3 차원 복원을 위한 비디오에서 고품질 프레임 추출

최종호, *유지상
광운대학교 전자공학과
mrchoi90@kw.ac.kr, * jsyoo@kw.ac.kr

The extraction of high-quality frame from video for 3D reconstruction

Choi, Jongho *Yoo, Jisang
Department of Electronic Engineering, Kwangwoon University

요 약

비디오 시퀀스에서 3D 모델을 복원하기 위해서는 기하 모델 추정이 용이한 프레임을 선택해야 한다. 본 논문에서는 안정 장치 도움을 받는 전문 비디오가 아닌 일반 비디오에서 고품질의 프레임을 손쉽게 자동 추출하는 방법을 제안한다. 제안하는 기법은 optical flow 기반 매칭 분석, 프레임 간 적당한 기준선 거리 판단, 비디오 내에서 빠른 탐색을 위한 고속 도약, 두 프레임 간의 호모그래피와 기본 행렬에 대한 GRIC 점수, 모션 블러 프레임 제거 방법 모두를 결합한다. 실내 공간에 촬영된 비디오를 이용한 실험을 통해, 우리의 방법이 모션 블러와 저하 움직임이 있는 상황에서 더 강건하게 3D point cloud 를 생성하는 것을 보여준다.

1. 서론

비디오 기반 복원(video-based reconstruction) 기법 [1-5, 9, 10]은 저비용으로 장면 또는 객체를 3 차원 복원하는 기법 중 하나이다. 비디오 내 장면 복원을 적용 할 경우 견고한 결과를 얻기 위해 비디오 프레임 내 일부만을 활용한다. 단 비디오에서 적절한 프레임을 추출하는 것이 중요하다. 장면을 비디오로 녹화하면 대용량 비디오 파일에 흐릿하고 잡음이 많은 중복 프레임으로 빈번하게 검출된다. 추출한 프레임의 품질이 저조하면 3 차원 복원의 성능은 현저하게 줄어든다. 제한된 실내 환경에서 복원과 달리, 실외 환경은 이를 제어하는 것이 더욱 어렵다. 특히 공간을 누빌 때 카메라의 갑작스런 움직임으로 인해 모션 블러(motion blur)의 발생은 불가피하다. 모션 블러는 프레임 내 적은 수의 특징점 검출을 유발하고 프레임 간 특징점 매칭의 신뢰도를 저하시켜 복원 성능의 저하를 야기한다. 게다가 복원 대상의 크기와 복잡성을 감안할 때 비디오로 녹화하는 데 보통 수 분 이상이 소요된다. 즉, 처리해야하는 프레임이 수천 개가 있음을 의미한다. 비디오의 모든 프레임에 대해 카메라 포즈(pose)와 3 차원 장면 구조를 추정한다면 상당한 연산량이 요구될 수 밖에 없다.

이러한 문제를 해결하기 위해 본 논문에서는 최적화 된 수의 고품질의 프레임을 자동으로 선택하고 추출한 다중 이미지로부터 3 차원 복원을 수행하는 방법을 제안한다. 제안하는 기법은 3 차원 복원을 위한 전처리 단계로 그림 1 은 고품질의 프레임을 자동 추출하는 기법의 흐름도를 보여준다. 본 논문의 구성은 다음과 같다. 2 절에서는 비디오에서 기본 행렬(fundamental matrix) 추정이 용이한 프레임을 우선적으로 추출하는 방법에 대해 다루고, 3 절에서 후보 프레임 기반 모션 블러 평가를 통해 최적의 키 프레임 집합을 자동 추출하는 과정에 대해 살펴본 후, 4 절에서 이러한 기법을 비디오에 적용하여 생성한 3 차원 복원 결과와 그 성능을 확인한다. 마지막으로 5 절에서는 본 논문에 대한 결론을 맺는다.

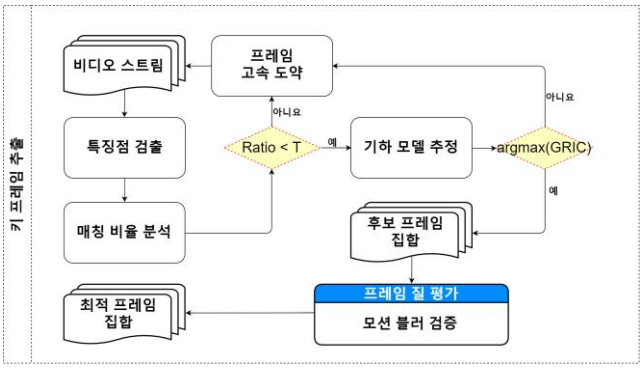

그림 1. 비디오로부터 키 프레임 추출 흐름도

2. 기본 행렬 추정이 용이한 후보 프레임 선택

대체로 비디오에서 특정 시간의 프레임과 그 이후의 프레임은 카메라 움직임이 유사한 특성이 있다. 여기서 카메라 움직임을 광류로 해석하면 프레임 간 대응점을 강인하게 검출할 수 있다. 영상 간 대응점을 찾아내면 삼각비 공식을 이용하여 간단히 깊이를 계산할 수 있다. 이때 깊이 계산의 불확실성을 줄이기 위해서는 카메라 간 기준선 거리를 늘려야 한다. 키 프레임 역시 이 조건을 충족해야 한다. 이러한 기준선 거리를 판단하는 척도로 Seo et al. [6, 7]처럼 식 (1)을 활용한다.

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

N_f 는 기준 프레임의 특징점 수를, N_t 는 다음 프레임에서 추적한 특징점 수를 나타낸다. 기준선과 비율 R_m 은 반비례 관계로 기준선을 충분히 확보하기 위해 R_m 이 낮은 프레임을 선택해야 하지만, 프레임 간 기본 행렬 추정에 필요한 대응점 수도 줄어들어 카메라 위치 추정에 영향을 미친다. 따라서 비율 R_m 이 상한 문턱치 T_{upper} 와 하한 문턱치 T_{lower} 사이를 만족하는 프레임들은 기본 행렬 추정을 위한 후보로 간주한다.

후보 프레임들을 대상으로 기준 프레임과의 기본 행렬을 추정한다. 그 중 기준 프레임과 최적의 기본 행렬을 형성하는 후보 프레임을 다음 기준 프레임으로 선택한다. 잘못된 대응점 집합은 모델 추정의 불안 요소로 작용하지만 애초부터 기본 행렬 추정이 수치적으로 불안정한 상황도 존재한다. 이를 저하 상황이라 부른다. 두 가지 경우로 하나는 3 차원에 모든 대응점들이 동일 평면(coplanar)상에 위치하는 구조 저하이고, 다른 하나는 평행이동없이 오로지 주점(focal point)에 대해 회전 움직임만 갖는 움직임 저하(motion degeneracy)이다. 이런 경우에는 대응점 집합에 대해 기본 행렬이 아닌 평면 간 투영 변환을 표현하는 호모그래피로 표현하는 것이 더 적합하다. 따라서 기준 프레임과 후보 프레임 간에 기본 행렬과 호모그래피 중 어느 모델이 더 적합한지 판단하고 후자의 모델이 더 적합한 후보 프레임은 거를 필요가 있다.

저하 상황을 거르기 위한 모델 선택 기준으로 geometric robust information criterion (GRIC)[8]을 활용한다. 대응점이 주어지면 GRIC 는 각 모델의 검정 결과를 점수로 환산하고 더 작은 GRIC 점수를 갖는 모델이 주어진 데이터에 더 적합하다고 판단한다. 후보들 중에서 기준 프레임과 기본 행렬의 GRIC 점수가 호모그래피의 GRIC 보다 작으며, 둘 간의 점수 차이가 가장 큰 후보 프레임을 다음 기준 프레임으로 선택하는 것이 좋다. 식 (2)는 키 프레임 선택 함수를 나타낸다.

$$K_{i+1} = \underset{j \in \forall(K_i)}{\operatorname{argmax}} \left(\frac{\left| GRIC_F(i,j) - GRIC_H(i,j) \right|}{GRIC_H(i,j)} \right) \quad (2)$$

i , j 는 비디오의 프레임 인덱스, K_{i+1} 은 다음 키 프레임, $\forall(K_i)$ 은 키 프레임 K_i 에 대한 모든 후보 프레임, $GRIC_{model}(i,j)$ 은 i , j 프레임 간 기하 모델의 GRIC 점수이다. 대개 $GRIC_{model}(i,j)$ 는 모델 추정에 사용된 대응점 개수에 비례하며 그에 따른 프레임 간 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. 이를 통해 비디오에서 빠르게 도약하면서 키 프레임 집합을 일차적으로 추출하게 된다.

3. 모션 블러 기반 키 프레임 선택

일반 비디오 촬영 방식은 공간을 이동하며 복원 대상을 녹화할 때 손 떨림과 카메라 이동 속도에 민감하다. 이에 따라 촬영할 때 불규칙한 움직임이 잦을수록 비디오 내 모션 블러가 두드러지고 프레임 간 장면 변화는 부드럽지 않게 된다. 결국 키 프레임 집합에서 모션 블러 프레임을 검출하고 이를 제거하는 과정이 필요하다. 모션 블러의 정도를 수치화할 수 있는 블러 측정법[10-13]이 중요한데 주로 영상의 밝기 변화에 초점을 맞춰 측정을 하게 된다. 보통 영상이 선명할수록 영상 내 에지 성분은 증가한다. 이를 통해 흐림의 정도를 영상 내 에지 성분의 역으로 정의할 수 있다. 식 (3)은 영상의 블러 측정(metric)을 나타낸다.

$$B_t = 1 \Big/ \sum_{P(x,y)} \{ (\frac{\partial f}{\partial x})^2 + (\frac{\partial f}{\partial y})^2 \} \quad (3)$$

모든 화소 위치 $P(x,y)$ 에서 영상 f 에 대한 x , y 축 방향 편미분의 제곱합으로 에지를 표현하고 이에 대한 역을 영상의 흐릿함 B_t 로 정의할 수 있다. 즉 영상 내 에지 성분이 덜 검출될수록 B_t 는 더 큰 값을 갖는다.

프레임 간 검출되는 에지 또는 잡음의 양은 상대적이기 때문에 오로지 B_t 만으로 흐릿함을 판단하는 것은 바람직하지 않다. 비록 프레임의 흐릿한 정도를 절대 평가할 수 없지만, 프레임 간 장면 변화가 극심하지 않을 경우 한정적인 이웃 프레임 내에서 상대적 흐릿함을 측정할 수 있다. Matsushita et al. [11]은 t 프레임의 상대적 흐릿함 RB_t 을 B_t / B_{t-1} 로 두고 그 값이 1 보다 작을 경우 t 프레임은 $t-1$ 프레임보다 더 선명하다고 판단했다. 그러나 프레임 간 에지 성분의 대소 관계만 보는 이분법적 판단은 모션 블러를 잘못 검출할 여지가 있다. 게다가 매 프레임 마다 기준 흐릿함이 다르게 적용되어 전체 프레임의 RB_t 변화를 관찰하기도 어렵다. 따라서 상대적 흐릿함 검사에 추가 제약조건을 부여하여 모션 블러를 검출한다.

4. 실험 결과

표 1. 실험 결과

	Seo et al. [7]	Ahmed et al. [10]	Proposed method
비디오 개수	7	7	7
복원 실패 횟수	1	1	0
프레임 평균 개수	3,380	3,380	3,380
추출한 프레임 평균 개수	98	81	52
복원 소요 평균 시간	6.2	4.5	5.5
저질 프레임 평균 비율	17.1	12.0	5.5

비디오로부터 추출한 키 프레임들은 VisualSfM toolbox [14, 15]을 활용하여 복원을 시도하였다. 표 1 은 기존 기법과의 성능을 비교한 것이다. 저조한 복원을 줄이기 위해 프레임 간 기하학적 관계를 고려한 기법[6, 9]은 더 나은 성능 보여주었다. 다만 추출한 프레임 집합 내 저질 프레임의 비율이 각각 17.1% [6]와 12.0% [9]로 그 영향을 무시할 수 없다.

이들은 모션 블러가 빈번한 일반 비디오의 특성을 온전히 감안하지 않았기에 특정 비디오에서는 복원 실패하였다. 반면에 제안하는 기법은 blur metric 에 기반해 motion blur 프레임을 제거함으로써 모든 비디오에서 3D point cloud 생성에 성공하였다. 저질 프레임 비율도 평균 5.5%까지 낮춘 것을 볼 수 있다. 제안하는 방법의 한 가지 제한 사항은 몇몇 파라미터의 값을 지정해야 하는데, 현재는 실험을 통해 경험적으로 값들을 결정하였습니다. 그러나 모든 파라미터는 영상에서 획득한 대응점의 수나 처리 시간 등과 관련되므로 몇 번의 실험을 통해서 특정 비디오 데이터에 적합한 파라미터를 결정할 수 있을 것입니다.

5. 결론

본 연구는 3 차원 복원에 적합한 프레임을 찾고 키 프레임 추출을 통해 비디오에서 불필요한 프레임을 제거하는 것에 중점을 둔다. 비디오에서 장면을 복원 할 때 제안하는 기법은 복원 성능을 향상시키고 계산 시간을 최소화한다. 실험 결과 모션 블러가 자주 발생하는 일반 비디오에서 저품질 프레임을 적절히 제거하는 것이 3D 복원에 효과적이라는 것을 보여준다. 향후에는 야외 환경 및 임의의 비디오 환경을 고려할 수 있는 연구가 수행될 수 있다. 이를 위해 움직이는 객체 분할, 샷 경계 검출 및 자동 교정과 같은 프로세스들과 연계할 필요가 있다.

ACKNOWLEDGEMENT

이 논문은 2017 년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.R0132-15-1005, 온-오프라인에서의 콘텐츠 비주얼 브라우징 기술개발)

6. 참고 문헌

[1] L. Ling, Ian S. Burrent, E. Cheng, "A dense 3D reconstruction approach from uncalibrated video sequences" ICMEW, pp. 587-592, 2012.

[2] J. Frahm, M. Pollefeys, S. Lazebnik, D. Gallup, B. Clipp, R. Raguram, C. Wu, C. Zach, T. Johnson, "Fast robust large-scale mapping from video and internet photo collections" ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 65, No. 6, pp. 538-549, 2010.

[3] M. Pollefeys et al., "Detailed real-time urban 3D reconstruction from video", International Journal of Computer Vision, Vol. 78, pp. 143-167, 2008.

[4] S. Gibson, J. Cook, T. Howard, R. Hubbold, D. Oram. "Accurate camera calibration for off-line, video-based augmented reality", in: IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR), Darmstadt, Germany, 2002.

[5] J.K. Seo, S.H. Kim, C.W. Jho, H.K. Hong, "3D Estimation and Key-Frame Selection for Match Move", ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications, pp. 1282-1285, July 2003.

[6] Y.H. Seo, S.H. Kim, K.S. Doo, J.S. Choi, "Optimal keyframe selection algorithm for three-dimensional reconstruction in uncalibrated multiple images", Journal of the Society of Photo-Optical Instrumentation Engineers, Vol. 47, No. 5, pp. 53201- 53400, 2008.

[7] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, R. Koch, "Visual modeling with a hand-held camera", International Journal of Computer Vision, Vol. 59, No. 3, pp. 207-232.

[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, pp. 485-491, 1998.

[9] M.T. Ahmed, M.N. Dailey, J.L. Landabaso, N. Herrero, "Robust key frame extraction for 3D reconstruction from video streams", in: Proc. The VISAPP, pp. 231-236, 2010.

[10] R. Fergus, B. Singh, A. Hertzmann, S.T. Roweis, W.T. Freeman, "Removing camera shake from a single photograph", ACM Transactions on Graphics, Vol. 25, No. 3, pp. 787-794, 2006.

[11] Y. Matsushita, E. Ofek, X. Tang, H.Y. Shum, "Full-frame video stabilization", in: Proc. Computer Vision and Pattern Recognition, pp. 50-57, 2005.

[12] S. Cho, J. Wang, S. Lee, "Video deblurring for hand-held cameras using patch-based synthesis", ACM Transactions on Graphics, Vol. 31, No. 4, pp. 1-9, 2012.

[13] S. Yang, M. Lizhuang, "Detecting and Removing the Motion Blurring from Video Clips", I.J.Modern Education and Computer Science, Vol. 1, pp. 17-23, January 2010.

[14] Changchang Wu, "Towards Linear-time Incremental Structure from Motion", in: Proc. International Conference on 3D Vision, pp. 127-134, 2013.

[15] Changchang Wu, "VisualSFM: A Visual Structure from Motion System", http://ccwu.me/vsfm/, 2011