Panoramic Video Generation Method Based on Foreground Extraction

김 상 환* · 김 창 수†
(Sang-Hwan Kim · Chang-Su Kim)

Abstract - In this paper, we propose an algorithm for generating panoramic videos using fixed multiple cameras. We estimate a background image from each camera. Then we calculate perspective relationships between images using extracted feature points. To eliminate stitching errors due to different image depths, we process background images and foreground images separately in the overlap regions between adjacent cameras by projecting regions of foreground images selectively. The proposed algorithm can be used to enhance the efficiency and convenience of wide-area surveillance systems.

Key Words : Panoramic video, Foreground extraction

1. 서 론

방법용 폐쇄회로 TV(CCTV)는 1980년대 중반 영국에서 교통법규 위반차량 단속을 위해 최초로 상용화된 이래로, 현재는 방법 및 출입 보안, 도난방지, 주차 관리 등 인적·물적 자산과 인명의 보호를 위한 다양한 목적으로 설치, 운영되고 있다. 특히 2001년 미국의 9·11 사태와 2005년 영국의 지하철 폭탄테러 사건과 같은 불특정 다수에 대한 테러 위협으로 인해 전 세계적으로 CCTV 설치는 폭발적으로 증가해 왔으며, 최근 국내에서도 아동 대상 성범죄가 사회적 이슈로 떠오르면서 많은 관심을 받고 있다. 현재 전국적으로 200만 대 이상의 CCTV가 설치된 것으로 경찰청은 추산하고 있고, 이 중 경찰청 관리 하의 CCTV의 경우 올해 6월 현재 2만8천397대로서 2008년 대비 3.2배 증가하는 등 범죄 및 사회 불안 요인으로 인한 공공과 민간 부문의 CCTV 설치는 가파른 증가세를 유지할 것으로 전망 된다. 국내·외에서의 이와 같은 CCTV의 폭발적인 증가 추세는 자연스럽게 이를 유지·관리할 감시 인력의 수급을 요구하게 되지만, 다수의 감시 영상을 하나씩 일일이 감시하는 비효율적인 현재의 방식으로는 지속적이며 신뢰도 높은 감시가 이루어지기 어렵다. 시장조사기관 IDC의 2008년 보고서는 2011년에 전 세계 네트워크 카메라 판매 비율이 아날로그 CCTV 카메라를 추월할 것으로 전망하고 있으며, 이에 따라 가까운 미래에는 대부분의 지역에 대한 영상 취득이 네트워크를 통해 이루어질 것으로 예상된다. 따라서 이처럼 넘쳐나는 영상 정보를 개별적으로 감시·관리하기보다는 인접한 다수 카메라 간의

영상 합성을 통한 통합된 시점의 영상 정보를 제공함으로써 효율성을 제고할 필요가 있다.

영상 합성에 관한 기존의 연구는 대부분 정지 영상으로의 합성에 집중되어 있다. 기존의 연구에서는 입력 영상을 받는 카메라의 종류가 고정된 카메라, PTZ 카메라, 어안 렌즈 카메라, 일반 디지털 카메라 등으로 다양할 뿐, 다수의 정지 영상 혹은 단일 동영상을 입력으로 삼아 단일 정지 영상으로 합성하는 연구가 주류를 이루고 있다. 그러나 앞에서 언급한 감시 시스템은 다수 카메라로부터의 동영상 입력을 단일 동영상으로 합성하는 파노라마 비디오의 생성 기법을 필요로 한다. 또한 통상의 정지 영상으로의 합성 기법이 주로 원경만으로 구성된 풍경 영상, 혹은 깊이가 비교적 고른 영상 등을 그 소재로 삼는 반면에, 감시 시스템 용도의 파노라마 비디오 생성 기법은 상이한 깊이를 지니는 전경과 배경이 혼재하는 일상생활 주변의 영상을 대상으로 한다. 이러한 차이점으로 인해, 영상 내의 상이한 깊이를 무시한 채 영상 전반에 걸쳐 동일한 평면으로의 투시 변환만을 취하는 합성은 특히 전경 부분에서의 고스팅 (ghosting) 현상을 일으키게 된다.

본 논문에서는 기존의 합성 기법을 차용하되 영상 내의 상이한 깊이 차이를 고려한 보다 자연스러운 합성 결과를 얻기 위해 배경의 추정 및 전경의 추출에 기반한 파노라마 비디오 생성 기법을 제안한다. 2장에서 영상 합성의 일반적인 원리에 대해 알아보고 3장에서 제안하는 기법을 설명한다. 4장에서는 구현된 시스템과 실험 결과에 대해 설명하고 향후의 연구 방향에 대해 논한다.

* 비 회 원 : 고려대 공대 전기전자전파공학부 석사과정

† 교신저자, 정회원 : 고려대 공대 전기전자전파공학부 부교수

E-mail : changsukim@korea.ac.kr

접수일자 : 2010년 12월 18일

최종완료 : 2011년 1월 24일

2. 대응점 정합과 전역 투시 변환

본 장에서는 단일 프레임의 영상 합성에 관련된 기본 원리와 이를 활용한 합성 과정에 대해 간략히 기술한 후, 이러

한 기법을 동영상에 적용했을 때 나타나는 현상으로부터 문제점을 도출한다.

2.1 특징점 기반 대응점 정합

서로 다른 시점을 지닌 인접한 다수 카메라 간의 합성을 위해 특징점 추출에 기반한 대응점 정합이 수행된다. 특징점을 추출하는 기법에 따라서 카메라 간 시점, 회전, 스케일, 광량의 차이 등과 같은 요소에의 강인함과 알고리즘 복잡도가 달라진다. 본 논문에서는 동영상 합성 시의 시간적 일관성을 유지하기 위해 초기화 과정에서 투시 변환을 산출하여 지속적으로 적용하므로 알고리즘 복잡도를 크게 고려할 필요가 없다. 이는 입력 동영상의 매 프레임마다 특징점을 추출하고 정합할 경우, 각 프레임마다 상이한 투시 행렬(perspective matrix)이 산출되어 투영 변환의 결과가 지속적으로 변하는 역효과를 일으키기 때문이다. 따라서 여러 요소에 가장 강인하다고 알려진 SIFT (Scale-Invariant Features Transform) [1] 기법에 의한 인접 영상 간 특징점 추출 및 정합을 다음과 같이 수행한다.

입력 영상에 $\sigma, 2\sigma, \dots, k\sigma$ 의 서로 다른 스케일링 팩터 (scaling factor)를 가지는 가우시안 커널에 의한 컨벌루션(convolution)을 취하여, 실제 영상의 해상도는 같지만 스케일은 다른 가우시안 영상의 집합을 생성한다. 식 (1)에서 $I(x, y)$ 는 입력 영상을 의미하며, $L(x, y, k\sigma)$ 은 양의 정수 k 와 분산 σ 를 가지는 식 (2)의 가우시안 커널에 의해 스무딩(smoothing)된 가우시안 영상을 의미한다.

$$L(x, y, k\sigma) = G(x, y, k\sigma) * I(x, y) \quad (1)$$

$$G(x, y, k\sigma) = \frac{1}{2\pi(k\sigma)^2} e^{-\frac{(x^2+y^2)}{2(k\sigma)^2}} \quad (2)$$

동일 옥타브 내의 인접한 가우시안 영상 간에 차분을 취함으로써 DOG(difference of Gaussians) 영상을 구성한다. 2개의 인접한 DOG 영상을 갖는 각 DOG 영상의 모든 화소에 대해서, 그 인접한 DOG 영상의 동일 위치의 화소 주변 3×3 블록을 포함한 26개의 이웃한 화소와 비교하여 극점인 화소를 특징점으로 삼는다. 현재의 옥타브를 다운샘플링(down sampling)하여 상기의 과정을 반복함으로써 특징점을 추출한다. 이와 같이 가우시안 커널 및 다운샘플링과 차분 및 극점 탐색에 의해 특징점을 추출함으로써, 스케일과 광량 변화에 강인한 특징점의 추출이 이루어진다.

각 입력 영상으로부터 추출된 특징점들을 정합하기 위해, 각 특징점이 속한 스케일에서의 이웃한 화소들의 그래디언트(gradient)에 기반하여 방향 히스토그램을 구성한다. 즉 16×16 크기의 블록을 4×4 크기의 서브-블록(sub-block) 16개로 분할하고, 각 서브-블록 내의 16개의 모든 화소에 대해서, 그래디언트의 방향을 기준으로 구성한 8개의 빈(bin)에 그래디언트의 크기 요소를 누적함으로써, 각 서브-블록을 8차원의 벡터로 표현한다. 결과적으로 각 특징점을 그러한 서브-블록 16개를 갖는 128차원의 벡터로 표현하게 되므로, 산출된 128차원의 벡터를 근거로 하여 가장 근접한 특징점을 정합함으로써 영상 간 대응점 탐색을 이루게 된다.

그림 1 인접 카메라 간 합성 결과 예시

Fig. 1 Examples of stitching images between adjacent cameras

2.2 투시 행렬의 산출 및 전역 변환에 의한 합성

추출되고 정합된 특징점 간의 변환 관계는 식 (3)과 같이 투시 행렬로써 표현될 수 있다 [2]. 그러나 영상 내 모든 화소가 동일한 깊이를 지니는 동일 평면 위에 위치하는 것은 아니므로, 식 (3)을 만족하는 완전한 해는 존재하지 않는다. 따라서 최소제곱법 근사, 혹은 RANSAC (Random Sample Consensus)[3]을 활용한 이상치(outlier) 제거를 통해 최적해를 산출하고, 이에 의한 전역 투영 변환과 블렌딩(blending) 등의 후처리를 통해 영상을 합성하게 된다. 식 (3)에서 ‘ $\sim$ ’ 기호는 정규화까지 고려했을 때 양변이 동치임을 나타내며, $X(x, y, 1)$ 와 $Y(u, v, 1)$ 는 각 입력 영상 간 대응되는 특징점을, H 는 그들 간의 변환 관계를 나타내는 투영 행렬을 의미한다.

$$HX \sim Y$$

$$H = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix}, X = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, Y = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (3)$$

그림 1의 (a)는 1절과 2절에서 기술한 방식으로 정지 영상 3장을 합성한 예이며, (b)는 인접한 두 카메라로부터 취득한 동영상의 한 프레임을 같은 방식으로 합성한 결과이다. 투영 변환의 적절함을 알아보기 쉽도록 평균 처리 외의 별다른 블렌딩은 취하지 않은 결과 영상이며, (a)가 자연스러운 합성 결과를 보이는 반면에 (b)는 특히 인물과 선인장 부분에서 크게 어긋나 있는 모습을 확인할 수 있다.

그림 1(b)의 부적절한 합성 결과는 투영 변환이 본질적으

로 한 평면을 또 다른 평면으로 투사하는 과정인 반면에, 영상 내의 화소들은 다양한 깊이를 가지므로 서로 다른 평면에 위치하는 것에 기인한다. 그림 1(a)와 같이 모든 화소들이 원경에 위치하면 그 깊이 차이가 무시할 수 있을 정도로 작아서 사실상 같은 평면상에 있는 효과를 받게 된다. 같은 논리로 모든 화소가 원경에 위치하지 않더라도 비슷한 깊이에 위치한다면 최적해 산출에 의해 좋은 합성 결과를 얻을 수 있다. 그러나 그림 1(b)와 같이 원경과 근경이 혼재하는 일상생활이 담긴 동영상의 경우, 깊이 차이가 현저한 영역 간의 오차로 인해 최적화에 의한 전역 변환으로는 더 이상 좋은 결과를 얻을 수 없다. 모든 대응점을 고려한 최적화를 수행하는 최소제곱법 근사를 수행한다면 영상 전반에 걸쳐 조금씩 들어지는 결과를 얻게 되며, RANSAC과 같이 가장 많은 대응점이 소속된 대응관계만을 고려한다면 그림 1(b)와 같이 배경 부분은 잘 들어맞지만 전경 부분에서 크게 어긋나는 결과를 얻는다.

3. 제안하는 방법

일상생활 주변의 동영상 합성에서 발생하는 부자연스러운 결과를 개선하기 위해, 본 논문에서는 인접한 카메라 간의 중첩 영역에서 배경과 전경을 개별적으로 처리하는 합성 기법을 제안한다. 우선 MOG (mixture of Gaussians) [4]에 의한 배경 추정을 개별 카메라별로 수행한다. 모든 카메라에 대해서 초기 배경 추정이 완료되면, 앞 장에서 기술한 방식으로 각 카메라의 배경 영상 간의 투시 행렬을 산출하고, 이로부터 인접한 카메라 간의 중첩 영역을 규명한다. 각 카메라의 입력 프레임에 대해서, 중첩 영역에 속하지 않는 영역은 앞서 언급한 깊이 차이로 인한 어긋남이 발생하지 않으므로 산출된 투영 변환을 그대로 적용한다. 반면에 중첩 영역에 속하는 영역에 대해서는 지속적으로 갱신되는 배경 추정 영상을 투영 변환한 후, 각 카메라로부터 추출된 객체 중 하나만을 취사선택하여 투영 변환함으로써 합성을 수행한다. 결론적으로 동영상의 시간적 일관성을 유지하기 위해 배경의 초기 추정 영상으로부터 산출된 투영 변환을 이후 합성 과정에서 그대로 적용되되, 인접한 카메라 간 오버랩 영역에서는 각 카메라로부터 추출된 전경 중 하나만을 취사선택하여 투영 변환함으로써 배경과 전경의 깊이 차에 의한 합성 오차를 방지하게 된다. 그림 2는 제안하는 알고리즘을 표현하는 개념도이다.

3.1 객체의 카메라 간 전환 시점 선택

인접한 카메라 간 중첩 영역에서 전경의 취사선택은 다음과 같이 수행한다. 객체가 오버랩 영역으로 진입하는 시점에서 해당 객체는 인접한 두 카메라 중 한쪽에서만 온전한 정보를 포함한다. 따라서 개별 객체에 대한 취사선택의 초기 부분과 말기 부분에서는 해당 객체가 진입해 온 카메라, 혹은 진입해 나갈 카메라를 선택한다. 그리고 이러한 양 말단 간의 중간 시점에서 카메라를 전환할 유일한 시점을 선택함으로써 카메라 간의 시점 일치성과 프레임 간의 시간적 일관성을 최대한 유지한다. 이러한 방식은 전경의 전환 시점에서 객체가 급격히 이동하는 부자연스러운 현상을 유발

그림 2 제안하는 알고리즘의 개념도

Fig. 2 The concept of the proposed algorithm

할 수 있다. 그러나 해당하는 한 프레임을 제외한 동영상 전반에 걸쳐서 자연스러운 결과를 제공할 수 있으며, 특히 복잡도가 낮아서 실시간 구현이 가능한 효율적인 알고리즘을 구성하게 된다.

3.2 객체의 카메라 간 전환 시점 선택

인접한 카메라 간 중첩 영역에서 전경의 취사선택은 다음과 같이 수행한다. 객체가 오버랩 영역으로 진입하는 시점에서 해당 객체는 인접한 두 카메라 중 한쪽에서만 온전한 정보를 포함한다. 따라서 개별 객체에 대한 취사선택의 초기 부분과 말기 부분에서는 해당 객체가 진입해 온 카메라, 혹은 진입해 나갈 카메라를 선택한다. 그리고 이러한 양 말단 간의 중간 시점에서 카메라를 전환할 유일한 시점을 선택함으로써 카메라 간의 시점 일치성과 프레임 간의 시간적 일관성을 최대한 유지한다. 이러한 방식은 전경의 전환 시점에서 객체가 급격히 이동하는 부자연스러운 현상을 유발할 수 있다. 그러나 해당하는 한 프레임을 제외한 동영상 전반에 걸쳐서 자연스러운 결과를 제공할 수 있으며, 특히 복잡도가 낮아서 실시간 구현이 가능한 효율적인 알고리즘을 구성하게 된다.

MOG에 의한 전경 추출은 화소별 전경·배경 판정 여부만을 제공하므로 이를 객체 단위로 고려하기 위해 연결구성요소 (connected components) [5] 기법을 활용하여 각 카메라별로 직사각형 형태의 마스크를 생성한다. 합성 영상에서 두 객체가 나타내는 상이성의 정도를 알아내기 위해 생성된 각 카메라의 마스크를 투영 변환한 후, 변환된 영역 간에 식 (4)와 같이 SSD (Sum of Squared Difference)를 산출하여 직전 프레임과 비교함으로써 객체의 카메라 전환 시점의 선택 기준으로 삼는다. 식 (4)에서 D_c 는 카메라 간 객체의 상이성을, n 은 두 마스크의 변환 영역 간 겹치는 화소의 총 개수를, $G_1(i)$ 은 첫 번째 카메라의 마스크의 i 번째 화소의

그레이 채널 값, $G_2(i)$ 는 두 번째 카메라의 마스크의 i 번째 화소의 그레이 채널 값을 의미한다.

$$D_c = \sum_{i=1}^n (G_1(i) - G_2(i))^2 \quad (4)$$

식 (4)는 카메라별 객체 간의 상이성을 효과적으로 표현한다. 이는 각 마스크가 직사각형으로 이루어져 있기 때문에 전경뿐만 아니라 주변의 배경도 다소 포함하고 있지만, 마스크 내의 배경 영역은 이미 배경 기준으로 산출된 일치 함수에 의해 정확히 정합되어 있기 때문에 식 (4)의 계산에서 효과적으로 제거되기 때문이다. 또한 전경으로 추출되는 그림자의 경우, 배경과 영상 깊이가 동일하기 때문에 역시 배경 기준으로 산출된 투영 변환에 의해서 효과적으로 제거된다. 결과적으로 크게 어긋나 있는 전경 영역에 의해 D_c 의 크기가 결정되므로, D_c 의 크기가 클수록 카메라별 객체 간의 상이성이 높다고 판별할 수 있다. 객체의 움직임이 비선형적인 급격한 변화를 보이지 않음을 전제하면 객체가 투영 변환의 오차가 커지는 방향, 혹은 그 반대 방향으로 꾸준히 이동한다고 가정할 수 있다. 따라서 현재 프레임의 SSD $D_{c,t}$ 와 직전 프레임의 SSD $D_{c,t-1}$ 을 비교하여 $D_{c,t}$ 가 $D_{c,t-1}$ 보다 크다면 카메라별 객체의 틀어짐이 심해지는 방향으로 이동함을 의미하므로, 현재 프레임을 직전 프레임에서의 선택을 전환하는 시점으로 선택함으로써 국부적인 오차를 최소화하게 된다. 반대로 $D_{c,t}$ 가 $D_{c,t-1}$ 보다 작다면 카메라별 객체의 마스크의 틀어짐이 약해지는 것이므로, 직전 프레임에서 선택한 카메라를 그대로 유지한다. 단, 추출된 전경의 마스크의 크기가 프레임마다 다소의 차이를 보일 수 있기 때문에 $D_{c,t}$ 와 $D_{c,t-1}$ 은 식 (5)와 같이 정규화를 취한 후 비교한다.

$$D_{c,t} = \frac{1}{n} \sum_{i=1}^n (G_1(i,t) - G_2(i,t))^2 \quad (5)$$

4. 실험결과 및 결론

실험 동영상은 PC와 1394 비전 카메라 3대를 이용하여 캠퍼스 내에서 직접 촬영하여 활용하였다. 각 카메라의 하드웨어 동작을 소스 코드 레벨에서 제어하여 프레임 누락 여부 및 동기화 여부를 실시간으로 확인하면서 촬영 및 저장을 병렬처리에 의해 동시 진행함으로써, 카메라 간에 동기화된 동영상을 실시간으로 획득하였다. 카메라 간 동기화 검증은 각 카메라의 레지스터(register)에 내장된 영상 획득 시간 정보를 비교함으로써 이루어졌으며, 카메라가 제공하는 최대 1/8000초 단위의 오차 범위 내에서 유효하다.

그림 3은 MOG에 의한 배경 추정 결과를 나타내는 영상으로서, 입력 프레임은 (a), (b), (c), (d)의 순서로 증가한다. 각 소번호의 영상 중 위의 두 영상은 인접한 카메라의 입력 영상을, 아래 두 영상은 그들로부터 추정된 배경 영상을 나타내며, 프레임의 증가에 따라 배경 영상의 추정이 이루어지는 과정을 보여주고 있다. 그림 4는 추출된 전경과 그 마스크

그림 3 배경 영상 추정 결과

Fig. 3 The results of estimating background images

그림 4 마스크 생성에 의한 객체 추출 결과

Fig. 4 The results of extracting objects by masking method

크 영상을 보여주는 결과로써, (a)는 입력 영상을, (b)는 MOG에 의한 전경·배경 판정 결과로써, (c)는 그림자를 제거한 객체 추출 결과로써, (d)는 추출된 객체의 마스크를 나타낸다. 그림 5는 제안하는 기법에 의한 최종적인 파노라마 동영상으로서, (a)는 전경 추출에 기반한 제안하는 알고리즘에 의한 합성 결과로써, (b)는 전역 투영 변환에 의한 일반적인 합성 결과를 나타낸다. (a)와 (b)의 비교로부터 제안하는 알고리즘이 앞서 언급한 그림 1에서와 같은 합성 오차를 크게 줄임으로써 시각적으로 자연스러운 합성 결과를 제공할 수 있다. 또한 매 프레임마다 투시 변환을 산출하는 대신에 배경 추정 및 전경 추출 결과를 활용하여 합성함으로써 보다 낮은 알고리즘 복잡도를 가능케 한다. 따라서 각 카메라별로 배경 영상 추정의 초기화 기간 이후로는 매 프레임마다 영상 합성이 바로 수행되므로 실시간 합성이 가능해진다.

제안하는 알고리즘은 배경 영역에 위치한 화소가 그 깊이

(a)

(b)

그림 5 파노라마 동영상 생성 결과 비교

Fig. 5 A comparison of results of generating panoramic videos

평면으로 투사하는 투영 변환에 의한 합성이 가능함을 전제하고 있다. 배경 영상 내에서도 현저한 깊이 차를 지니는 영역이 있을 수 있기 때문에, 엄밀히 말해서 이러한 전제는 사실이 아니다. 그러나 이러한 전제를 취함으로써 문제의 단순화를 꾀할 수 있으며, 더욱이 전경과 배경만을 구분하는 본 논문의 알고리즘을 차후 배경 내에서도 오차를 유발하는 영역을 구분하는 방식으로 확장함으로써 극복될 수 있는 부분이다. 차후의 연구는 이러한 부분을 개선하는 방향으로 이루어질 것이다.

감사의 글

본 연구는 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. 2010-0001958) 및 서울시 산학연 협력사업(No. ST090818)의 지원을 받아 수행된 연구임.

참 고 문 헌

- [1] D. Lowe, "Distinctive image features from scale-invariant keypoints", Int. J. Comput. Vision, vol. 60, no. 2, pp. 91-110, 2004.
- [2] R. Szeliski, "Image alignment and stitching: A tutorial", Tech. Rep. MSR-TR-2004-92, Microsoft Research, December 2004.
- [3] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography", Communications of the ACM, vol. 24, no. 6, pp. 381-395, June 1981.
- [4] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking", in Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 246-252, June 1999.
- [5] B. K. P. Horn, Robot Vision, The MIT Press, pp. 66-69, 299-333, 1986.

저 자 소 개

김 상 환 (金 相 煥)

2009년 8월 : 고려대학교 전기전자전파공학부 졸업 (공학사), 2010년 3월 ~ 현재 : 고려대학교 대학원 전기전자전파공학과 석사과정
<주관심분야> 영상처리

김 창 수 (金 昌 洙)

1994년 2월 : 서울대학교 제어계측공학과 졸업 (공학사), 1996년 2월 : 서울대학교 대학원 제어계측공학과 졸업 (공학석사), 2000년 8월 : 서울대학교 대학원 전기공학부 졸업 (공학박사), 2000년~2001년 : Visiting Researcher, USC, 2001년~2003년 : 서울대학교 전기공학부 박사후연구원, 2003년~2005년 : Assistant Professor, Chinese University of Hong Kong, 2005년 9월~현재 : 고려대학교 전기전자전파공학부 부교수
<주관심분야> 비디오 및 3D 영상처리, 멀티미디어통신