의미적 개념 기반 비디오 트랜스코딩 방법 및 시스템

정용주*, 김영석*, Truong Cong Thang*, 노용만*, 김태희**, 김재곤**
* 한국정보통신대학교 멀티미디어 그룹
** 한국전자통신연구원 전파방송연구소
{yro, yjjung}@icu.ac.kr

Semantic Concept-based Video Transcoding Method and System

Yong Ju Jung, Young Suk Kim, Truong Cong Thang, Yong Man Ro, Tae-hee Kim, and Jea-Gon Kim
Information & Communications University, Multimedia Group

Abstract: 본 논문에서는 다양한 사용자 환경에서 비디오의 범용적인 서비스를 위한 다차원 비디오 트랜스코딩의 판단에 관하여 논한다. 효율적인 판단을 위해 여러 영화 비디오 클립들을 비슷한 의미적 개념을 가지는 비디오들과 비슷한 장면 복잡도를 가지는 비디오들로 분류하고, 각 종류별로 주관적인 테스트(subjective test)를 실시하여 비디오 트랜스코딩에 있어서 사용자 인지(perception)의 특성을 분석한다. 이렇게 분석된 인간의 시각 특성들을 이용해 비디오 트랜스코딩 판단 궤적(trajectory)을 만들고 이를 다차원 비디오 트랜스코딩 판단 시에 적용하기 위한 방법을 제안한다.

1. 서론

최근에는 멀티미디어 소비 환경이 다양해짐으로써 사용자에게 멀티미디어 컨텐츠의 QoS(Quality of Service) 제공이 필수적인 요소로 부각되고 있다. 즉, 컨텐츠 제공자는 최종 사용자가 주어진 환경 자원에 가장 적합한 최대한의 품질을 가지는 멀티미디어 컨텐츠의 소비가 가능하도록 지원해야 할 필요가 있다. 사용자 환경에 맞는 최적의 품질 제공을 위한 최적의 비디오 트랜스코딩 전략을 찾는 문제는 적응 변환된 개체의 품질을 최대화하기 위해 기본적으로 네트워크나 터미널, 그리고 사용자 특성에 의한 제한 조건들을 충족시켜 줄 수 있는 최적의 트랜스코딩 연산(operation)들의 조합 정보를 찾는 것이다. 비디오 트랜스코딩 연산에는 공간-SNR-시간(spatio-SNR-temporal) 해상도(resolution)를 변화시키는 연산들이 있으며, 보다 효율적으로 비트를 할당하기 위해서는 이러한 다차원적 트랜스코딩 연산들 상호간에 트레이드오프(tradeoff)를 고려할 필요가 있다. 예를 들어서, 비트율이 감소하는 경우에 낮은 SNR 품질을 가진 컨텐츠를 좀 많이 전송하는 것과 높은 SNR 품질을 가진 컨텐츠를 좀 적게 전송하는 것 중에서 선택을 하거나, 또는 적은 용량의 프레임을 좀 많이 전송하는 것과 큰 용량의 프레임을 좀 적게 전송하는 것 중에서 선택을 해야 한다.

본 논문에서는 방송 통신 융합 환경과 같은 다양한 사용자 환경에서 비디오의 범용적인 서비스를 위한 다차원 비디오 트랜스코딩의 판단에 관하여 논한다. 본 논문에서는 여러 영화 비디오 클립들을 같은 의미적 개념을 가지는 비디오들로 분류하고 각 종류별로 주관적인 테스트(subjective test)를 실시하여 비디오 트랜스코딩에 있어서 사용자 인지(perception)의 특성을 분석한다. 이것은 비슷한 의미적 특성을 지닌 비디오들은 비슷한 트랜스코딩 전략을 가질 것이라는 기본적인 아이디어에서 출발한다. 예로, 액션(action)을 많이 포함한 비디오와 대화(dialog) 장면이 주를 이루는 비디오에 대해 사용자들은 트랜스코딩 된 결과의 선호에 있어서 각각의 고유한 선호 특성을 가짐을 생각할 수 있다. 따라서 다른 의미적 개념을 갖는 비디오들에 각기 다른 트랜스코딩 전략을 적용함으로써 사용자에게 최상의 품질을 제공할 수가 있다. 이렇게 분석된 인간의 시각 특성들을 다차원 비디오 트랜스코딩 판단 시에 적용하기 위한 비디오 트랜스코딩 판단 궤적 (transcoding decision trajectory: 이하 TDT)을 이용하는 시스템적인 판단 방법에 대해 논한다. 즉, 각 비디오의 특성에 따라 어떠한 TDT를 가지는지를 분석하고 주관적 테스트에 의해 얻어진 TDT를 이용하여 최적의 트랜스코딩 연산 조합이 결정 될 수 있음을 보인다.

2. 의미적 개념 기반 비디오 트랜스코딩

다양한 멀티미디어 소비 환경의 사용자들에게 환경의 제약 조건들을 만족하는 범위 내에서 최상의 멀티미디어 품질을 제공하기 위해서는 그림 1의 트랜스코딩 엔진에서 최적의 판단이 이루어져야 한다. 예로, 입력된 영화 클립을 무선 환경의 모바일 폰을 소지한 사용자에게 서비스하기 위해 비디오 트랜스코딩 엔진에서는 무선 네트워크와 터미널 환경에 맞는 영화 클립을 재 생산해서 전송할 필요가 있다. 이러한 트랜스코딩 판단을 위해 판단 엔진에서는 다음과 같은 질문의 답을 찾아야 한다.

입력된 이런 영화 클립은 이런 사용자 환경 하에서는 어떻게 트랜스코딩을 해서 서비스해야 하는가?

이러한 트랜스코딩 판단 문제를 풀기 위한 여러 연구들이 진행되어왔다. 트랜스코딩 판단을 위해서는 기본적으로 트랜스코딩 된 비디오의 품질 측정이 필요하며, 측정된 품질을 바탕으로 최상의 품질을 제공할 수 있는

트랜스코딩 연산들을 찾을 필요성이 있다. 품질 측정에는 크게 PSNR(peak signal to noise ratio)이나 MSE(mean square error)와 같은 객관적인 측정 방법과 인간의 주관적인 판단에 의한 측정 방법이 있으며, 두 가지 측정 결과에는 다소 차이가 있을 수 있다. 또한 움직임(motion) 또는 공간적 정보(spatial detail)와 같은 저 레벨(low level) 특징(feature)에 따른 사용자들의 트랜스코딩 선호 특성을 파악하여 위의 판단 문제를 풀고자 시도한 연구가 있었다 [3][4]. 하지만 비록 같은 시공간적 특성을 지닌 비디오라 할지라도 트랜스코딩에 있어서 사용자들이 다른 결과를 선호할 수도 있다. 예를 들어, 움직임과 공간적 정보가 많지 않은 특성을 가지는 대화(dialog) 비디오 클립과 이와 비슷한 저 레벨 특징을 갖는 교육 방송에서 주로 등장하는 텍스트가 오버레이(overlay) 된 비디오 클립에 대해서 사용자들은 각기 다른 트랜스코딩 선호도를 가진다. 즉, 같은 의미(semantic)를 갖는 비디오에 따라 판단을 하는 것과는 차이가 존재한다.

본 논문에서는 MSE와 같은 객관적 품질 측정에 의한 트랜스코딩 결과와 인간의 인지(human perception) 의해서 직접 선택 된 결과와의 차이를 보상해줌과 함께, 비디오의 의미에 따른 차이를 보상해주기 위한 접근법을 사용한다. 이를 위해, 비디오들을 각각의 의미적 개념에 따라 비디오들을 분류하고 각 종류별로 주관적인 테스트(subjective test)를 실시하여 트랜스코딩에 있어서 사용자 인지(perception)의 특성을 분석할 필요가 있다. 이렇게 사용자의 주관적 테스트에 의해 측정된 품질과 입력 비디오의 의미적 정보를 트랜스코딩 판단에 고려함으로써 최종 사용자에게 좀더 좋은 품질을 제공할 수 있다.

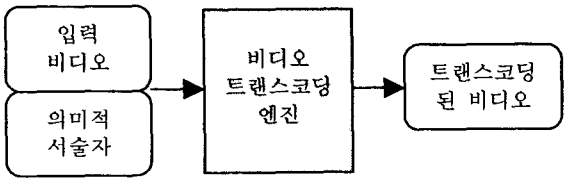


그림 1. 의미적 비디오 트랜스코딩 시스템 개념도

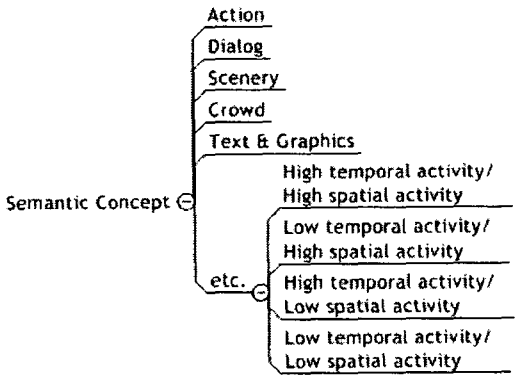


그림 2. 트랜스코딩 판단을 위한 영화 비디오 분류도

인간의 시각적 특성 분석을 위해 우리는 지난 논문 [1]에서 모바일 환경의 사용자들에게 영화를 서비스하기 위한 트랜스코딩 어플리케이션을 위해 의미적 개념을 그림 2에서와 같이 정의하고 세분화하였다 [5]. 이렇게 정의된 5개의 개념들에 따라 실험용 비디오들을 분류하고 테스트용 데이터를 작성하여 주관적 테스트(subjective test)를 수행하였다. 본 논문에서는 이와 더불어 5개의 범주에 속하지 않는 예외 비디오 클립들을 그것들이 가지는 저 레벨 특징들에 의한 장면 복잡도(scene complexity)로 분류하고, 이에 따라 각각 테스트를 진행하였다. 이렇게 수행된 테스트의 결과 분석을 통해 우리는 같은 의미적 개념 또는 같은 장면 복잡도를 갖는 비디오 종류에 따라 사용자들의 트랜스코딩 연산에 있어서 선호 특성들을 얻는다.

3. 주관적 테스트 및 인간의 시각적 특성 분석

우리는 대략 15초 정도의 길이를 가지는 27개의 영화 클립들을 가지고 주관적 테스트를 진행하였고, 이를 5종류의 다른 의미적 개념을 가지는 클래스들과 어느 의미적 개념에도 속하지 않는 예외 클래스들로 분류하였다. 즉, 5개의 각 개념 클래스마다 5개 영화 클립들과 예외 클래스를 위해 2개의 영화 클립들로 테스트 세트를 구성하였다. 테스트 원본 비디오들은 30 또는 25 프레임 율과 CIF 프레임 크기를 가지고 MPEG-4 심플 프로파일로 인코딩 되었다. 그리고 실험에서는 다음과 같은 다차원 트랜스코딩 연산들이 사용되었다.

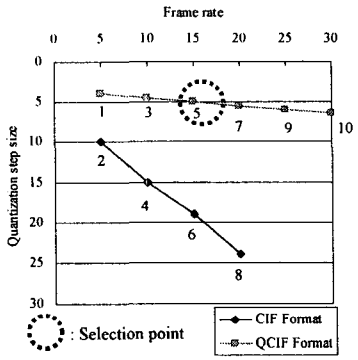
1) 재양자화(requantization) 파라미터: {1,…, 31},
2) 공간적 해상도: {CIF, QCIF},
3) 시간적 해상도: {30f/s, 25f/s, 20f/s, 15f/s, 10f/s, 5f/s}

이러한 트랜스코딩 연산들을 조합해서 실험을 위한 특정 비트율마다 여러 비디오 버전들을 생성하였다. 각 테스트마다 15명 내외명의 대학원생들이 참여하였으며, 테스트 방법으로는 강제 선택 방법(forced choice methodology)이 사용 되었다 [3]. 모든 사람에 대해 테스트를 마친 후에 각 사용자들의 선택의 수를 합하여 사용자들이 어떠한 버전(들)을 선택하는지 분석하였다. 그림 3은 대화 개념(Dialog concept)을 갖는 비디오에 대한 주관적 테스트의 결과의 예를 보여준다.

요약하여 각 개념의 특징을 비교하여 보면 다음과 같은 차이점을 발견할 수 있다. 의미적 개념에서 살펴보면, 등장인물 혹은 중심 목적물의 움직임이 매우 많다는 것을 특징으로 하는 액션 개념은 다른 개념들에 비하여 역동적인 느낌이 훨씬 강하게 표현된 트랜스코딩 조합이 선택 포인트로 분포되며, 프레임 율에 대한 선호도가 매우 강하게 적용되어 선택 포인트가 분포함을 살펴볼 수 있고, 주로 중심 배우들이 등장하여 대화하는 대화 개념은 다른 개념들에 비하여 배우들의 세부적인 표정변화 등을 표현할 수 있는 SNR 품질이 유지되는 트랜스코딩 조합이 선택 포인트로 선택됨을 볼 수 있다. 군중 개념은 군중의 느낌을 전할 수 있도록 큰 사이즈와 군중의 움직임을 표현 할 수 있을 정도의 프레임 율을 유지하는 트랜스코딩 조합이 선택 포인트로 선호되고, 자연경관 개념은 다른 개념들과 비교하여 보면 특별한 중심 목적물 또는 배우가 등장하지 않는 풍경의 특성을 살릴 수 있도록 SNR이 다른 개념들에 비하여 많이 떨어져도 큰 사이즈를 유지한 트랜스코딩 조합이 선택포인트로 분포함을 살펴볼 수 있다. 텍스트 & 그래픽스 개념은 다른 개념들과는 다소 차이가 존재하는 정보 또는 지식의 습득에 대한 의미적 개념이 존재하므로, 이러한 의미적 개념을 살릴 수 있도록 정보 전달의 주요 수단인 텍스트의 표현특성이 강하게 나타남을 확인할 수 있다. 보다 자세한 분석은 우리의 지난 논문 [1]을 참조할 수 있다.

4. 트랜스코딩 판단 궤적

트랜스코딩 판단 궤적(Transcoding Decision Trajectory: TDT)을 기술함에 있어서 우리는 비트율의 감소의 시점에 트랜스코딩 연산에 있어서 선택의 특징을



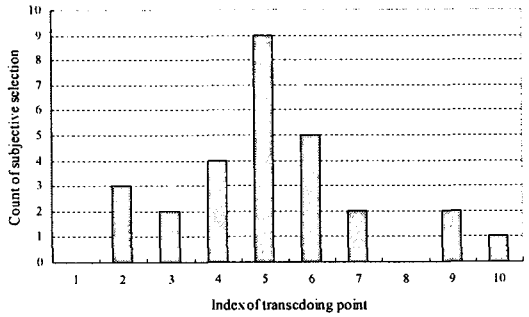


그림 3. 대화 개념 비디오의 트랜스코딩 선택 포인트(위)와 사용자에 의한 선택의 분포도 (아래)

분석해야 한다. 각 의미적 개념을 위한 궤적을 만들기 위해 우리는 각 특정 비트율에서 선택점(selection point)을 모두 표시하고 이 점들을 서로 연결함으로써 각 의미적 개념에 대한 TDT 선을 그린다. 실제로, 하나의 선을 그리는 것은 불가능하다. 왜냐하면 한 특정 비트율에서 사용자들이 선호하는 두 개 이상의 선택점이 동시에 존재하는 경우가 있기 때문이다. 그러므로 우리는 그림 4와 그림 5에서 보듯이 가장 높은 양자화 단계 크기(quantization step size)의 한 점만을 취함으로써 임계값(threshold)에 의한 경계선의 궤적을 그렸다. 그림 4와 5에서 각 선은 하나의 의미적 개념을 위한 TDT를 나타낸다. 또한 그림 4의 각 선의 왼쪽 아래의 정점은 그림 5의 오른쪽 위쪽 정점과 연결된다. 즉, 이것은 각 의미적 개념의 TDT가 CIF에서 QCIF형식까지 계속된다는 것을 의미한다. 이렇게 TDT를 생성함에 따라 우리는 의미적 개념에 따른 비디오 트랜스코딩의 특징을 관찰하고 비교 할 수 있다.

예를 들어, 그림 4에서 보듯 대부분의 유저들은 15f/s에서 25f/s범위의 트랜스코딩 된 영상의 품질을 구별하지 못한다. 그러므로 트랜스코딩 시스템은 SNR품질을 줄이고 원래의 25f/s를 유지함으로써 비트율을 감소시켜 대처할 수 있다. 다른 의미로 트랜스코딩 시스템은 TDT를 이용함으로써 가장 좋은 트랜스코딩 연산의 조합을 찾아내는 결정을 내릴 수 있다. 그림 8의 예에서와 같이 많은 트랜스코딩 후보(candidate)들이 존재할 때 시스템은 궤적 선에서 가장 가까운 후보점을 검색함으로써 최적의 트랜스코딩 연산 조합을 간단히 결정할 수 있다 [4].

또한, 앞서 설명한 저 수준의 특징(low level feature)을 분석하여 장면 복잡도(scene complexity)에 따라 비디오 클립들을 분류하기 위해 우리는 입력비디오의 공간적/시간적 활성도(spatial/temporal activity)를 계산한다. 공간적 활성도를 측정하기 위해서 우리는 각 프레임 타입에 따라 DCT AC 계수의 분산(variance)을 측정하였다 [2]. 그리고 시간적 활성도에 대해 평균 모션 강도(motion activity)를 측정하고 AC 계수의 분산에 따른 카메라 모션을 판별한다. 표 1은 테스트 영화클립에 대한 장면 복잡도의 예를 보인다. 표 1에 보인 "클립1"은 [2]의 설명에 의거하여 높은 공간적 상세도(spatial detail)를 가지며 느린 카메라 패닝(panning) 효과와 작은 움직임을 가지는 장면으로 설명 될 수 있다. 또한 "클립2"는 높은 공간적 상세도를 가지며 중간 정도의 카메라 패닝(panning) 효과와 아주 작은 움직임을 가지는 장면으로 설명 될 수 있다. 그림 6은 "클립1"과 같은 장면 복잡도를 가지는 비디오 클립들을 위해 앞서 설명한 주관적 실험에 의해 얻은 TDT 결과를 보인다.

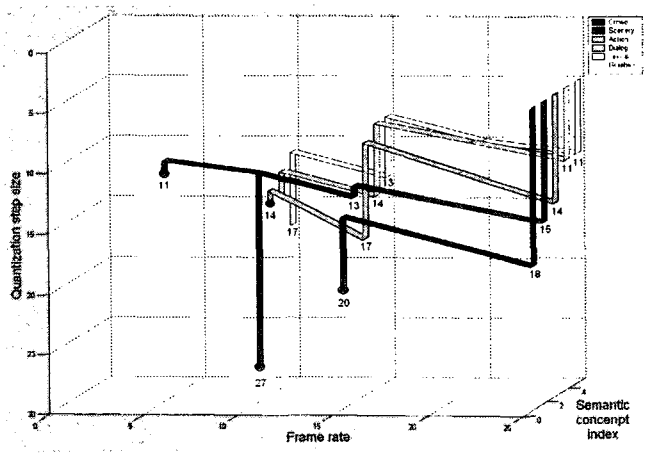


그림 4. CIF 해상도에서 각 의미적 개념에 따른 트랜스코딩 판단 궤적 (TDT)

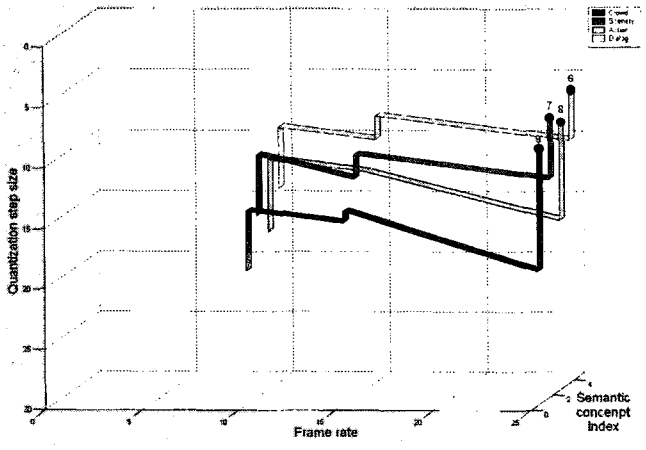


그림 5. QCIF 해상도에서 각 의미적 개념에 따른 트랜스코딩 판단 궤적 (TDT)

표 1. 저 수준의 특징에 따른 장면 복잡도

	Variance of I Frame	Variance of P Frame	Variance of B Frame	Motion activity
반지의 제왕 클립1	6,468	20	5	1.558

반지의 제왕 클립2	6,612	143	88	0.153

그림 1의 비디오 트랜스코딩 시스템에 대한 설명은 다음과 같다. 입력 비디오와 일종의 메타데이터로 구성된 의미적 개념들을 담고 있는 서술자(description)를 입력받아 트랜스코딩 엔진에서 비디오 트랜스코딩 과정을 거친 후 트랜스코딩 된 결과를 생성해낸다. 이를 위해, 우선 입력 비디오들은 같은 의미적 개념을 담고 있는 비디오 세그먼트들로 분류되고 각 세그먼트는 최적의 트랜스코딩 연산 조합을 찾기 위해 트랜스코딩 판단엔진을 거친다. 그림 7은 비디오 트랜스코딩 판단 엔진의 알고리즘도를 나타낸다. 먼저 입력된 서술자를 분석하여 어느 의미적 개념에 속하는 비디오 세그먼트인지 분류를 한다. 만약 그림 2의 5개 의미적 개념에 속하지 않는 세그먼트라면 "Low feature based Classification" 을 통해 시간적/공간적 활동도(activity)를 분석하여 분류한다. 이렇게 분류된 세그먼트를 위해 사전에 만들어진 TDT를 매핑(mapping) 시킨다. 최종적으로 그림 8에서와 같이 TDT를 이용해 여러 트랜스코딩 후보점(candidate)들 중에 최적의 트랜스코딩 연산 조합을 찾는다. 이렇게 검색된 최적의 연산 조합을 통해 트랜스코더에서 입력 비디오를 변환하게 되며 각 변환 된 세그먼트들을 다시 하나의 비디오 스트림으로 생성하여 출력하게 되는 알고리즘을 사용한다.

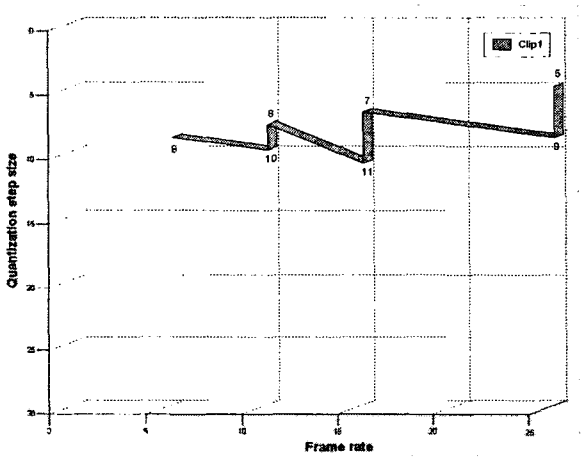


그림 6. 장면 복잡도에 따른 분류에서 표 1에 보인 "클립 1" 과 같은 복잡도를 갖는 비디오 클래스의 트랜스코딩 판단 궤적 (CIF 해상도)

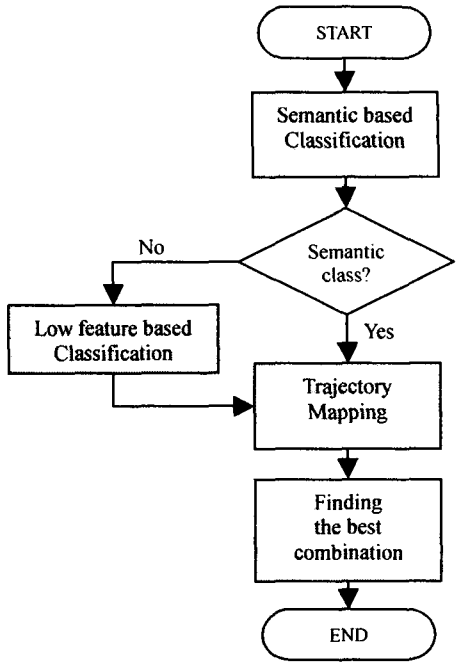


그림 7. 트랜스코딩 판단 엔진의 순서도

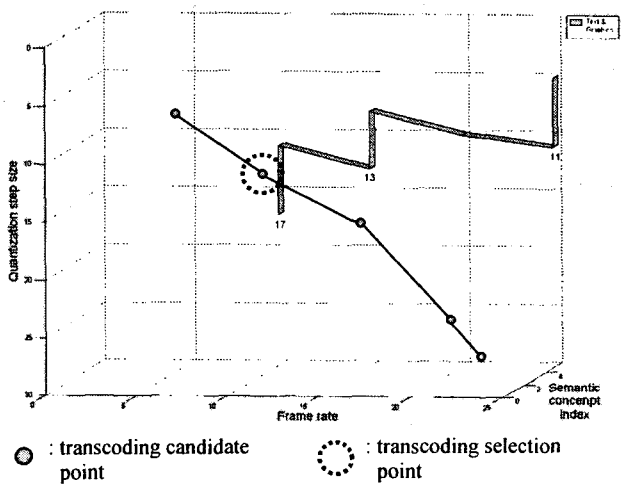


그림 8. 트랜스코딩 판단 궤적을 이용한 트랜스코딩 판단의 예제도

5. 결론

본 논문에서는 비디오 트랜스코딩 판단 문제를 풀기 위해 비디오 클립들을 의미적 개념에 따라 분류하고 예외 클립들은 장면 복잡도에 따라 분류하여 주관적인 테스트를 통해 사용자들의 인지 특성을 분석하였다. 이렇게 얻은 결과인 트랜스코딩 판단 궤적을 트랜스코딩 판단 시에 반영함으로써 최종 사용자들에게 보다 나은 품질의 비디오를 서비스 할 수 있을 것이다. 다음 연구에서는 좀 더 정확한 트랜스코딩 판단을 위해 저 레벨 특징에 따른 장면 복잡도를 좀더 세분화하고자 한다.

참고문헌

[1] 정용주, 김영석, 김덕연, 노용만, "의미적 개념 기반 비디오 트랜스코딩을 위한 인간의 시각적 특성 분석," 신호처리합동학술대회논문집, 제 15 권, 1 호, 2004.

[2] A. Puri and R. Aravind, "Motion-Compensated video Coding with Adaptive Perceptual Quantization," *IEEE Trans. Circuits Syst. Video Technol.*, vol.1, no.4, Dec. 1991.

[3] N. Cranley, L. Murphy, and P. Perry, "User-Perceived Quality-Aware Adaptive Delivery of MPEG-4 Content," in *Proc. NOSSDAV' 03*, pp. 42-49, June, 2003.

[4] Y. Wang, S.-F. Chang, and A. C. Loui, "Subjective Preference of Spatio-Temporal Rate in Video Adaptation Using Multi-Dimensional Scalable Coding," in *Proc. ICME*, June, 2004.

[5] J.-L. Koh, C.-S. Lee, and A. L. P. Chen, "Semantic Video Model for Content-based Retrieval," in *Proc. IEEE ICMCS*, pp. 472-478, June, 1999.