해 Weight(w, $0 \le w \le 1$)를 주어 수식(2)와 같이 나타낼 수 있다.

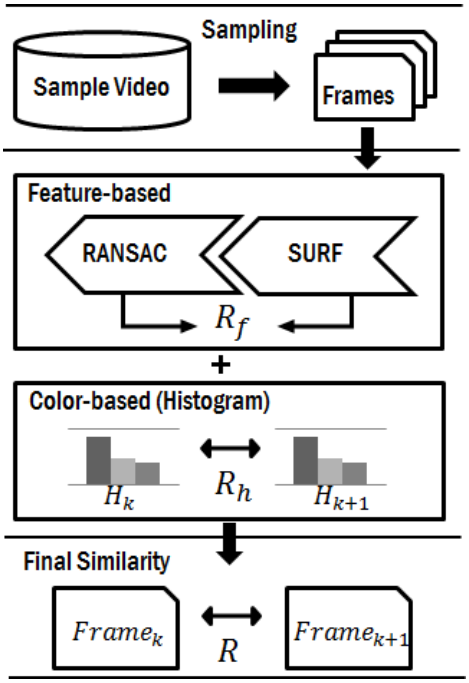
$$R = wR_f + (1-w)R_h \qquad (2)$$

최종 유사도 R을 계산하고 나면 이 값을 이용하여 두 프레임이 유사한지를 판정한다. 이 때 Threshold를 주어 Threshold보다 R이 크다면 두 프레임은 같다고 판정하고 작다면 두 프레임은 다르다고 판정한다.

5. 실험 및 결과

실험을 진행하기에 앞서 실제 실험에 쓰이는 여러 파라미터 중 프레임 유사도 R을 판정하기 위해 사용되는 Threshold를 제외한 나머지 파라미터는 모두 고정된 상태에서 진행한다.

실험의 각 프로세스 단계를 그림으로 나타내면 그림 2와 같다.



(그림 2) 실험 프로세스 개요도

동영상 샘플링을 수행한 후 추출된 1초간격의 프레임들을 제안한 유사 프레임 측정 기법을 사용하여 각 프레임 간의 유사도를 측정한다. 측정된 유사도를 기반으로 Threshold를 적용시켜 유사하다고 판정된 두 프레임을 묶는 방법으로 진행한다.

실험에 사용한 프레임 데이터는 표 2와 같다.

<표 2> 실험에 사용한 데이터

종류	총 프레임 수	샘플 프레임 수	샘플링 주기
영화 "오늘" [6]	215,546	7,184 (사용 7,166)	1초

표 2를 보면 영화의 재생시간은 약 2시간이며 총 프레임 수는 215,546개 (FPS 30)이다. 여기서 1초단위로 프레임을 추출하면 7,184개의 프레임이 추출된다. 여기에는 영화의 인트로(제작사 로고)와 아웃트로(크레딧)이 포함되어 있다. 본 유사 프레임 추출의 경우 인트로와 아웃트로에 대한 정보는 생략하고 영화의 본 내용에 대한 유사 프레임 추출에 대한 실험을 진행한다. 인트로와 아웃트로를 제외하면 총 프레임 수는 7,166개가 된다.

유사도를 판정하기 위해 추출한 샘플을 1000개씩 7개의 그룹으로 나누어 실험한 후 최종적인 유사도 판정 결과를 도출한다. 이 때 발생하는 각 그룹 경계부분의 이미지의 유사도는 최대 그룹의 개수인 7개의 오류가 나타날 수 있지만 전체 실험 데이터(7,166개)에 비해 적은 수 이므로 무시한다.

그룹으로 나누어진 프레임들에 대해 제안한 유사도 판정 알고리즘을 적용하기 위해 모든 프레임들을 크기 2의 Window를 이용하여 순회한다. 즉, 총 프레임 수가 N일 때, 크기 2의 Window를 1칸씩 Shift하며 N-1번의 유사도 판정을 수행한다. 모든 프레임을 N:N으로 비교하지 않고도 이러한 방법으로 유사도 판정이 가능한 이유는 앞서 설명했듯이 영화 등의 동영상은 일반적으로 시간적 순서에 따라 배열되어 있고 공간적 응집도(Spatial-Coherency)가 있기 때문이다.

본 논문에서 제안하는 알고리즘을 샘플링한 프레임에 대해 적용한 후 그 결과를 검증하기 위해 Precision과 Recall을 사용하였다. Precision은 시스템이 유사하다고 판정한 샘플들 중 실제 유사한 샘플의 비율을 나타내는 지표이며, Recall은 시스템에서 유사하거나 유사하지 않다고 판정한 샘플들 중 실제 유사한 샘플의 비율을 나타내는 지표이다.

Precision과 Recall은 각 수식(3), 수식(4)와 같이 나타낸다.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

여기서 True Positive, False Negative와 같은 인자는 표 3과 같이 정의한다.

<표 3> TP, FN, FP, TN의 정의

구분	정답 집합	결과 집합
True Positive(TP)	T	T
False Negative(FN)	T	N
False Positive(FP)	N	T
True Negative(TN)	N	N

정답 집합은 사람에 의해 미리 유사한 프레임들을 구분한 집합이고, 결과 집합은 시스템에 의해 생성된 실험결과이다. T는 비교 대상이 되는 두 프레임이 유사하다고 판정한 것이고, F는 두 프레임이 유사하지 않다고 판정한 것을 의미한다. 즉, TP(True Positive)의 경우 제대로 유

사도를 판정한 것이고, FN(False Negative)의 경우 실제로는 두 프레임이 유사한데 시스템에서는 유사하지 않다고 판정한 것이다. 그리고 FP(False Positive)는 실제로는 유사하지 않은데 시스템에서는 유사하다고 판정한 것이고, TN(True Negative)는 실제로 유사하지 않은 것을 시스템에서도 유사하지 않다고 판정한 것이다.

본 실험의 결과는 표 4와 같다.

<표 4> 그룹별 Precision과 Recall

그룹(프레임범위)	프레임 수	Precision	Recall
G1(1～1000)	1,000	0.9770	0.9749
G2(1001～2000)	1,000	0.9839	0.9797
G3(2001～3000)	1,000	0.9775	0.9881
G4(3001～4000)	1,000	0.9801	0.9823
G5(4001～5000)	1,000	0.9600	0.9686
G6(5001～6000)	1,000	0.9812	0.9747
G7(6001～7166)	1,166	0.9785	0.9560
계, 평균	7,166	0.9768	0.9749

결과를 보면 7개의 그룹에 대해 평균적으로 0.97정도의 Precision과 Recall을 보여주고 있다. 이 결과를 바탕으로 Object기반의 유사도를 측정하기 유리한 특징점기반의 유사도 측정 기법과 영상의 전반적인 분위기의 측정에 유리한 히스토그램기반의 유사도측정방법을 Hybrid하게 조합하여 영상의 유사도를 비교하는 방법이 동영상과 같은 특징을 가지는 프레임의 유사도를 측정하는데 좋다는 것을 알 수 있다.

6. 결론

실험을 통해 도출한 결과를 바탕으로 우리가 제안하는 유사도 비교 방법의 적합성을 확인할 수 있다. 앞서 소개한 동영상이 가진 여러 가지 장점들을 이용하여 특징점기반의 유사도 비교방법으로 유사도를 측정하고 특징점기반으로는 찾아내기 힘든 부분을 히스토그램기반으로 유사도를 측정하는 Hybrid형식의 유사도 측정기법을 통해 두 기법 상호간의 단점을 보완하였다. 실제 특징점으로는 Object기반으로 유사도를 판단하기 때문에 노이즈에 상당히 민감하다고 할 수 있다. 이러한 부분을 노이즈에 비교적 강한 히스토그램을 이용하여 보완하였다.

본 실험의 최종적인 목표는 동영상의 요약 시퀀스를 생성하는 것으로 핵심부분인 유사도 측정 부분에 대한 실험만을 다루었다. 추후 연구를 통해 유사도 측정의 세밀한 분석과 정확도 향상을 통해 우수한 성능의 동영상 요약 시퀀스 생성 시스템을 구현할 것이다.

참고문헌

[1] Cooper M., Foote J., Girgensohn A., and Wilcox L., “Temporal Event Clustering for Digital Photo Collections,” J. ACM Trans. Multimedia Comp. Commu. and Appl. (TOMCCAP), vol.1, no.3, pp.269-288, 2005.

[2] Goldberger J., Gordon S., and Greenspan H., “An Efficient Image Similarity Measure Based on Approximations of KL-Divergence Between Two Gaussian Mixtures,” Computer Vision, IEEE International Conf., vol.1, pp.487-493, 2003.

[3] Baeza-Yates, R., and Valiente G., “An Image Similarity Measure based on Graph Matching,” String Processing and Info., pp.28-38, 2000.

[4] Wang, Z., and Simoncelli, E.P., “Translation Insensitive Image Similarity in Complex Wavelet Domain,” Acoustics, Speech, and Signal Proce. (ICASSP), pp.573-576, 2005.

[5] Kim, K.H., Kim, S.H., and Cho, H.G, “A Fast Summarization Method for Smartphone Photos Using Human-Percaption Based Color Model,” Multimedia, Computer Graphics and Broadcasting (MulGraB), pp.98-105, 2011.

[6] 감독 이정향, 주연 송혜교, 영화 “오늘”, 개봉 2011년 10월 27일, 다운로드 “네이버 영화 다운로드”, 2011.

[7] Bay, H., Tuytelaars, T., and Gool, L.V., “SURF :Speeded Up Robust Features,” Lecture Notes in Computer Science, vol.3951/2006, pp.404-417, 2006.

[8] Fischler, M.A., and Bolles, R.C., “Random Sample Consensus: A Paradigm for Model Fitting with Appli.,” Comm. ACM, vol.24(6), pp.381-395, 1981.

[9] Lowe, D.G. “Object recognition from local scale-invariant features,” ICCV’99, vol.2, pp.1150-1157, 1999.