다중 사용자 포즈 추정 및 트래킹 알고리즘의 구현

*김승렬 **안소윤 *서영호

*광운대학교

*kimsr_96@naver.com **ays613@naver.com *yhseo@kw.ac.kr

Development of Multi-Person Pose-Estimation and Tracking Algorithm

*Kim, Seung-Ryeol **Ahn, So-Yoon *Seo, Young-Ho

*Kwangwoon University

요약

본 논문은 3D 공간에서 사용자를 추출한 뒤, 체적 정보 분석을 통한 3D 스켈레톤(skeleton) 분석 과정을 통해 정확도 높은 다수 사용자의 위치 추적 기술에 대해 연구하였다. 이를 위하여 YOLO(You Only Look Once)를 활용하여 실시간으로 객체를 검출(Real-Time Object Detection)한 뒤 Google의 Mediapipe를 활용해 스켈레톤 추출, 스켈레톤 정규화(normalization)를 통한 스켈레톤의 크기 및 상대적 비율 계산, RGB 영상 스케일링(Scaling) 후 주요 마디 인접 영역의 RGB 색상 정보를 추출하는 방법을 통해 정확도가 개선된 높은 성능의 다중 사용자 추적 기술을 연구하였다.

1. 서론

자세 추정에는 크게 Top-down 방식과 Bottom-up 두 가지 방식이 있다. 이미지에서 사람의 명수를 먼저 파악한 뒤 각 사람에 해당하는 관절 위치를 추정하는 경우가 전자이고, 모든 관절 위치를 검출한 다음 검출된 관절의 위치와 방향을 토대로 연결해 각 사람을 추정하는 경우가 후자이다.[1]

하향 방식을 사용하는 PoseNet[2], RMPE[3], Mask R-CNN[4]은 사람의 명수에 따라 처리 시간이 급격히 증가하고, 사람을 잘못 검출할 시에 복구하기 어렵다는 단점이 있다. 기존의 상향 방식을 사용하는 Deepcut[5], OpenPose[6], MultiPoseNet[7] 또한 다수의 사람들이 검출될 경우 관절을 매칭하기 어렵고 이전의 알고리즘의 경우 마찬가지로 처리 시간이 증가하는 단점이 있다.

본 논문은 다중 사용자 추적 기술에 활용되는 YOLO, Mediapipe 기반의 개별 사용자 감지, 스켈레톤 추출 및 정규화, 주요 관절 인접 영역의 색상 정보 저장, 매 프레임 사용자의 이동 거리에 대한 기술적인 절차를 나타내고, 그에 따른 결과를 제시한다.

2. 제안한 알고리즘

다중 사용자 트래킹 기술로서 그림 1 순서의 방식을 제시한다. 첫째, YOLO를 사용한 기존 객체 감지 모델과 Mediapipe를 사용한 자세 추정 모델을 제시한 뒤, 개별적으로 감지된 Bounding box에 대해 스켈레톤을 추출 후 ID를 매칭하는 절차적 결과물을 제시한다. 둘째, RGB 영상을 스케일링해 스켈레톤을 정규화하고 상대적인 비율 정보를 저장, 주요 마디 인접 영역의 RGB 색상 정보를 저장한 결과물을 제시한다. 셋째, 위에서 저장된 비율 정보와 RGB 색상 정보, 감지된 박스의 위치 변화를 통해 다수의 사용자 중 한 명의 사용자를 특정할 수 있는지에 관한 결과를 ID 값으로써 제시한다. 마지막으로, 결과로서 제시된 특정된 ID 값을 가진 사용자와 실제 공간에서의 사용자가 일치하는 정도를 비교해 정확도와 최대 사용자 수를 척도로써 표현하고, 이에 대한 기대효과를 제시한다.

그림 1. 다중 사용자 트래킹 순서도

2.1. Object detection & Pose Estimation

본 단계에서는 YOLOv4를 사용해 감지된 사용자의 bounding box를 영상 내에 개별 RoI(Region of Interest)로 설정, Mediapipe를 사용해 각 Roi별로 스켈레톤을 추출함으로써 여러 사람에 대해 스켈레톤을 추출할 수 있도록 한다.

2.2. Get person information for Re-Identification

Id를 부여하고 추적하기 위하여 감지된 신체에서 관절 주변 색상 정보, 매 프레임 이동하는 거리를 추출한다.

Mediapipe로 신체의 11(left_shoulder), 12(right_shoulder), 13(left_elbow), 14(right_elbow), 23(left_hip), 24(right_hip), 25(left_knee), 26(right_knee) 총 8개의 각 관절 주변 3x3 픽셀값(neighborhood값)의 HSV 색 영역에서의 hue값, bounding box의 x, y, width, height에 해당하는 위치 정보값을 저장한다. HSV 색 공간은

명도와 채도를 제외한 hue값만의 추출이 가능하므로 불규칙적으로 명도나 채도의 변화가 나타날 수 있는 실시간 추적에서 효율적이다.

2.3. ID-matching

2.2절의 결과와 감지된 박스의 위치 변화로 다수의 사용자 중 한 명의 사용자를 특정할 수 있는지에 관한 결과를 ID 값으로써 제시한다. 결과값은 코사인 유사도(cosine similarity), 위치 변화(distance), hamming distance 형식으로 제공되며, 스켈레톤의 특정한 8개 landmark의 hue 값과 좌표를 통해 위치 정보를 프레임 단위로 갱신한다.

Hamming distance는 이전 프레임에서 저장되어 있던 특정 ID값을 가진 사람의 hue 값과 현재 프레임의 hue 값의 차이가 20 이상일 경우 1, 그 미만일 경우 0을 추가하는 배열을 만들어 최종값으로 배열 내의 값들을 반환하여 관절 색상 정보를 정규화한 배열이 얼마나 유사한지 판단한다. 코사인 유사도[11]를 사용하여 스켈레톤의 특정한 8개 landmark를 벡터로써 표현해 이전 프레임과 현재 프레임의 landmark의 유사도를 계산한다. Bounding box의 위치 변화는 유클리드 거리(Euclidean distance)(1)를 통해 구현했다.

$$\sqrt{(x_1 - x_2)^2} + \sqrt{(y_1 - y_2)^2} = d(x, y) \quad (1)$$

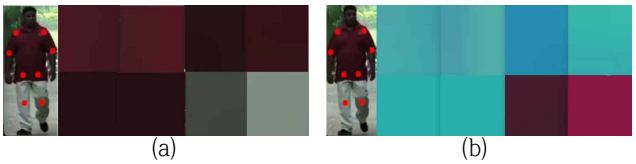
3. 실험 결과

실험 영상으로는 두 명이 포함된 영상을 사용했다. 그림 2(a)은 원본 영상, 그림 2(b)는 YOLOv4를 활용해 영상 내에 bounding box를 그린 결과, 그림 2(c)는 Mediapipe를 사용해 각각의 박스 내에서 자세 추정을 진행한 결과이다.

(a) (b) (c)

그림 2. (a) 원본 영상 (b) YOLOv4 모델 객체 검출 결과 (c) Mediapipe 자세 추정 결과

위의 결과에서 2.2절에 언급된 총 8개의 관절 주변 3x3 픽셀값의 hue값을 추출한다. 그림 3(a)는 추출된 픽셀의 RGB 형상, 2(b)는 HSV 형상을 표현한 결과값이다.



(a) (b)

그림 3. 관절 주변 3x3 픽셀 색 검출 결과 (a) RGB 색 공간에서의 검출 (b) HSV 색 공간에서의 hue값 추출 (좌측 상단부터 차례대로 landmark 11, 12, 13, 14, 23, 24, 25, 26)

2.3절의 코사인 유사도, 위치 변화, hamming distance로 최종 박스에 대한 ID값을 추정한다. 표 1은 예시 4개의 프레임에서의 코사인 유사도와 위치 변화값을 표시한 결과, 그림 4는 특정 사용자의 추적 결과이다.

Previd	Current id	Frame 1			Frame 2			Frame 3			Frame 4		
		cs	d	hd	cs	d	hd	cs	d	hd	cs	d	hd
0	0	1	1.41	0	0.99	1.41	1	0.97	2.24	2	0.98	0	1
0	1	0.70	155.12	5	0.75	158.32	6	0.71	157.26	5	0.76	157.1	6
1	0	0.72	153.12	5	0.72	153.12	6	0.80	156.16	6	0.71	156.16	5
1	1	0.99	1.41	1	0.98	5	3	0.99	0	0	0.99	2.24	1

cs = cosine similarity, d = euclidean distance, hd = hamming distance

표 1. 4개 프레임에서의 코사인 유사도와 위치 변화 값

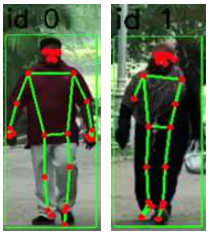


그림 4. 특정 사용자 트래킹 결과

결과로써 코사인 유사도 값이 높고 위치 변화값이 낮을 경우 기존 id와 현재 id가 같다고 추정되고, 추정된 ID값을 bounding box 상단에 표기해 특정 사용자 위치 추적이 가능하다.

4. 결론

다중 사용자 포즈 추정 및 트래킹을 위해 본 논문에서는 YOLOv4의 객체 감지와 Mediapipe를 각 사용자들에 대한 자세 추정 결과를 제시하고 스켈레톤의 상대적 비율 정보와 주요 마디 인접 영역의 색상 정보인 hue값을 추출, bounding box의 위치 변화 정보를 통해 구체화함으로써 다수 사용자에 대해 정확도 면에서 좋은 결과를 보인다는 것을 입증했다. 이는 한 사용자의 특징을 추출해 구체화함으로써 다중 사용자로부터 특정할 수 있다는 결과를 보여 수많은 군중 내에서 모든 사람에 대한 추적 가능성을 보였고, 앞으로의 추적 알고리즘에서의 정확도 향상에 대한 연구들에 도움을 줄 것으로 보인다.

5. 감사의 글

본 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2018R1D1A1B07043220).

6. 참고문헌

[1] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, Jingdong Wang, “Bottom-Up Human Pose Estimation via Disentangled Keypoint Regression”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14676-14686

[2] Alex Kendall, Matthew Grimes, Roberto Cipolla, “PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2938-2946

[3] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu, “RMPE: Regional Multi-Person Pose Estimation”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2334-2343

[4] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick, “Mask R-CNN”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961-2969

[5] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, Bernt Schiele, “DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4929-4937

[6] G. Hidalgo, Z. Cao, T. Simon, S.-E. Wei, H. Joo, Y. Sheikh, “OpenPose library”, https://github.com/CMU-Perceptual-Computing-Lab/openpose.

[7] M Kocabas, S Karagoz, E Akbas, “MultiPoseNet: Fast Multi-Person Pose Estimation using Pose Residual Network”, Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 417-433

[9] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection”, https://github.com/AlexeyAB/darknet

[10] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, Matthias Grundmann, “MediaPipe: A Framework for Building Perception Pipelines”, https://github.com/google/mediapipe

[11] Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi, “Semantic Cosine Similarity”, The 7th International Student Conference on Advanced Science and Technology (ICAST), 2012