영상 통화 상황에서 안정적인 사람 영역 검출 방법

허선, *구형일, 조남익
서울대학교, *아주대학교
hsfra111@ispl.snu.ac.kr, *hikoo@ajou.ac.kr, nicho@snu.ac.kr

요 약

본 논문에서는 영상 통화나 웹캠 혹은 화상 회의 상황의 비디오 영상에서 안정적으로 사람 영역과 배경을 분리하는 방법을 제안한다. 이 방법은 카메라가 고정이라는 등의 제약을 두지 않고 자유롭게 움직이는 비디오 영상에서 사용자의 입력도 필요 없이 자동으로 사람 영역을 분리해 내게 된다. 첫 프레임에서 얼굴 검출을 통해 사람의 대략적인 위치를 추측하여 배경과 사람 영역을 Gaussian Mixture Model 로 모델링하고, 매 프레임 이 모델을 효율적으로 갱신한다. 그리고 비디오 영상의 연속성을 에너지 함수 설계에 적용하여 프레임간 사람 영역의 변화가 크지 않고 안정적으로 나오게 된다. 제안하는 방법은 기존 방법들에 비하여 제약이 적고, 사용자 입력이 필요 없으며 안정적으로 사람 영역을 분리함을 실험을 통하여 확인하였다.

1. 서론

주변에서 흔히 볼 수 있는 비디오 영상들은 크게 두 부분으로 이루어져 있다. 사용자가 관심 있어 하는 전경 영역과 그 외의 배경 부분이 그것이다. 따라서 전경과 배경을 분리해 내는 일은 영상을 분석하는데 도움을 줄 수 있고, 이를 이용하여 여러 다른 응용 프로그램들에 도움을 줄 수 있다. 예를 들면 주어진 비디오 영상에서 배경을 다른 영상으로 대체하여 사용자가 원하는 영상을 새롭게 만들어 내거나 화상 회의와 같은 상황에서 배경을 조정할 수 있고, 영상 통화나 웹캠을 이용한 통신에서 배경을 없앰으로써 사용자의 사생활을 보호할 수 있다. 또한, 영상을 배경과 전경으로 계층 분리할 수 있기 때문에 영상 압축 등에 활용될 수 있고, 데이터 전송에서 전송할 데이터의 용량을 줄이는 이득을 볼 수 도 있다.

특히, 스마트폰이 널리 보급되고 PC 를 이용한 온라인 커뮤니케이션이 많아지면서 일반 사용자들도 비디오 영상을 이용하는 사례가 많아지게 되었고, 따라서 영상에서 배경과 전경을 분리하려는 기술이 더욱 필요하게 되었다. 그리고 이러한 일반 사용자들이 많이 접하거나 사용하게 되는 비디오 영상 대부분이 사람을 전경으로 생각하는 영상들이므로 우리는 영상 통화나 웹 캠 혹은 화상 회의의 상황을 목표로 비디오 영상에서 사람을 분리하는 것을 목적으로 하였다.

비디오 영상에서 배경과 전경을 분리하는 알고리즘은 다음과 같은 몇 가지 사항을 고려해야 한다. 첫째는 전경 분리의 정확성이다. 전경 분리가 목표이므로 정확한 결과를 내주어야 한다. 둘째는 알고리즘의 동작 속도이다. 비디오 영상은 사진과 다르게 매우 많은 영상 시퀀스들이므로 알고리즘의 동작 속도가 빨라야 사용이 가능하다. 셋째는 사용자의 상호 작용이다. 전경 분리를 위하여 사용자가 얼마나 많은 입력을 주어야 하는지가 중요하다. 마지막으로는 알고리즘이 동작하기 위한 비디오 영상의 촬영 조건 제약이다. 많은 방법들이 카메라가 고정이어서 배경의 움직임이 없다는 가정을 하고 있다. 이러한 제약은 전경 분리 결과 성능을 향상시켜주지만 일반적인 상황에서 적용하기에는 무리가 있다. 이러한 사항들을 모두 고려하여 알고리즘을 설계해야 한다. 만약 정확성만을 중시하면 사용자가 비디오 영상 모든 프레임에 대하여 전경 부분을 일일이 지정하면 되지만, 이는 시간도 오래 걸리고 사용자의 입력이 많아야 하므로 사용하기 어려운 방법이다.

우리는 이 논문을 통하여 카메라가 움직이는 일반적인 상황의 비디오 영상에서 화상 통화에 지장이 없는 정확성을 주면서 빠르고 사용자의 입력이 전혀 없이 동작하는 방법을 제안한다. 기존 방법은 에너지 함수를 graph-cut 으로 풀어서 전경을 분리한다. 이런 방법은 한 장의 영상에서 상당히 정확하게 전경 분리를 해주지만 이 방법을 사용하기 위해서는 사용자 입력이 있어야 한다[1]. 또한, 이를 비디오 영상에 그대로 적용하면 모델을 학습하는데 시간이 오래 걸리고 시간이 지날수록 에러가 누적되어 나중에는 매우 부정확한 결과를 주게 되는 단점이 있다. 본 논문에서는 사용자의 입력 없이 사람 영역을 자동으로 추측하여 초기값을 설정하고, 새로운 프레임을 처리할 때 모델을 학습하는 시간을 줄였으며, 인접 프레임간의 관계를 이용하여 안정적으로 전경을 분리하도록 하였다. 다양한 영상에서 실험한 결과 좋은 품질의 전경 분리 성과물을 얻을 수 있었다.

2. 관련 연구

영상에서 배경과 전경을 분리하는 기본적인 방법은 에너지를 설계하고 이를 최적화하는 것이다. 이는 Boykov 가 graph-cut 을 활용하여 에너지를 최적화하는 방법을 제안한 이후 배경/전경 분리의 기본 알고리즘이 되었다[1,2]. 이후 다양하게 에너지를 설계하는 방법들이 제안되었고, 단일 영상에서는 매우 정확한 성능을 보여주었다[3,4]. 이 방법을 비디오 영상으로 확장하는 알고리즘들이 제안되었는데, 대부분 카메라가 고정이라는 제약을 두거나, 비디오 영상의 일정 프레임 간격으로 키 프레임을 주어야 하는 등의 사용자의 입력이 많아야 하고 꾸준히 있어야 한다는 제한이 있다. 그리고 비디오 영상을 다루다 보니 동작 속도가 느리다는 단점이 있다[5,6,7,8].

또한, 비디오 영상이다 보니 시간이 지날수록 첫 프레임과 현재 프레임 사이의 차이가 심해지므로 배경과 전경에 대한 모델이 어긋나게 된다. 이는 온라인 모델 갱신을 통하여 해결하게 되는데 갱신을 빠르고 정확하게 하는 것이 주요 관심사가 된다. [9]에서는 배경과 전경에 대한 모델 갱신을

각각 따로 진행하지 않고, 현재 프레임의 정보를 이용하여 프레임 전체를 학습하는 방법을 사용하였다. 이 방법은 프레임 사이에서 가려진 영역에 대해서도 잘 학습을 하여 모델 갱신이 꽤 정확하지만 배경과 전경 분리는 현재 프레임만 이용해서 하기 때문에 프레임간 전경 분리 결과가 안정적이지 않다는 단점이 있다.

우리는 영상 통화 혹은 웹 캠 상황의 비디오 영상에서 사람 영역을 분리시키는 것을 목표로, 카메라 고정에 대한 제약을 두지 않고 적정 동작 속도로 수행되며 프레임간 사람 영역 분리 결과가 크게 변화하지 않고 안정적으로 나오는 방법을 제안한다.

3. 사람 영역과 배경 영역의 추측

영상 통화나 웹 캠 상황에서는 일반적으로 한 명의 사람이 영상의 정면을 응시하고 있다. 이 사람이 우리가 관심을 두는 전경 영역이다. 영상에서 사람은 몸을 많이 기울이지 않고 대체로 똑바른 자세로 있게 되며, 사람의 크기는 영상의 전체 영역 중 많은 부분을 차지하고, 대체로 상반신까지만 영상에 나오게 된다. 이러한 가정을 이용하면 아래에 소개된 순서대로 쉽게 사람 영역을 추측할 수 있다. 일단, 영상의 첫 프레임에서 얼굴 검출 알고리즘을 사용하여 사람의 얼굴을 검출한다[10]. 얼굴 검출 알고리즘은 영상의 한가운데를 중심으로 일정 크기의 관심 영역을 설정하여 이 관심 영역에서만 적용한다. 이는 사람의 얼굴이 영상 가운데 있을 가능성이 높기 때문에 적용 가능한 일이고, 첫 프레임에서 얼굴이 검출되지 않으면 다음 프레임으로 넘어가게 되고 얼굴이 검출될 때까지 이 과정을 반복한다. 얼굴이 검출되면 사람 영역을 안전하게 추측하기 위하여 검출된 네모 박스의 크기의 80% 영역만 얼굴이라고 추측한다. 박스 전체를 얼굴 영역으로 추측하면 얼굴 주변의 배경 부분이 같이 사람 영역으로 들어갈 수 있기 때문에 이를 방지하기 위한 것이다. 또한, 같은 이유로 직사각형으로 추측된 얼굴 영역에서 턱 부분을 고려하여 직사각형의 아래쪽 테두리를 깎아낸다. 이렇게 얼굴의 위치를 추측한 후 우리는 사람 영역을 위 아래 2 가지 방향으로 확장하여 추측한다. 위 방향은 사람의 머리 부분을 추측하는 것이고, 아래 방향은 사람의 상반신 영역을 추측하는 것이다. 대체로 머리 부분은 사람 얼굴 길이의 30% 정도 확장하면 되고, 상반신 부분은 마찬 가지로 얼굴 길이의 30% 정도 밑에서 얼굴 넓이의 2 배 정도 좌우로 확장한 길이로 영상의 밑부분까지를 모두 상반신으로 추측한다.

영상에서 사람 영역을 추측하고 나면 배경 영역을 추측하는 일은 모폴로지 연산을 통하여 간단히 추측하게 된다. 추측된 사람 영역을 얼굴 크기에 비례하는 모폴로지 확장 연산을 통하여 확장하고, 전체 배경에서 확장된 영역을 제외한 부분을 배경이라고 가정하게 된다. [그림 1.]은 이렇게 사람 영역과 배경 영역을 추측하는 일련의 과정을 나타내고 있다.

[그림 1.] 얼굴 검출을 통한 사람 영역과 배경 영역의 추측

오른쪽 그림에서 흰색 부분이 추측된 사람 영역이고, 검은 색 부분이 추측된 배경 영역이다.

4. 사람 영역 분리

영상에서 배경과 전경을 분리하는 방법으로 잘 알려진 에너지 최적화 방법을 사용하여 사람 영역을 분리할 것이다. 각 픽셀에 할당되는 라벨을 x_i라 하면 $x_i \in \{0,1\}$이고 0 이면 배경을 나타내고 1 이면 사람 영역을 의미한다. 사람 영역을 분리하기 위하여 설계한 에너지는 다음과 같다.

$$\mathrm{E}(X) = \lambda_1 \sum_i E_1(x_i) + \lambda_2 \sum_{(i,j)\in V} E_2(x_i, x_j) + \lambda_3 \sum_i E_3(x_i, y_i)$$

여기서 $\mathrm{X} = \{x_i\}_{i=1}^{N}$ 는 현재 프레임에서 영상 전체의 라벨 정보이고, N 은 영상 전체의 픽셀 수이다. V 는 영상에서 인접한 픽셀의 집합을 나타내고, y_i 는 이전 프레임에서의 라벨 정보이다. $E_1(\cdot)$ 은 현재 픽셀의 정보가 배경과 전경의 모델 중에서 어느 모델에 더 알맞은 것인지를 보는 데이터 항목의 에너지가 되고, $E_2(\cdot,\cdot)$ 는 인접한 픽셀끼리는 같은 라벨 값을 가져야 한다는 제한에서 오는 에너지이다. 일반적인 단일 영상에서 배경/전경 분리는 위의 2 개 에너지를 갖고 에너지 최적화 알고리즘을 통해 라벨을 구하지만 비디오 영상에서는 이전 프레임과 현재 프레임의 관계를 더 볼 수 있으므로 이러한 관계를 $E_3(\cdot,\cdot)$ 의 에너지로 표현하였다. 이는 이전 프레임과 현재 프레임의 배경과 전경의 픽셀 숫자는 비슷해야 한다는 제한과 이전 프레임과 현재 프레임의 같은 위치의 라벨 값은 같아야 한다는 제한을 두는 항목이다.

$E_1(\cdot)$ 은 배경과 사람 영역을 모델링하여 에너지 값을 할당하면 되는데, 우리는 배경과 사람 영역의 모델을 원소의 개수가 5 개인 특징 벡터를 사용하여 Gaussian Mixture Model(GMM)으로 모델링 하였다. i 번째 픽셀의 특징 벡터 $z_i = (p_i, q_i, r_i, g_i, b_i)$ 는 2 개의 위치에 대한 값과 3 개의 컬러 값을 원소로 갖는다. 각 원소는 정규화(normalization)되어서 0 과 1 사이의 값을 갖게 된다. 배경과 전경은 아래와 같은 분포를 갖게 된다.

$$\mathrm{p}(z|l) = \sum_{i=1}^{M_l} \alpha_{i,l} G(z; \theta_{i,l})$$

여기서 $\mathrm{l} \in \{f, b\}$은 사람 영역과 배경을 나타내는 라벨이 되고, M_l 은 가우시안 분포의 개수, $\alpha_{i,l}$ 은 가우시안 분포의 가중치, $\theta_{i,l} = \{\mu_{i,l}, \Sigma_{i,l}\}$는 가우시안 분포의 파라미터로 평균과 공분산을 나타낸다. 이와 같은 분포에서 에너지는

$$E_1(x_i) = \begin{cases} \min_k \left(-log\left(\alpha_{k,b} G(z_i; \theta_{k,b})\right)\right) & if\ x_i = 0 \\ \min_k \left(-log\left(\alpha_{k,f} G(z_i; \theta_{k,f})\right)\right) & if\ x_i = 1 \end{cases}$$

이다.

$E_2(\cdot,\cdot)$는 인접한 두 픽셀 사이에서 발생하는 에너지이므로

$$E_2(x_i, x_j) = \frac{1}{d(i,j)} exp\left(-\frac{\|I_i - I_j\|^2}{2\sigma^2}\right)|x_i - x_j|$$

로 설계하면 된다. 여기서 $\mathrm{d}(\cdot,\cdot)$는 픽셀 사이의 거리를 알려주는 함수이고 I_i 는 픽셀의 컬러 값, σ는 인접 픽셀 사이의 컬러 값들의 분산이다.

$E_3(\cdot,\cdot)$ 는 이전 프레임과 현재 프레임 사이에서 발생하는

에너지로 비디오에서 한 프레임 사이에 많은 변화가 있지 않으므로 이전 프레임과 현재 프레임의 같은 위치는 같은 라벨을 갖도록 다음과 같이 설계한다.

$$E_3(x_i, y_i) = exp\left(-\frac{\|I_i - I_i^p\|^2}{2\omega^2}\right)|x_i - y_i|$$

여기서 I_i^p는 이전 프레임에서 픽셀의 컬러 값이고, ω는 이전 프레임과 현재 프레임 사이의 컬러 값들의 분산이다.

이렇게 에너지를 설계하고 잘 알려진 graph-cut 알고리즘을 사용하여 에너지 최적화 과정을 통해 현재 프레임에서 사람 영역을 분리해낸다[2,11]. 단, 에너지 최적화를 한 사람 영역 분리 결과가 여러 에러가 발생할 가능성이 있으므로 최종적으로는 connected component(cc)를 사용하여 가장 큰 cc 만을 사람 영역으로 할당한다.

5. 모델 갱신 과정

GMM 모델에서 사용한 특징 벡터 원소 중 위치와 컬러 사이에는 큰 상관관계가 없다고 해도 무방하므로 위치와 컬러 변수는 서로 독립이라고 가정할 수 있다. 따라서 공분산은

$$\Sigma_{i,l} = \begin{bmatrix} \Sigma_{i,l,s} & 0 \\ 0 & \Sigma_{i,l,c} \end{bmatrix}$$

가 된다. 공분산이 이처럼 블록 대각 행렬이면 가우시안 분포의 특성상 가우시안 분포를 2 개의 가우시안 분포의 곱으로 나타낼 수 있다.

$$G(z; \theta_{i,l}) = G(z; \mu_{i,l,s}, \Sigma_{i,l,s})G(z; \mu_{i,l,c}, \Sigma_{i,l,c})$$

매 프레임 GMM 파라미터 갱신은 잘 알려진 EM 알고리즘을 통하여 이루어진다[12]. 영상 통화나 웹 캠 상황에서 사람의 색이나 배경이 급격하게 변하지 않으므로 컬러 모델은 유지하고 위치에 대한 가우시안 분포의 파라미터만 갱신하도록 한다. 이 방식으로 파라미터를 갱신하면 비디오에서 사람 영역 분리 도중 에러가 발생했을 경우, 시간이 흐르면서 에러가 누적되지 않고 원래대로 복귀될 가능성이 많게 된다. 또한 위치 변수 2 개에 대해서만 EM 알고리즘을 수행하므로 5 개 변수를 사용할 때보다 빠른 속도로 파라미터 갱신을 수행할 수 있다.

파라미터 갱신을 위하여 모델의 샘플을 추출해야 하는데 새로 들어온 영상에서 어느 부분이 사람 영역이고 배경인지 알 수가 없다. 이전 프레임의 배경과 사람 영역 모델을 이용하여 사람 영역 분리를 수행할 수도 있지만, 이는 현재 프레임과의 괴리가 생길 수 있다. 따라서 배경과 사람 영역에 대한 GMM 을 나누지 않고, 하나로 합쳐서 영상 전체에 대한 GMM 을 만들고 현재 프레임의 정보를 사용하여 모델 파라미터를 갱신한다. 2 개의 GMM 을 하나로 합칠 경우 주의할 것은 각 가우시안 분포의 가중치를 조정하는 것인데, 이는 이전 프레임에서 배경과 사람 영역의 크기에 비례하여 조정한 후 전체 가중치 합이 1 이 되도록 정규화(normalization) 시켜주면 된다[9].

하지만 모델 파라미터 갱신 과정에서 현재 프레임의 모든 픽셀을 샘플로 EM 알고리즘을 수행하면 수행 속도가 매우 느린 단점이 있다. 그러므로 현재 프레임을 일정 크기 이하로 리사이즈한 영상을 샘플로 이용한다. 이 경우 현재 프레임 전체를 샘플로 수행한 파라미터와 정확도에서 큰 차이가 없으면서 수행 속도 면에서는 매우 빠른 속도로 수행할 수 있다는 장점이 생기고, 이는 비디오 영상을 다루는데 있어 매우 큰 이점이다.

6. 실험 결과

제안하는 방법을 직접 촬영한 비디오 영상과 인터넷에서 얻을 수 있는 영상에 대하여 적용하여 실험하였다. 일반적으로 사용되는 얼굴 검출기[10]를 이용하였고, 영상 통화나 웹 캠의 보통 해상도인 30 만 픽셀 정도의 영상을 이용하였다. 이 때, 알고리즘은 약 3~5 프레임/초 의 속도로 동작하였고, [9]와 비교했을 때, 프레임간 전경 분리 결과의 변화가 적고 안정적인 결과를 얻을 수 있었다. 그리고 [9]에서는 모델 파라미터를 갱신할 때 현재 프레임을 그대로 이용하는데 반해, 우리는 현재 프레임을 작게 리사이즈된 영상을 샘플로 이용하기 때문에 더욱 빠른 속도로 동작한다. [그림 2.]는 제안하는 방법의 실험 결과를 보여주고 있다. [그림 3.]은 [9]의 방법과 제안한 방법을 비교한 결과이다.

7. 결론

본 논문에서는 영상 통화 같은 상황의 비디오 영상에서 사용자의 입력 없이 자동으로 사람 영역을 배경으로부터 분리해내는 효율적인 방법을 제안하였다. 얼굴을 검출한 후 이로부터 사람 영역을 추측하고, 효율적인 GMM 모델 갱신 방법과 인접 프레임간 사람 영역을 안정적으로 분리하도록 해주는 에너지 함수를 설계하여 전체 비디오 영상에서 사람 영역을 효과적으로 분리 해내었다. 제안하는 방법은 기존 방법들보다 제약이 적어 더 다양한 상황에 적용될 수 있을 뿐 아니라, 보다 안정적으로 사람 영역을 분리해 냄을 확인할 수 있었다

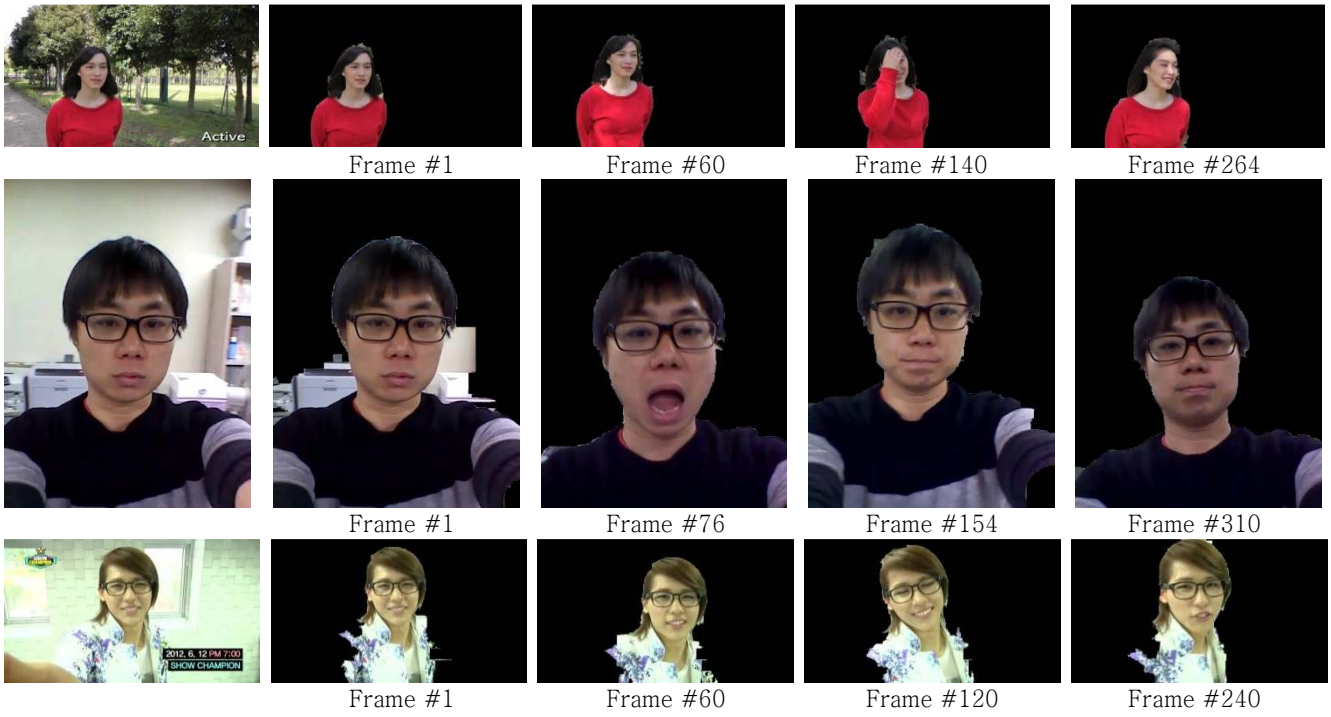
감사의 글

이 논문은 2013 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2009-0083495)

참고문헌

[1] Y. Boykov and M.-P. Jolly, “Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images,” In ICCV, 2001.
[2] Y. Boykov, O. Veksler and R. Zabih, “Fast approximate energy minimization via graph cuts,” In IEEE Transactions on PAMI, vol.23, no.11, pages pp1222-1239, Nov. 2001.
[3] Y. Li, J. Sun, C.-K. Tang and H.-Y. Shum, “Lazy snapping,” In SIGGRAPH, 2004.
[4] C. Rother, V. Kolmogorov and A. Blake, “Grabcut-interactive foreground extraction using iterated graph cuts,” In SIGGRAPH, 2004.
[5] A. Criminisi, G. Cross, A. Blake and V. Kolmogorov, “Bilayer segmentation of live video,” In CVPR, 2006.
[6] J. Sun, W. Zhang, X. Tang and H.-Y. Shum, “Background cut,” In ECCV, 2006.
[7] Y. Li, J. Sun and H.-Y. Shum, “Video object cut and paste,” In SIGGRAPH, 2005.

[8] M. Gong, and L. Cheng, “Forground segmentation of live videos using locally competing 1SVMs,” In CVPR, 2011.
[9] T. Yu, C. Zhang, M. Cohenm Y. Rui and Y. Wu, “Monocular video foreground/background segmentation by tracking spatial-color Gaussian mixture models,” In WMVC, 2007.
[10] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” In CVPR, 2001.
[11] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen and C. Rother, “A comparative study of energy minimization methods for markov random fields,” In ECCV, 2006.
[12] X. Meng and D. Rubin, “Maximun likelihood estimation via the ecm algorithm,” A general framework, In Biometrika, 80(2), 1993.



[그림 2.] 비디오 영상에 대한 사람 영역 분리 결과

[그림 3.] SCGMM[9]과 제안하는 방법의 비교.
위 행이 SCGMM 이고 아래 행이 제안하는 방법이다. 제안하는 방법이 SCGMM 보다 더 안정적으로 사람 영역 분리를 한다