3차원 재구성을 위한 키 프레임 추출

*최종호 **유지상

광운대학교 전자공학과

*mrchoi90@kw.ac.kr **jsyoo@kw.ac.kr

Extraction of Key Frames for 3D Reconstruction

*Choi, Jongho **Yoo, Jisang

Department of Electronic Engineering, Kwangwoon University

요약

키 프레임 추출 기법은 2차원 비디오 영상을 3차원으로 재구성하기 위해 꼭 필요한 프레임을 선택하는 방법이다. 본 논문에서는 비디오에서 빠르게 프레임을 검사하며 최적의 키 프레임을 선택하는 기법을 제안한다. 제안하는 기법은 3차원 재구성을 위한 전처리 과정에 초점을 둔 것으로 프레임 간 대응점 비율 검사를 통해 프레임의 도약 강도를 결정하고 기하 모델 추정이 원활한 프레임을 선택한다. 이로부터 3차원 복원 후처리 과정을 통해 최종적인 3차원 점군(point cloud) 데이터를 획득한다. 실험을 통해 다른 기법과 성능을 비교했을 때, 제안하는 기법이 복원 소요 시간도 적게 들고 보다 밀집된 3차원 데이터를 얻을 수 있었다.

1. 서론

현실의 물체 또는 장면을 3차원으로 재구성하는 것은 컴퓨터 비전에서 가장 도전적인 분야 중 하나이다. multi-view stereo(MVS) 또는 structure from motion(SfM) 방법이 대표적이며 다양한 분야에서 진전을 보였다[1,2,3]. 다양한 위치에서 획득한 영상들을 바탕으로 카메라 파라미터와 영상 간 카메라 위치(pose) 추정을 통해 대상을 3차원으로 복원한다. 다만 MVS 또는 SfM은 일반적으로 고정 시점에서의 영상을 이용하는데, 각 영상은 대상의 부분적인 형상만을 나타내기에 자연스럽게 은면(hidden surface)이 발생한다. 따라서 대상의 표면을 온전히 복원하기 위해서라도 풍부한 다시점(multi-view) 영상의 획득은 필수적이다.

그러나 다시점 영상의 수는 3차원 복원 성능에 비례하지 않는다. 영상의 수가 증가할수록 3차원 복원의 밀도와 완성도가 증가하지만, 그 정도가 일정 수준을 넘어가면 추가 영상은 불필요하며 오히려 복원 과정에서 계산량만 늘어난다. 사실상 적정량의 다시점 영상을 확보하는 것이 관건이다. 효과적인 영상 획득 가이드라인을 바탕으로 적절한 위치에서 다시점 영상을 얻는 것이 가능하지만 촬영에 상당한 시간이 소모된다. 멀티 카메라 시스템을 활용해 영상 확보 시간을 줄일 수 있지만 고비용 및 경량화 문제가 남아 있다.

이에 대한 대안으로 비디오 스트림을 활용하는 방법이 있다. 제멋대로 획득한 영상들과 비교했을 때, 비디오는 더 유용한 정보를 제공할 수 있다. 게다가 복원 대상을 스캔하듯 촬영하면 앞선 영상 획득 방법들에 비해 효율적으로 다시점 영상을 단시간에 얻을 수 있다. 단 비디오에서 키 프레임들을 선택하는 전처리 과정이 필요하다. 이때 키 프레임은 카메라의 전체 움직임 및 카메라 파라미터 추정 시 발생하는 오차를 최소화해야 한다. 키 프레임은 대상의 일부 형태를 공유한 채 비디오 스트림에서 최대한 적은 수로 선택되어야 한다. 보통 1분 분량의 비디오는 최소 1000 프레임 이상으로 구성되는데 모든 프레임에 대해 키 프레임 검사를 수행할 수 없다. 그러나 비디오는 시간적 중복성(temporal redundancy)을 갖기 때문에 특정 조건 하에서 일부 프레임을 거르며 키 프레임 검사를 수행하는 것이 가능하다.

본 논문에서는 2차원 비디오로부터 3차원 공간을 재구성하기 위하여 필요한 키 프레임을 추출하는 기법을 제안한다. 그림 1은 제안하는 기법의 흐름도로 3차원 재구성의 전처리 단계이다. 먼저 영상 내 특정점 검출 및 영상 간 대응점 비율 분석을 통해 비디오 스트림에서 빠르게 도약하며 기하 모델 추정을 위한 후보 프레임을 찾는다. 이후 후보 프레임 중에서 기존 키 프레임과 기하 모델 추정이 용이한 프레임을 다음 키 프레임으로 선택한다. 이 과정을 통해 비디오 스트림에서 키 프레임 집합을 획득하게 된다.

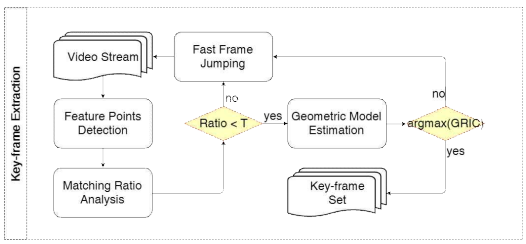

그림 1. 제안하는 기법의 흐름도

2. 방법

서로 다른 둘 또는 그 이상의 영상을 매칭하여 대응점을 찾는 문제는 3차원 재구성의 중요한 선결 과제이다. 대체로 비디오에서 특정 프레임과 그 이후의 프레임은 카메라 움직임이 유사한 특징이 있다. 여기서 카메라 움직임을 광류(optical flow)로 해석하면 프레임 간 대응점을 강인하게 검출할 수 있다. 본 논문에서는 광류 추정의 하나인 KLT(Kanade-Lucas-Tomasi) 특징 추적기[4,5]를 이용한다.

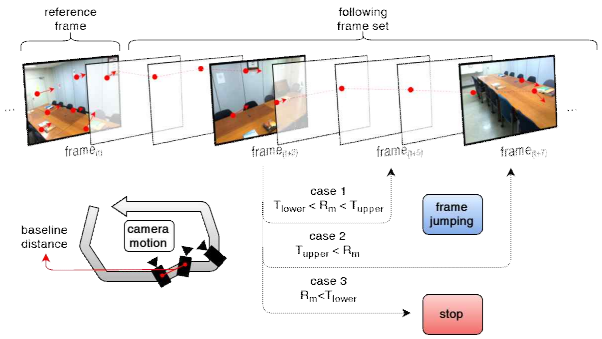

그림 2. 광류 추정 및 고속 프레임 도약 방법

프레임 간 대응점을 찾아내면 삼각비(triangulation) 공식을 이용하여 깊이를 계산할 수 있다. 이때 깊이 계산의 불확실성을 줄이기 위해서는 카메라 간 기준선(baseline) 거리를 늘려야 한다. 키 프레임 역시 이 조건을 충족하면서 추출되어야 한다. 기준선 거리를 판단하는 척도로 식 (1)을 이용한다.

$$R_m = \frac{N_t}{N_f} \quad (1)$$

여기서, N_f는 기준 프레임의 특징점 수를, N_t는 다음 프레임에서 추적에 성공한 특징점 수를 나타낸다. 기준선과 비율 R_m은 반비례 관계로 기준선을 충분히 확보하기 위해 R_m이 낮은 프레임을 선택해야 하지만, 기하 모델 추정에 필요한 대응점 수도 줄어들어 카메라 위치 추정에 영향을 미친다. 따라서 R_m이 상한 문턱치 T_{upper}와 하한 문턱치 T_{lower} 사이를 만족하는 해당 프레임들은 기하 모델 추정을 위한 후보 프레임으로 간주한다.

동시에 R_m이 이중 문턱치를 만족하는 프레임들을 빠르게 찾는 것도 중요하다. 그림 2에서 기준 프레임과 다음 프레임과의 특징점 추적을 통해 R_m을 계산하면, 프레임이 시간이 지날수록 그 비율은 감소하고 기준 프레임과 더 이상 겹치는 영역이 없을 경우 대체로 R_m은 0으로 수렴하는 것을 알 수 있다. 이를 통해 모든 프레임들에 대해 R_m을 계산하지 않고 일부 프레임에서만 계산을 한다. 비디오에서 시간적으로 인접한 프레임 간 정보 유사도는 상당히 높으며 특정 프레임을 취사선택하여도 카메라 움직임 정보의 큰 손실은 없다. 기준 프레임 $frame_{(t)}$이 $frame_{(t)}$와 $frame_{(t+3)}$ 간 R_m으로부터 $frame_{(t+3)}$에서 얼마나 도약(jumping) 할지는 세 가지 경우로 생각할 수 있다. 첫 번째 경우는 적절한 기준선 거리로 간주하고 $frame_{(t+3)}$을 기하 모델 추정을 위한 후보로 삼고 $frame_{(t+5)}$로 약한(weak) 도약 후 $frame_{(t)}$와 R_m을 계산한다. 두 번째 경우는 기준선 거리가 충분하지 않은 경우로 $frame_{(t+7)}$로 강한(strong) 도약 후 $frame_{(t)}$와 R_m을 계산한다. 마지막 경우는 기하 모델 추정을 위한 대응점이 부족한 것으로 추가적인 도약을 중단하고, 첫 번째 경우를 만족하는 후보 프레임들 중에서 기본 행렬 추정에 더 적합한 후보를 다음 기준 프레임으로 채택한다. 문턱치와 도약 강도는 실험을 통해 경험적으로 결정한다.

서로 다른 위치와 방향에서 획득한 두 영상의 관계는 에피폴라 기하(epipolar geometry) 기반의 수학적 모델로 정의된다. 이를 기본 행렬(fundamental matrix)이라 부른다. 식 (2)에서 기본 행렬(F)은 두 영상의 시점 간에 상대적인 회전($R^{'}$,R), 평행이동(S_b) 그리고 내부 파라미터($K^{'}$,K) 정보를 요약하고 장면 구조에 독립적인 3X3 특이행렬(singular matrix)로 표현된다.

$$F = (K^{'})^{-T} R^{'} S_b R (K)^{-1} \quad (2)$$

기본 행렬은 7개 또는 8개의 대응점이 주어지면 에피폴라 제약조건으로부터 선형 방정식을 만들어 빠르게 직접 계산할 수 있으나, 대응점 간 잘못된 정합 탓에 부정확한 기하 모델링으로 이어질 수 있다. 본 논문에서는 이상치(outlier)에 대처하도록 강인한 추정 기법 중 하나인 LMedS(least median of squares)[6]을 통해 오차 함수를 반복 개선함으로써 최적화된 기본 행렬을 얻는다. 이것은 비선형(non-linear) 최적화 문제로서 Levenberg-Marquardt 기법[7]을 활용해 복잡한 추정의 계산비용을 줄인다. 위의 과정은 기본 행렬 추정 'FUNDEST'와 비선형 최적화 'LEVMAR' 라이브러리를 토대로 구현한다.

후보로 꼽힌 프레임들을 대상으로 기준 프레임과의 기본 행렬을 계산한다. 그 중 기준 프레임과 최적의 기하 모델을 형성하는 후보 프레임을 다음 기준 프레임으로 선택한다. 잘못된 대응점 집합은 최적 추정의 불안 요소로 작용하지만 애초부터 기본 행렬 추정이 수치적으로 불안정한 상황도 존재한다. 이를 저하(degenerate) 상황이라 부른다. 크게 두 가지 경우로 하나는 3차원에 모든 대응점들이 동일 평면(coplanar)상에 위치하는 구조 저하이고, 다른 하나는 평행이동 없이 오로지 카메라 주점(focal point)에 대해 회전 움직임만 갖는 움직임 저하이다. 저하 조건에서는 기본 행렬보다 평면 간 투영 변환을 표현하는 호모그래피(homography) 모델로 표현하는 것이 더 적합하다. 하지만 호모그래피는 대응점 집합에 의존적이다. 다시 말해 동일한 장면 내에서도 대응점을 다르게 선택할 때마다 서로 다른 호모그래피를 얻게 된다. 따라서 기준 프레임과 후보 프레임 간에 기본 행렬과 호모그래피 중 어느 모델이 더 적합한지 판단하고 후자가 더 적합한 후보 프레임은 거른다.

본 논문에서는 두 모델 간 선택 기준으로 Geometric Robust Information Criterion(GRIC)[8]을 활용한다. 대응점이 주어지면 GRIC는 각 모델의 검정(test) 결과를 점수로 환산하고 더 작은 GRIC 점수를 갖는 모델이 주어진 대응점 집합에 최적 모델로 판정한다. 저하 상황을 피하기 위해서는 기본 행렬의 GRIC 점수가 호모그래피의 GRIC 점수보다 작아야 한다. 후보 프레임 중에 앞의 조건을 만족하면서 모델 간 GRIC 점수 차이가 가장 두드러진 프레임을 취한다. 따라서 식 (3)을 키 프레임 선택 함수로 이용한다.

$$K_{i+1} = \mathrm{argmin}_{j \in \forall(K_i)} \left(\frac{\left| GRIC_F(i,j) - GRIC_H(i,j) \right|}{GRIC_H(i,j)} \right) \quad (3)$$

여기서, i,j는 비디오의 프레임 인덱스, K_{i+1}은 새 키 프레임, $\forall(K_i)$은 키 프레임 K_i 기준의 후보 프레임 집합, $GRIC_{model}(i,j)$은 i,j 프레임 간 기하 모델의 GRIC 점수이다. 보통 은 모델 추정에 사용된 대응점 개수에 비례한다. 이에 따른 점수 차이를 보정하기 위해 모델 간 GRIC 차이에 $GRIC_H(i,j)$로 나누어 정규화한다. 결과적으로 정규화 값이 가장 큰 프레임을 다음 키 프레임으로 선택하게 된다.

3. 실험

그림 3. 비디오 스트림에서 추출한 중요 프레임 집합

그림 3은 세미나실을 촬영한 비디오 스트림에서 제안하는 기법을 토대로 선택한 중요 프레임 집합을 보여준다. 비디오는 1분 33초 분량으로 총 2817 프레임으로 구성된다. 표 (1)은 비디오 스트림에서 제안하는 기법과 uniform sampling 기법의 성능 비교를 보여준다. 키 프레임 수 측면에서는 uniform sampling의 경우 30 프레임 당 하나의 키 프레임을 추출하여 총 93장의 영상을 얻은 반면에 제안하는 기법은 그보다 적은 69장의 다시점 영상을 획득하였다. 3차원 복원은 VisualSFM[9,10]을 토대로 수행하였다. 복원의 질 측면에서는 제안하는 기법의 복원 시간이 덜 소요되었고 상대적으로 더 밀집한(dense) 점군(point cloud) 데이터를 얻었다. 이는 그림 4-5에서 확인할 수 있다. 비디오 스트림에서 키 프레임을 선택할 때, 제안하는 기법은 카메라의 전체 움직임 및 카메라 파라미터 추정 시 발생하는 오차를 줄이고 적절한 수의 다시점 영상을 획득했기 때문에 성능 차이를 보였다. 또한 복원 대상 촬영 시 발생할 수 있는 저하 상황을 고려하지 않았기에 uniform sampling 기법의 경우 영상의 수가 더 많더라도 오히려 저조한 복원 성능을 보였다.

표 1. 성능 비교

	키 프레임 수	복원 소요 시간	복원된 점의 수
uniform sampling	93	2203s	16,229
proposed method	69	1817s	43,861

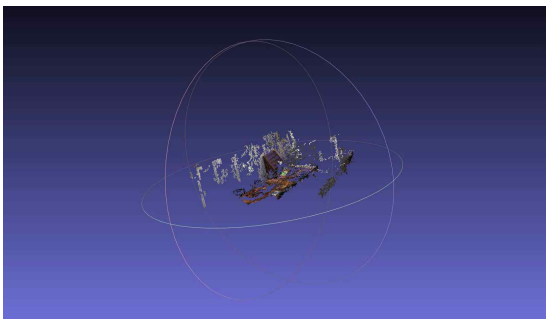

그림 4. uniform sampling 기법의 재구성 결과

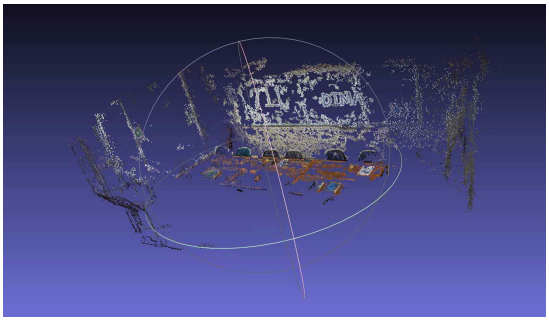

그림 5. 제안하는 기법의 재구성 결과

4. 결론

본 논문에서는 2차원 비디오로부터 3차원 공간을 재구성하기 위하여 키 프레임을 선택하는 방법에 대해 논한다. 풍부한 다시점 영상은 효율적 3차원 복원으로 귀결되므로 2차원 비디오에서 카메라의 전체 움직임을 반영할 수 있는 최적의 영상들을 추출하는 것이 중요하다. 제안하는 기법은 영상 간의 충분한 기준선 거리를 확보하면서 기본 행렬 추정이 용이한 키 프레임을 선택하고 실험을 통해서 그 성능을 확인하였다.

ACKNOWLEDGEMENT

"이 논문은 2016년도 장은공익재단 지원에 의해 연구되었음"

5. 참고문헌

[1] Ling, L., Burrent, I. S., & Cheng, E. (2012). A dense 3D reconstruction approach from uncalibrated video sequences. *ICMEW 2012*, 587 - 592.

[2] Frahm, J. M., Pollefeys, M., Lazebnik, S., Gallup, D., Clipp, B., Raguram, R. Johnson, T. (2010). Fast robust large-scale mapping from video and internet photo collections. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(6), 538 - 549.

[3] Ahmed, M. T., Dailey, M. N., Landabaso, J. L., & Herrero, N. (2010). Robust key frame extraction for 3D reconstruction from video streams. *VISAPP 2010*, 231 - 236.

[4] Shi, J. (1994). Good features to track. *IEEE Conference on Computer Vision and Pattern Recognition, 1994*(June), 593 - 600.

[5] Tomasi, C. (1991). Detection and tracking of point features. *Carnegie Mellon University Technical Report, 1991*(April), 1 - 22.

[6] Rousseeuw, P. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, *79*(388), 871 – 880.

[7] Lourakis, M. I. a. (2005). *A brief description of the Levenberg-Marquardt algorithm implemented by levmar. Foundation of Research and Technology (FORTH),* 4.

[8] P.H.S. Torr, A.W. Fitzgibbon, A. Zisserman, Maintaining multiple motion model hypotheses over many views to recover matching and structure, in: Proc. The 6th International Conference on Computer Vision, Bombay, India, 1998, 485-491.

[9] Wu, C., Agarwal, S., Curless, B., & Seitz, S. M. (2011). Multicore bundle adjustment. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 3057 – 3064.

[10] Wu, C. (2013). Towards linear-time incremental structure from motion. In *Proceedings - 2013 International Conference on 3D Vision, 3DV 2013,* 127 – 134