다중레벨(Multi-Level) 분할 매칭을 이용한 뮤직비디오 자동 생성

윤종철° 이인권
연세대학교 컴퓨터과학과
media19@cs.yonsei.ac.kr°, iklee@yonsei.ac.kr

Automatic Music Video Generation using the multi-level temporal segment matching

Jongchul Yoon°, In-kwon Lee
Dept. of Computer Science, School of Engineering, Yonsei University

요 약

뮤직 비디오란 주어진 음악과 비디오가 동기화 된 형태의 창작물을 뜻한다. 기존의 뮤직비디오 제작방식에서는 만들어진 음악을 위해 영상 촬영에 전문적인 촬영 기술을 요구하였다. 본 논문에선 보다 쉬운 뮤직비디오 생성을 위하여 비디오와 음악의 특성을 분석하여 자동적인 뮤직비디오 생성시스템을 소개한다. 두 개체의 연속성을 보장하는 비교를 위해 우리는 각각의 객체의 흐름을 분석하고, 흐름의 유사성을 기준으로 분할하는 기법을 제시한다. 분할된 영상과 음악의 특성 비교를 통한 최적화된 매칭기법 을 비롯하여, 보다 다양한 조각 생성을 위한 다중 레벨(multi-level)분할 기반의 매칭 기법을 소개한다. 본 논문의 기술을 사용하여, 일반인이 홈비디오 등을 사용하여 손쉽게 뮤직비디오를 제작할 수 있다.

1. 서 론

기존의 뮤직비디오 제작방식은 미리 만들어진 음악에 동기화되기 위한 영상 촬영을 위하여 전문적인 촬영 기술을 요구하였다. 음악과 영상의 분리된 작업환경의 제약 때문에 수많은 시행착오를 요구하였고, 일반적인 홈비디오 사용자에겐 동기화된 뮤직비디오의 제작이 쉽지 않은 일이다. 따라서 본 논문은 전문적인 기술이 없는 일반 홈비디오 사용자가 손쉽게 영상과 음악을 자동적으로 연계시킬 수 있는 시스템을 소개한다.

본 논문의 목적은 비디오와 음악의 매칭을 통해 주어진 음악에 맞는 최적화된 비디오 클립을 추출하는 뮤직 비디오 자동 생성기법의 제시이다. 연속성을 보장하는 음악과 비디오의 매칭을 위해서 우리는 자동분할을 기준으로 한 매칭방법을 제시한다. 촬영자에게 있어서 하나의 비디오 샷 또는 클립은 특정한 연속적인 정보전달을 위해서 제작된다. 즉 비디오에 함축된 정보를 보호하기 위해선 촬영자에 의해 만들어진 이야기의 흐름을 깨지 않아야 한다. 따라서 본 논문에선 흐름을 측정하기 위한 유사성 측정 방법을 제안하고 이것의 분석을 통해 단계별 조각 매칭기법을 제시한다.

우리의 시스템은 크게 매체 분석 모듈과 매체 매칭모듈로 나눌 수 있다. 입력 영상은 일반적인 촬영물이고, 음악의 입력은 웨이브파일을 기준으로 하였다. 데이터 분석 모듈에서는 음악과 비디오를 분할한 뒤 각 조각의 속도정보와 밝기정보를 분석하게 된다. 분석된 데이터베이스를 통해 주어진 음악의 조각에 가장 유사한 정보를 가지는 영상 조각을 찾아내는 작업을 매칭 모듈에서 해주게 된다. 만약 주어진 음악 조각에 맞는 적당한 영상을 찾지 못했을 경우, 다중 레벨(multi-level) 기반의 분할기법을 사용하여 음악 조각을 재분할해준 뒤 다시 매칭 되는 영상을 찾는다.

그림 1에서 보듯, 영상과 음악 조각은 서로 다른 길이를 가지기 때문에 길이를 정규화 하는 과정을 거치게 된다. 이때의 각 매체의 길이 정보를 결과 합성 시 타임와핑 해 줌으로서 원래 영상에 없었던 다이나믹한 장면의 생성도 가능해진다.

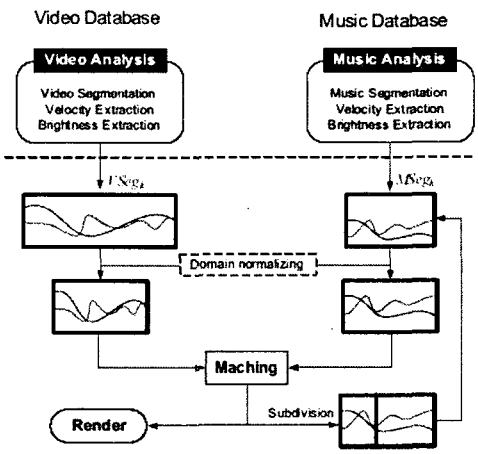

그림 1. 뮤직비디오 자동 생성 시스템의 개요도: 영상과 음악의 속도(red)와 밝기(blue)의 비교를 통한 자동 매칭기법

2. 기존 연구

비디오와 음악의 동기화에 대한 연구는 대부분 주어진 음악에 맞도록 영상 데이터를 수정 또는 재조합하는 방법으로 진행되었다. Foote 등 [1]은 음악의 반복적인 특성을 이용하여, 음악의 유사행렬(similarity matrix)을 계산하고 여러 조각으로 나눴다. 또한 비디오의 밝기 변화와 카메라 움직임을 분석하여 비디오도 여러 조각으로 나눈 후, 각 조각의 변환점(transition point)을 맞추는 방법을 소개하였다. 비디오 조각 각각의 특징과 음악의 특징을 고려하는 방법은 Hua 등 [2]에 의해서 시도되었다. Hua 등은 일반인이 찍은 홈비디오의 경우 화면의 질이 낮고 필요 없는 부

분들이 많을 것이라는 전제를 바탕으로 물체의 동작, 카메라 동작, 오디오 등을 토대로 비디오 샷(shot)마다 집중도(attention score)를 계산하여 중요한 샷만을 요약하는 기법을 소개하였다. 선택된 비디오 샷들은 박자(beat)의 세기를 기준으로 나눠진 음악 조각들의 빠르기(tempo)에 맞도록 대응시킴으로써 동기화를 시도하였다. Mulhem 등 [3]은 비디오 편집 전문가들이 통상적으로 사용하는 몇 가지 미적인 규칙(aesthetic rules)들을 사용하여 음악의 변화에 적절한 내용(특성)의 비디오 조각을 붙여나가는 방법을 제시하였다.

본 연구에서 제시하는 비디오와 음악의 동기화 기법은 분할 매칭이라는 점에서 Foote 등 [1]의 기법과 유사하지만, 다중 레벨 기반의 조각 매칭을 통해 원본영상의 흐름을 최대한 유지한다는 점에서 장점을 가진다.

3. 분할 기반의 영상분석

영상은 정보 전달을 위한 움직이는 이미지의 집합이다. 만약 음악에 맞는 비디오의 제작을 위해 비디오 클립의 임의의 부분을 붙여내어 가는 과정을 계속한다면 촬영자가 원하고자 했던 정보전달의 의미가 깨질 위험성이 있다. 따라서 우리는 비디오의 흐름을 보장하기 위한 조각단위로 비디오를 분석하겠다.

3.1 영상 분할

프레임간의 유사성이란 두개의 이미지가 얼마나 비슷한 색을 가지고 있느냐로 구분된다. 하지만 움직이고 있는 영상에서는 단순히 같은 좌표의 색상차이만으로는 유사성을 따지기 힘들다. 따라서 우리는 Contour shape matching [4]을 통한 유사성 측정을 통해 비디오를 분할한다. 임의의 N 개의 프레임으로 이루어진 영상 V_i (i=1,...,N) 가 주어졌을 때, 우리는 Canny edge detector [5]를 통해 이미지를 외곽선 맵 F_i로 변환 시켰다. 노이즈로 부터 발생할 수 있는 작은 외곽선을 방지하기 위해 영상클립을 미리 가우시안 필터링 해주었다.

F_i는 픽셀 단위의 외곽선들로 이루어져 있다. 외곽선으로 부터 얻어진 7개의 Hu-moment를 h^i_u (u=1,...,7) 라 했을 때 두 이미지형태의 차이는 다음과 같이 나타낼 수 있다.

$$I_{i,j} = \sum_{k=1}^{7} |1/m^{i_k} - 1/m^{j_k}|$$
$$where$$
$$m^i_k = sign(h^i_u)\log_{10}|h^i_u|$$

여기서, Hu-moment는 이동, 회전, 그리고 크기에 독립적이기 때문에 움직이는 영상에서 흐름이 끊기지 않는다면 유사하다고 판정을 한다 [4]. 위의 식에서 얻어진 유사행렬 $I_{i,j}$는 그림 2(a)에 나타나 있다.

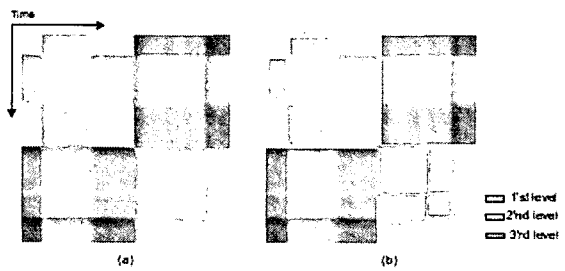

그림 2. 비디오 분할: (a) 유사행렬 $I_{i,j}$. (b) Radial symmetric kernel의 크기 조절을 통한 다중분할 결과.

Foote 등 [1]은 Radial symmetric kernel를 통해 유사행렬을 분할하는 기법을 소개하였다. 우리는 앞에서 얻어진 $I_{i,j}$을 Radial symmetric kernel을 통하여 분할한다. kernel의 크기를 조정함으로서 그림 2(b)와 같이 다중 레벨로 분할이 가능하다. 총 3가지 레벨로 다중 분할을 적용하여, 결과를 모두 데이터베이스에 저장한다.

3.2 영상의 특성 분석

영상은 촬영 기술 또는 압축정도에 따라서 많은 노이즈를 가지기 때문에 단순한 픽셀의 색 비교로는 정확한 움직임 분석이 힘들다. 따라서 본 논문에선 이미지의 외곽선을 기반의 비디오 움직임 측정기법을 제시한다. 외곽선 기반의 움직임 측정을 위해 외곽선 위의 한 점에 대하여 w의 크기를 가지는 영역 $\phi_{x,y}(p,q)$ (p,q는 윈도우 내에서의 좌표)를 선언해 준 후, 다음과 같은 색의 L^2-거리를 통해 움직임 벡터를 추출한다.

$$D^2 = (\phi^i_{x,y}(p,q) - \phi^{i+1}_{(x,y)+ve^i_{x,y}}(p,q))^2$$

여기서, i번째 프레임과 i+1번째 프레임의 컬러의 차이인 D^2를 최소화하는 벡터 $ve^i_{x,y}$가 외곽선위의 한 점 (x,y)에 대한 움직임 벡터가 된다. 외곽선이 아닌 부분의 움직임은 무시하기 위해 $F_{i(x,y)}=0$인 픽셀에서는 벡터를 (0,0)으로 고정해 준다. 보다 안정적인 결과를 위하여 지역적인 Lucas-Kanade [6]기법을 적용하여 결과를 보완하였다.

외곽선 움직임 정보를 통해 우리는 프레임 i의 속도 V^i_{vel}를 다음과 같이 구할 수 있다.

$$V^i_{vel} = \frac{1}{n \times m} \sum_{x=1}^{n} \sum_{y=1}^{m} \|ve^i_{x,y}\|$$

영상의 다른 특성인 밝기정도의 측정은 프레임을 이루고 있는 전체 픽셀의 밝기 분포를 통해 얻어내야 한다. 우리는 고전적으로 이용되고 있는 Histogram분석 방식을 사용하여 비디오의 밝기 정보 V^i_{bri}를 추출하였다. 외곽선의 움직임 벡터 $ve^i_{x,y}$의 크기는 그림 3(c)에 나타나 있다.

그림 3. 에지 기반의 비디오 속도 분석 : (a) 비디오 클립, (b) 외곽선 결과, (c) 움직임 벡터 맵

4. 분할 기반의 음악분석

분석된 비디오와의 매칭을 위해선 음악 역시 조각단위로 잘라서 분석을 해야 한다. 하지만 웨이브 형식의 음악에서 얻을 수 있는 정보는 특정 시간 t에 대한 진폭(amplitude)정보 밖에 존재하지 않는다. 따라서 주어진 진폭에 대한 시그널 분석을 통해 음악을 분할하고 속도와 밝기 정보를 측정해야한다.

음악의 분할을 위해 영상 분할과 유사한 방식으로 유사행렬을 구하도록 하겠다. Foote 등 [1]은 음악의 변화량을 구하기 위해 푸리에 변환을 사용한 Novelty라는 개념을 제안하였다. 우리는 Novelty score M^i_{nvl}를 음악

의 속도로 가정하였고 속도 변화량이 급격한 부분을 Radial symmetric kernel을 사용하여 분할의 기준으로 삼았다. 영상에서 사용한 다중 레벨 분할은 실제 매칭이 이루어질 때 순차적으로 이루어진다.

음악의 밝기 정보를 얻기 위해 일반적으로 음향의 밝기 측정하는데 사용되어 지는 spectral centroid [3]를 측정하여 M^i_{bri}을 구해내었다.

5. 음악과 비디오의 다중 레벨 (Multi-level) 매칭

음악과 비디오의 조각을 각각 $Vseg_m$, $Mseg_n$라 가정했을 때, 각 조각은 속도, 밝기, 그리고 길이라는 3가지 속성을 가지고 있다. 우리는 이 3가지 속성 사용하여 두 개체를 매칭 하려한다. 각각의 조각은 서로 다른 길이를 가지기 때문에 비교를 위해선 전체 길이를 정규화 시키는 과정이 필요하다. 따라서 우리는 각 조각의 속도와 밝기 정보를 3차 허밋 곡선을 사용하여 보간한 뒤, 두 개체의 유사도를 측정하였다. 총 m개의 영상과 n개의 음악 조각이 있다 가정했을 때, 다음과 같은 유사도 함수가 성립한다.

$$C_{m,n} = w_1 \sum |V^m_{vel}(t) - M^n_{vel}(t)| + w_2 \sum |V^m_{bri}(t) - M^n_{bri}(t)| + w_3 |1 - \frac{\min(T^m, T^n)}{\max(T^m, T^n)}|$$

여기서 T^m과 T^n은 각각 영상과 음악조각의 길이를 뜻한다. w값의 제어를 통해 특정 항목의 중요도를 조절할 수 있다. (본 논문의 결과영상을 위해서 우리는 $w_i = 5,2,3$을 사용하였다). 주어진 음악 조각에 대하여 최소값을 가지는 영상 클립을 데이터베이스에서 선택함으로써 자동적인 뮤직비디오 생성이 이루어진다.

최소화한 C값이 일정 임계값 이상이면, 음악을 재분할하여 다시 매칭을 찾아주는 다중 레벨 매칭을 실행한다. 임계값 이하의 매칭을 찾을 수 없는 음악 조각에 대해 Radial symmetric kernel의 크기를 줄여서 세분화한 다음 다시 매칭을 찾아주는 과정을 통해 최적화된 매칭을 찾아낸다. 만약 적절한 matching을 구하지 못할 경우 계속해서 재분할될 가능성이 있기 때문에 우리는 3단계로만 재분할을 적용하였다.

6. 결론 및 결과

우리는 뮤직비디오의 생성을 위하여 그림 3(a)와 같은 단편영화의 추격장면을 사용하였다. 5분의 길이를 가지는 영상의 다중 레벨 분할을 통하여 총 32개의 조각을 분석하였고, 여기서 얻었진 조각을 바탕으로 하여 2분가량의 추격에 적합하게 작곡된 음악과 매칭을 시도하였다. 비디오의 분석시간은 총 2시간이 걸렸고 음악의 분석은 실시간으로 계산되었다. 영상과 음악의 분할 매칭결과는 그림 4에 나타나 있다.

본 논문에선 두 개체의 연속성을 보장하는 비교를 촬영자에 만들어진 비디오의 정보를 최대한 유지하는 최적화된 매칭을 찾아내고자하였다. 외곽선기반의 비디오 분석을 통해 보다 안정적인 속도 분석이 가능하게 하였고, 비디오의 타임 와핑 가능성을 주어 보다 다이나믹한 뮤직비디오 생성이 가능하였다.

본 기술의 한계점은, 영상 분석시 노이즈에 영향을 많이 받는 다는 점이다. 따라서 우리는 현재 보다 정확한 영상 분석을 위해 Liu 등 [7]이 소개한 영상 분석기술과 연계하여 클립 내에서 세밀한 움직임까지도 자동적으로 찾아내는 시스템을 구상중이다.

그림 4. 영상과 음악의 분할 매칭 결과.

감사의 글

본 연구는 정보통신부 대학IT연구센터 육성지원사업의 연구결과로 수행되었습니다.

참고문헌

[1] Foote J., Cooper M., Girgensohn A.: Creating music videos using automaticmedia analysis. In Proceedings of ACM Multimedia. 553-560 (2002)

[2] Hua X. S., Lu L., Zhang H. J.: Ave - automated home video editing. In Proceedings of ACM Multimedia. 490-497 (2003)

[3] Mulhem P., Kankanhalli M.S., Hassan H., Yi J.: Pivot vector space approach for audio-video mixing. In Proceedings of IEEE Multimedia 28-40 (2003)

[4] M. Hu.: Visual Pattern Recognition by Moment Invariants, IRE Transactions on Information Theory, 8(2), 179-187, (1962)

[5] J. Canny.: A Computational Approach to Edge Detection, IEEE Trans. on Pattern Analysis and Machine Intelligence, 8(6). 679-698 (1986).

[6] Lucas, B., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision, Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI), 674-679 (1981)

[7] Liu. C., Torralba. A., Freeman T. W., Durand. F., Adelson H. E.: Motion magnification, In Proceedings of ACM SIGGRAPH 05, 519-526 (2005)