

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH**

NGUYỄN THÀNH TRUNG

**KHÓA LUẬN TỐT NGHIỆP
ƯỚC LƯỢNG TƯ THẾ NGƯỜI
TRONG KHÔNG GIAN HAI CHIỀU**

CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH, 2020

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH**

NGUYỄN THÀNH TRUNG – 16521320

**KHÓA LUẬN TỐT NGHIỆP
ƯỚC LƯỢNG TỰ THẾ NGƯỜI
TRONG KHÔNG GIAN HAI CHIỀU**

CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH

**GIẢNG VIÊN HƯỚNG DẪN
TS. NGÔ ĐỨC THÀNH**

TP. HỒ CHÍ MINH, 2020

DANH SÁCH HỘI ĐỒNG BẢO VỆ KHÓA LUẬN

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số 523/QĐ-ĐHCNTT ngày 25 tháng 08 năm 2020 của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

1. Chủ tịch: PGS.TS. Lê Đình Duy
2. Thư ký: TS. Lương Ngọc Hoàng
3. Ủy viên: TS. Mai Tiến Dũng

LỜI CẢM ƠN

Lời nói đầu tiên, Tôi xin chân thành cảm ơn TS. Ngô Đức Thành đã tận tình hướng dẫn, giảng dạy và đóng góp ý kiến quý báu giúp chúng tôi có thể hoàn thành tốt khoá luận này.

Tôi cũng xin gửi lời cảm ơn đến các thầy cô, anh chị và bạn bè khoa Khoa học máy tính và các thành viên ở Phòng thí nghiệm Truyền Thông Đa phương tiện của trường Đại Học Công Nghệ Thông Tin đã hỗ trợ, tạo điều kiện thuận lợi nhất để Tôi hoàn thành khoá luận.

Cảm ơn các bạn lớp KHTN2016 đã luôn bên cạnh, động viên và ủng hộ Tôi trong suốt quá trình học tập và rèn luyện tại trường Đại học Công Nghệ Thông Tin.

Chúng tôi xin chân thành cảm ơn.

TPHCM, tháng 08 năm 2020.

ĐỀ CƯƠNG CHI TIẾT

TÊN ĐỀ TÀI: Ước lượng tư thế người trong không gian hai chiều	
TÊN ĐỀ TÀI TIẾNG ANH: 2D Human pose estimation	
Cán bộ hướng dẫn: TS. Ngô Đức Thành	
Thời gian thực hiện: Từ ngày 10/02/2020 đến ngày 20/07/2020.	
Sinh viên thực hiện:	
Nguyễn Thành Trung - 16521320	Lớp: KHTN2016
Email: 16521320@gm.uit.edu.vn	Điện thoại: 0979785754
Nội dung đề tài:	
Hiện nay với công nghệ phát triển một cách nhanh chóng, con người áp dụng công nghệ vào đời sống, ước lượng tư thế người trong không gian hai chiều cũng không ngoại lệ. Các phương pháp ước lượng tư thế người trong không gian hai chiều dựa trên học sâu gần đây đã chứng minh được tính ứng dụng và kết quả tốt hơn trước. Tuy nhiên, bài toán cũng gặp khó khăn nhiều thách thức khác nhau như trong đám đông, độ phân giải, ánh sáng,... Trong bài luận này, nhóm sinh viên phân tích ưu điểm và nhược điểm của bài báo “Pifaf: Composite Fields For Human Pose Estimation”, tác giả đã tập trung giải quyết thách thức che khuất trong đám đông và độ phân giải thấp. Chúng tôi phân tích ưu điểm và nhược điểm trên các bộ dữ liệu khác nhau: bộ dữ liệu COCO, ngoài ra chúng tôi phân tích các lỗi trên 1000 hình ảnh MPII, 2000 hình ảnh bộ dữ liệu thể thao do chúng tôi thu thập. Qua đó có cách nhìn tổng quan hơn về bài báo, và từ đó đưa ra các hướng phát triển nghiên cứu để cải thiện bài báo.	
Kế hoạch thực hiện:	

- 1. 10/02/2020 – 10/04/2020:** Khảo sát đề tài và các phương pháp liên quan đến bài toán. Thực hiện cài đặt lại phương pháp sẵn có.
- 2. 11/04/2020 – 25/04/2020:** Chính sửa lại các phương pháp đã lựa chọn với đầu vào là dữ liệu hình ảnh. Khảo sát những điểm hạn chế của phương pháp đó.
- 3. 26/04/2020 – 10/05/2020:** Phân tích lỗi và so sánh Pifpaf và openpose trên 1000 hình ảnh bộ dữ liệu COCO-val, đưa ra cái nhìn tổng quan về những thách thức mà pifpaf làm tốt, openpose làm tốt.
- 4. 10/05/2020 – 26/05/2020:** Thu thập 2,000 hình ảnh về thể thao và tiến hành phân tích các lỗi bài báo gặp phải trên các backbone model khác nhau và ở các độ phân giải khác nhau.
- 5. 27/05/2020 – 19/06/2020:** Tiếp tục phân tích kết quả trong bài báo và phân tích lỗi trên các backbone model khác nhau ở những độ phân giải khác nhau qua 4000 hình ảnh COCO test-dev, 1000 hình ảnh MPII.
- 6. 20/06/2020 – 20/07/2020:** Viết báo cáo khoá luận chi tiết và tiến hành tiếp các thí nghiệm cần thiết.

Xác nhận của CBHD	TP. HCM, ngày....thángnăm 2020 Sinh viên
--------------------------	---

Mục lục

TÓM TẮT KHOÁ LUẬN

xviii

1 TỔNG QUAN	1
1.1 Giới thiệu bài toán	1
1.2 Ứng dụng	3
1.2.1 Nhận diện hành động	3
1.2.2 Chụp chuyển động và thực tế ảo	4
1.2.3 Huấn Luyện robot	5
1.3 Thách thức của bài toán	6
1.4 Mục tiêu, nội dung thực hiện và đóng góp	9
1.4.1 Mục tiêu	9
1.4.2 Nội dung thực hiện	9
1.4.3 Đóng góp	10
1.5 Các công trình liên quan	10
1.5.1 Top down	10
1.5.2 Bottom Up	14
1.6 Cấu trúc Khóa luận tốt nghiệp	18
2 ƯỚC LƯỢNG TƯ THẾ NGƯỜI TRONG KHÔNG GIAN HAI CHIỀU	
SỬ DỤNG PifPaf	19
2.1 Tổng quan phương pháp	19
2.2 HeatMaps	21

2.3 Part Intensity Fields	22
2.4 Part Association Fields	27
2.5 Greedy Coding	30
3 THỰC NGHIỆM VÀ ĐÁNH GIÁ	33
3.1 Dữ liệu	33
3.2 COCO Dataset	34
3.3 Dữ liệu đánh giá hệ thống	37
3.4 Tiêu Chí đánh giá	41
3.4.1 Precision và Recall	41
3.4.2 Average Precision	42
3.4.3 Object Keypoint Similarity (OKS)	42
3.5 Đánh giá hệ thống	45
4 KẾT LUẬN	53
4.1 Kết luận	53

Danh sách hình vẽ

1.1	Minh họa về 17 joint trên cơ thể người	2
1.2	Minh họa bài toán ước lượng tư thế người trong không gian hai chiều	3
1.3	Ứng dụng: nhận diện hành động	4
1.4	Ứng dụng: Chụp chuyển động và thực tế ảo	5
1.5	Ứng dụng: Đào tạo Robot	5
1.6	Thách thức bài toán: Che khuất	6
1.7	Thách thức bài toán: Sự đa dạng trong tỉ lệ người	7
1.8	Lỗi từ các thách thức: Tư thế khó	7
1.9	Lỗi từ các thách thức: Missing part	8
1.10	Lỗi từ các thách thức: wrong connection	8
1.11	Qui trình hướng tiếp cận Top down	11
1.12	Qui trình mô phỏng hướng tiếp cận Top down	11
1.13	Công trình nghiên cứu liên quan theo hướng Top down: RMPE . .	12
1.14	Công trình nghiên cứu liên quan theo hướng Top down: Simple Baselines for Human Pose Estimation and Tracking	13
1.15	Công trình nghiên cứu liên quan theo hướng tiếp cận Top down: Simple Baselines for Human Pose Estimation and Tracking	14
1.16	Qui trình theo hướng tiếp cận Bottom Up	15
1.17	Minh họa mô phỏng về qui trình theo hướng tiếp cận Bottom Up . .	15
1.18	Công trình nghiên cứu liên quan hướng bottom up : Openpose . .	16
1.19	Công trình nghiên cứu liên quan hướng bottom up: PersonLab . .	17

2.1 Kiến trúc PifPaf	20
2.2 Minh họa ba kênh đầu ra trên mỗi keypoint	22
2.3 Minh họa quá trình tạo ra keypoint	24
2.4 Minh họa Confidence map của vai trái trên mỗi người trong hình ảnh	26
2.5 Minh họa Vector fields của joint (vai trái) trong hình ảnh	26
2.6 Kết quả của joint (vai trái) sau khi qua thành phần Pif	27
2.7 Ví dụ minh họa liên kết thành phần PAF	28
2.8 Minh họa 19 kết nối cho một người	30
2.9 Minh họa Greedy Coding	31
3.1 Panoptic segmentations	34
3.2 Object detection with segmentation masks	35
3.3 Person keypoint estimation	35
3.4 DensePose	35
3.5 COCO Dataset: Thông kê số người trong một ảnh	36
3.6 COCO Dataset: Thông kê số lượng keypoint trên một người trong ảnh	37
3.7 Minh họa về bộ dữ liệu COCO test-dev	38
3.8 Minh họa về bộ dữ liệu COCO test-dev	38
3.9 Minh họa về bộ dữ liệu thể thao	39
3.10 Minh họa về bộ dữ liệu thể thao	39
3.11 Minh họa về bộ dữ liệu MPII	40
3.12 Minh họa về bộ dữ liệu MPII	40
3.13 Biểu đồ quan hệ Precision và Recall	42
3.14 Các lỗi sử dụng cho phân tích và đánh giá phương pháp PifPaf bằng phương pháp thủ công(quan sát bằng mắt)	44
3.15 Minh họa kết quả phân tích 4000 hình ảnh MPII trên biểu đồ cột	48

3.16 Minh họa kết quả phân tích 4000 hình ảnh COCO test-dev trên biểu**đồ cột** 49**3.17 Minh họa kết quả phân tích 2000 hình ảnh thể thao trên biểu đồ cột** 51

Danh sách bảng

3.1	10 metrics đánh giá phương pháp tại các ngưỡng khác nhau	44
3.2	So sánh PifPaf với các phương pháp khác dựa trên kết quả thực nghiệm trên bộ dữ liệu COCO-val	45
3.3	So sánh PifPaf với các phương pháp khác dựa trên kết quả thực nghiệm trên bộ dữ liệu COCO-testdev	46
3.4	Kết quả thực nghiệm trên các hàm loss khác nhau	46
3.5	So sánh các model backbone với nhau	47
3.6	So sánh các model backbone với nhau	47
3.7	Phân tích thủ công các lối PifPaf trên 1000 hình ảnh MPII	48
3.8	Phân tích thủ công các lối PifPaf trên 4000 hình ảnh COCO testdev	49
3.9	Phân tích thủ công các lối PifPaf trên 2000 hình ảnh thể thao tự thu thập	50
3.10	thời gian chạy một ảnh trên các model backbone	52

Danh mục từ viết tắt

CNN	Convolutional Neural Network
Paf	Part association field
Pif	Part intensity field
IoU	Intersect over Union
NMS	Non-Max Suppression
COCO	Common Object in Context
AP	Average Precision
AR	Average Recall

TÓM TẮT KHOÁ LUẬN

Trong khoá luận này, chúng tôi nghiên cứu và phân tích bài toán ước lượng tư thế người trong không gian hai chiều. Đầu vào là một hình ảnh có ba kênh màu Red, Green, Blue và đầu ra là một hình ảnh mỗi người trong hình được ước lượng tư thế người. Đây là một bài toán được ứng dụng trong thực tế như phát hiện một người bị ngã, đóng vai trò như một huấn luyện viên cho các vận động viên thể thao tự tập ở nhà, phát hiện các động tác sai của yoga khi tự tập ở nhà với gương thông minh... Tại hội nghị CVPR2019 [8], bài báo “Pifpaf: Composite Fields for human pose estimation” đã đề xuất phương pháp ước lượng tư thế người trong không gian hai chiều. Đây là phương pháp hiện đại nhất hiện nay về sự che khuất trên độ phân giải thấp. Tác giả tập trung vào giải quyết hai thách thức chính đó là sự che khuất giữa người với người ở đám đông và độ phân giải thấp của hình ảnh. Phương pháp này là một phương pháp bottom up và nó gồm hai phần chính là Part Intensity Fields (PIF) phát hiện các joints (khuỷu tay, mắt,...) của mỗi người trong ảnh và Part Association Fields (PAF) dùng để kết nối hai joints với nhau. Tác giả đã thử nhiều backbone model khác nhau để trích xuất đặc trưng, thử nghiệm các hàm mắt mát khác nhau trong việc huấn luyện để chọn ra hàm mắt mát phù hợp nhất với phương pháp của mình. Qua đó chọn ra những phần phù hợp nhất cho kiến trúc của mình. Mục tiêu khoá luận: Đầu tiên chúng tôi tìm hiểu các công trình nghiên cứu liên quan đến ước lượng tư thế người trong không gian hai chiều, thứ hai chúng tôi chọn phương pháp “Pifpaf: Composite Fields for human pose estimation” để tiếp tục nghiên cứu, và tiếp theo đó là chúng tôi phân tích kết quả của phương pháp PifPaf do tác giả thử nghiệm, sau đó chúng tôi tiến hành phân tích ưu điểm và nhược điểm của pifpaf dựa vào các lỗi như wrong connection, missing part, no detect,... trên 4000 hình ảnh COCO test-dev, 1000 hình ảnh MPII và 2000 hình ảnh về thể thao như bóng đá, bóng chuyền, bóng bầu dục,... do chúng tôi thu thập, ngoài ra chúng tôi phân tích pifpaf trên nhiều

backbone model như Resnet50, shufflenetv2k16w, shufflenetv2k30w và trên các độ phân giải khác nhau: 256 pixel, 321 pixel, 641 pixel. Qua đó chúng tôi đưa ra cái nhìn tổng quan hơn về phương pháp Pifpaf và hướng phát triển nghiên cứu về phương pháp cho tương lai.

Chương 1

TỔNG QUAN

Trong chương này, chúng tôi sẽ giới thiệu về bài toán ước lượng tư thế người trong không gian hai chiều, những ứng dụng mà bài toán này có thể đem lại lợi ích cho con người và những thách thức mà bài toán gặp phải, những nhà nghiên cứu phải giải quyết để đem lại nhiều lợi ích cho cộng đồng nghiên cứu và áp dụng nó vào thực tế. Ngoài ra, để giải quyết thách thức thì chúng tôi cũng đề cập hai hướng tiếp cận chính đó là Top Down và Bottom Up và chúng tôi trình bày về các phương pháp liên quan đến bài toán này. Bằng cách tham khảo các phương pháp đã có, chúng tôi đưa ra cách tiếp cận, mục tiêu và phương pháp thực hiện của khoá luận.

1.1 Giới thiệu bài toán

Trước khi vào khái niệm của bài toán, chúng tôi giới thiệu sơ lược về khái niệm joint. Ở đây joint là một trong những điểm chính để miêu tả những bộ phận của cơ thể người (mắt trái, mũi, khuỷu tay,...) thì trên cơ thể sẽ có 17 joint theo định nghĩa mới nhất của COCO (hình 1.1).

⁰Ảnh tự thiết kế



HÌNH 1.1: Minh họa 17 joint trên cơ thể một người theo định nghĩa của COCO, Mỗi điểm xanh trên ảnh gọi là một joint.

Ước lượng tư thế người trong không gian hai chiều là ước tính các joint của một hoặc nhiều cơ thể trong không gian hai chiều từ một hình ảnh và sau đó các joint sẽ được kết nối với nhau để tạo thành một khung xương để mô tả tư thế người.

- Đầu vào: là một bức ảnh có ba kênh màu Red, Green, Blue.
- Đầu ra: một bức ảnh chứa các joint được liên kết với nhau để tạo thành tư thế người.

⁰Ảnh được lấy từ yoga journal



(a) Đầu vào



(b) Đầu ra

HÌNH 1.2: Minh họa bài toán. Mỗi điểm xanh trên ảnh đầu ra gọi là một joint. joint là một trong những điểm chính trên cơ thể người (mắt trái, khuỷu tay,...)

1.2 Ứng dụng

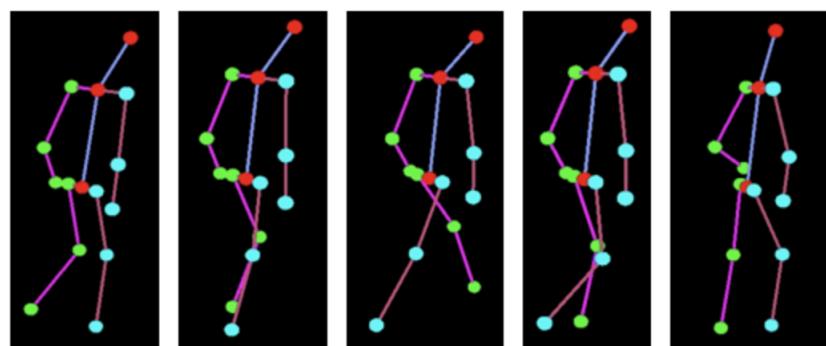
Hiện nay công nghệ phát triển một cách chóng mặt trên toàn thế giới và con người áp dụng công nghệ vào trong đời sống như: xe tự lái, nhận diện khuôn mặt.... Thì một trong các kỹ thuật và phương pháp để tạo nên những công nghệ như thế là ước tính tư thế người trong không gian hai chiều. Một trong số những ứng dụng trong đời sống mà phương pháp ước lượng tư thế người trong không gian hai chiều đem lại lợi ích cho con người là.

1.2.1 Nhận diện hành động

Theo dõi các chuyển động của các tư thế của một người trong khoảng thời gian nhất định là một ứng dụng thực tế mà bài toán ước lượng tư thế người được sử dụng để nhận dạng hoạt động, cử chỉ và dáng đi. Một số trường hợp được ứng dụng bao gồm:

- Ứng dụng để phát hiện nếu một người bị ngã hay không.

- Các ứng dụng về thay thế một huấn luận về cho các vận động viên thể thao và khiêu vũ nếu muốn tập thêm ở nhà, hay bất cứ nơi đâu.
- Ứng dụng về phát hiện tư thế sai trong yoga thông qua gương thông minh nếu tự tập ở nhà.
- Các ứng dụng có thể tăng cường bảo mật và giám sát.



HÌNH 1.3: Theo dõi dáng đi của người trong một thời gian nhất định
phụ vụ cho mục đích an ninh và giám sát¹

1.2.2 Chụp chuyển động và thực tế ảo

Một ứng dụng thú vị của ước tính tư thế người trong không gian hai chiều là dành cho các ứng dụng Computer generated imagery (CGI), hay Gaming. Đồ họa, kiểu dáng, cải tiến lõi mắt, thiết bị và các tác phẩm nghệ thuật có thể được đặt lên trên người nếu tư thế con người của họ có thể được ước tính, thay vì chơi game bằng bàn phím hay tay cầm thì có thể thay đổi bằng các chuyển động của tay, chân, và các bộ phận khác của người. Bằng cách theo dõi các biến thể khác nhau về tư thế của con người này, đồ họa được kết xuất có thể phù hợp với người đó khi họ di chuyển.

¹Ảnh được lấy từ Internet

²Ảnh được lấy từ Internet



HÌNH 1.4: ví dụ của CGI rendering²

1.2.3 Huấn Luyện robot

Thay vì lập trình cho robot chuyển động đi theo quỹ đạo, robot có thể được chế tạo để đi theo quỹ đạo của bộ xương người đang thực hiện một hành động. Con người có thể dạy robot một cách hiệu quả một số hành động nhất định chỉ bằng cách thể hiện tương tự. Sau đó, robot có thể tính toán làm thế nào để di chuyển các khớp nối của mình để thực hiện hành động tương tự.



HÌNH 1.5: Ứng dụng trong Robotics³

1.3 Thách thức của bài toán

Hiện nay với sự phát triển của công nghệ, đặc biệt là áp dụng mạng học sâu để giải quyết bài toán ước lượng tư thế người trong không gian hai chiều đem lại hiểu quả đáng kể. Bên cạnh đó, bài toán ước lượng tư thế người ứng dụng vào thực tế cũng gặp một số thách thức nhất định như: tính đa dạng trong tỉ lệ người, sự che khuất giữa người với nhau, đám đông,...

Chúng tôi liệt kê một số thách thức ảnh hưởng đến ước tính tư thế người trong không gian hai chiều.

Che khuất (occlusion): hình ảnh bị che khuất dễ gây ra những liên kết sai.



HÌNH 1.6: Thách thức bài toán: Che khuất giữa nhiều người với nhau

4

Tính đa dạng trong tỉ lệ người (Variance in person scales): sự đa dạng trong tỉ lệ người có thể gây khó khăn cho việc ước lượng các joint.

⁴Ảnh được lấy từ fanpage Leicester City Football Club trên facebook

⁵Ảnh được lấy từ Internet



HÌNH 1.7: Thách thức bài toán: Sự đa dạng của tỉ lệ người⁵

Dựa vào các thách thức bài toán ở trên thì chúng tôi liệt kê một số lỗi thường gặp trong bài toán ước lượng tư thế người trong không gian hai chiều như: tư thế khó (hình 1.8), nhầm lẫn các khớp với nhau trên một người (hình 1.9), nhầm lẫn các khớp giữa hai người trở lên (hình 1.10), ...



HÌNH 1.8: Tư thế khó. Khi tư thế khó thì việc ước lượng joint gấp khó khăn như trong hình là thiếu các joint và dễ gây ra những liên kết sai.⁶



HÌNH 1.9: Lỗi missing part:Nhầm giữa các joint với nhau trên một người.⁷



HÌNH 1.10: Lỗi Wrong connection: Nhầm lẩn giữa các joint với nhau từ hai người trở lên, joint người này nối qua joint người khác tạo thành liên kết sai.⁸

⁶Ảnh được lấy từ Website Yoga Journal

⁷Ảnh được lấy từ Ảnh được lấy từ fanpage Manchester United trên facebook

1.4 Mục tiêu, nội dung thực hiện và đóng góp

1.4.1 Mục tiêu

Khi thực hiện khoá luận chúng tôi đặt ra các mục tiêu sau:

- Tìm hiểu về phương pháp “PifPaf: Composite Fields for Human Pose Estimation”.
Phương pháp này được đề xuất tại hội nghị CVPR2019.
- Phân tích ưu điểm và nhược điểm của phương pháp.
- Phân tích chi tiết các lỗi mà PifPaf gặp phải ở các backbone model như Resnet50, shufflenet và thử nghiệm phương pháp ở nhiều độ phân giải khác nhau như: 256 pixel, 321 pixel, 641 pixel.
- Đánh giá phương pháp trên các tập dataset: COCO, MPII, bộ dữ liệu tự tạo.

1.4.2 Nội dung thực hiện

- Tìm hiểu bộ dữ liệu COCO, MPII phục vụ cho quá trình đánh giá.
- Tìm hiểu phương pháp PifPaf mà bài báo đã đề xuất tại hội nghị CVPR2019.
Cụ thể là tìm hiểu hai phương pháp Part Intensity Field (PIF) và Part Association Fields (PAF).
- Sau khi tìm hiểu, tiến hành thực hiện lại phương pháp PifPaf.
- Phân tích chi tiết các lỗi mà phương pháp thường gặp trên các bộ dữ liệu khác nhau, các độ phân giải khác nhau để đưa ra nhận xét tổng quan về hướng tiếp cận cải thiện phương pháp.

⁸Ảnh được lấy từ [Ảnh được lấy từ fanpage Liverpool trên facebook](#)

1.4.3 Đóng góp

Sau quá trình thử nghiệm và đánh giá, khoá luận chúng tôi có một số đóng góp như sau đóng góp như sau:

- Tìm hiểu phương pháp PifPaf.
- Phân tích chi tiết các lỗi mà PifPaf gặp phải ở những độ phân giải khác nhau như: 256 pixel, 321pixel, 641 pixel và các mô hình khác nhau như ResNet50, Shufflenetv2k16w, Shufflenetv2k30w.
- Dựa trên kết quả đánh giá thì chúng tôi đã đề xuất một số giả thiết về phương pháp PifPaf và hướng tiếp cận để phát triển và đạt được những kết quả tốt hơn trong tương lai.

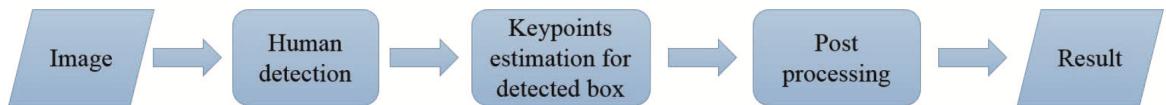
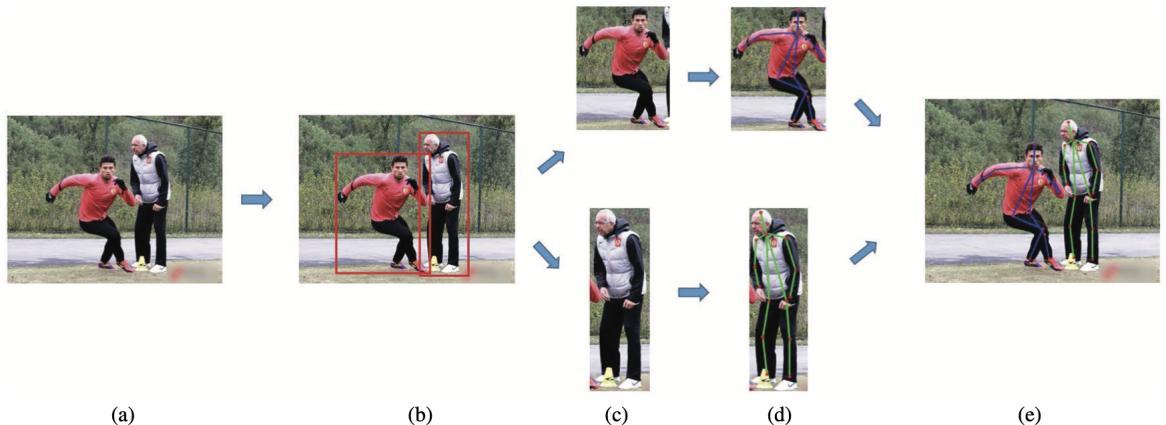
1.5 Các công trình liên quan

Dựa vào những thách thức đề cập phía trên và để giải quyết những thách thức này sẽ có hai hướng tiếp cận chính là Top down và Bottom up.

1.5.1 Top down

Phương pháp Top down là một trong những cách tiếp cận của bài toán ước lượng tư thế người trong không gian hai chiều. Đầu tiên là sử dụng kỹ thuật nhận diện con người để phát hiện tất cả người trong hình ảnh, sau đó mỗi người được gán nhãn bởi một bounding box. Tiếp đó ước tính tư thế người cho mỗi bounding box chứa người, cuối cùng là tổng hợp lại cho ra kết quả cuối cùng. Hình [1.11] thể hiện qui trình của cách tiếp cận Top down. Một số phương pháp theo hướng Top Down: PoseNet [12], CFN [6], Mask R-CNN [5], gần đây là CPN [2], MSRA [16], Pose Nerial Fabrics Search [17].

⁹Ảnh được lấy từ bài báo [3]

HÌNH 1.11: Qui trình hướng tiếp cận Top down⁹HÌNH 1.12: Minh họa chi tiết quy trình theo hướng tiếp cận Top down. (a) Hình ảnh đầu vào,(b) hai người được phát hiện bởi human detector, (c) cắt xén hình ảnh một người, (d) kết quả phát hiện tư thế của một người và (e) kết quả phát hiện tư thế nhiều người.¹⁰

RMPE: Regional Multi-Person Pose Estimation.

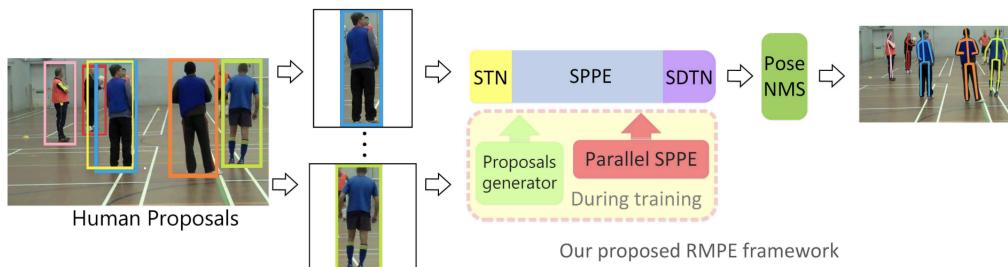
RMPE là nghiên cứu của Hao-Shu Fang và cộng sự của mình, bài toán được đề xuất ở hội nghị ICCV2017 [4].

RMPE là một trong những phương pháp Top down phổ biến của ước lượng tư thế người trong không gian hai chiều. Tác giả cho rằng các phương pháp theo hướng tiếp cận Top down thường phụ thuộc vào độ chính xác nhận diện người. Do đó các lỗi thường xảy ra ở dự đoán hộp giới hạn chồng lên nhau.

Để giải quyết vấn đề này, tác giả đã đề xuất một kiến trúc mạng Region multi person pose estimation (RPME) để tạo điều kiện cho việc ước tính tư thế người trong một bounding box chưa chính xác. Phương pháp gồm ba phần: Đầu tiên là Symmetric Spatial Transformer Network (SSTN) để trích xuất chính xác vùng có

¹⁰Ảnh được lấy từ bài báo [3]

người ở những hộp giới hạn không chính xác và Parametric Pose NonMaximum-Suppression (NMS) để giải quyết vấn đề redundant detection và cuối cùng là Pose-Guided Proposals Generator (PGPG) để tăng cường dữ liệu huấn luyện. Phương pháp này tập trung vào cải thiện hai vấn đề là inaccurate bounding box và redundant detection. Phương pháp đạt được kết quả 76.7 mAP trên bộ dữ liệu MPII.



HÌNH 1.13: Qui trình của kiến trúc RMPE. Symmetric STN bao gồm STN gắn trước SPPE và SDTN sau SPPE. STN nhận kết quả bounding box chứa người và SDTN tạo ra các đề xuất về tư thế cho người. SPPE hoạt động như là một bộ điều chỉnh bổ sung để đạt kết quả tốt hơn trong quá trình huấn luyện. Cuối cùng, Pose NMS (p-Pose NMS) được thực hiện để loại bỏ các tư thế không cần thiết. Không giống như những huấn luyện truyền thống tác giả huấn luyện SSTN+SPPE với hình ảnh từ PGPG.¹¹

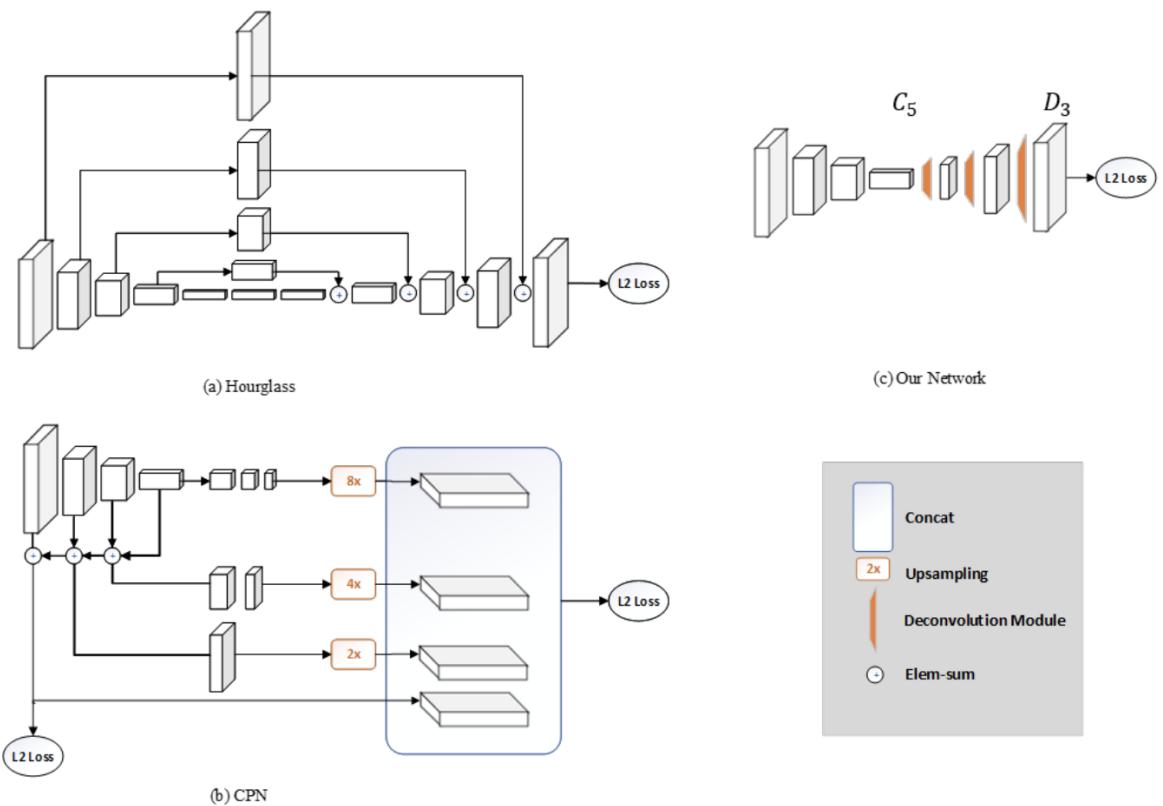
Simple Baselines for Human Pose Estimation and Tracking.

Phương pháp này là nghiên cứu của Bin Xiao và cộng sự của mình, bài báo được đề xuất ở hội nghị ECCV2018 [16].

Trong những năm gần đây, đã có những tiến bộ đáng kể về kết quả của bài toán ước lượng tư thế người, đồng thời thì độ phức tạp của thuật toán và các hệ thống cũng tăng theo, làm cho việc phân tích và so sánh các thuật toán và phương pháp trở nên khó khăn hơn. Để giải quyết vấn đề trên tác giả đã đề xuất một phương pháp đơn giản và hiệu quả hơn.

¹¹Ảnh được lấy từ bài báo [4]

Tác giả đã thêm một vài lớp deconvolutional vào giai đoạn tích chập cuối cùng trong kiến trúc ResNet. Kiến trúc này rất dễ để tạo ra các bản đồ nhiệt từ hình ảnh có độ phân giải thấp. Ba lớp deconvolutional với batch normalization và RELU activation được sử dụng mặc định Hình 1.14 minh họa cho qui trình của phương pháp.

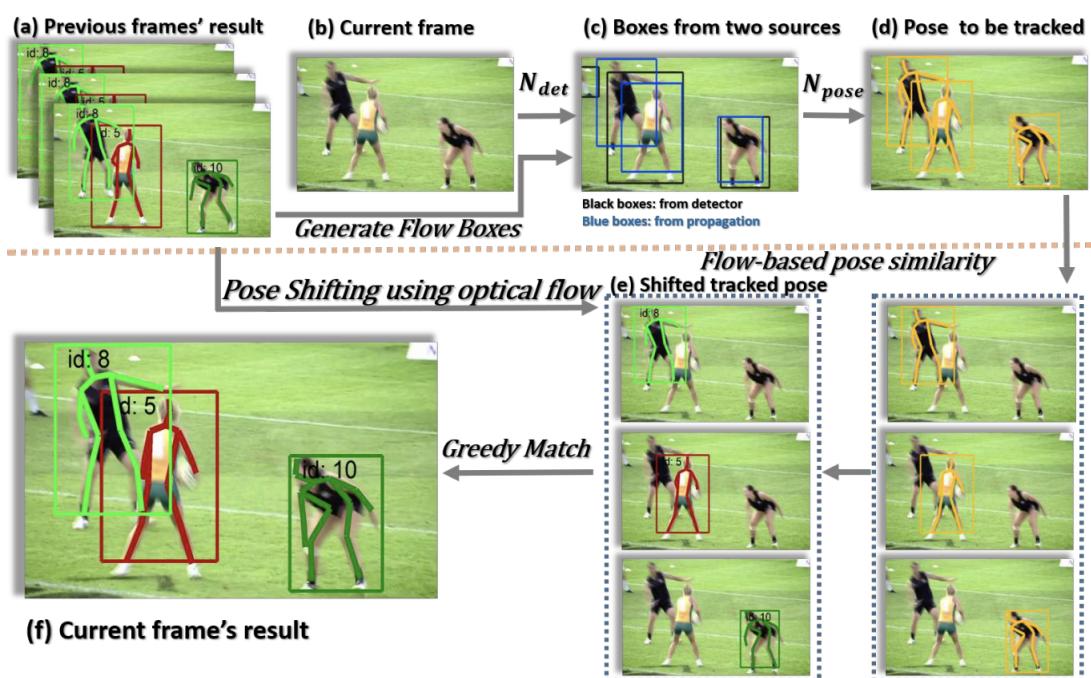


HÌNH 1.14: Minh họa của hai kiến trúc mạng hiện đại để ước tính tọa độ (a) một giai đoạn trong Hourglass [22], (b) CPN [6] và hệ thống đơn giản của tác giả (c).¹²

¹²Ảnh được lấy từ bài báo [16]

Hình 1.15 thể hiện quá trình của khung theo dõi tư thế. Theo dõi tư thế trong video được thực hiện bằng cách trước tiên ước tính tư thế của người và gán cho mỗi tư thế một mã định danh duy nhất và sau đó theo dõi nó qua các khung.

Phương pháp này đã đạt được kết quả tốt nhất với mAP là 74.6, điểm MOTA 57.8 và cải thiện kết quả 15% và 6% so với 59.6 và 51.8 của người dành chiến thắng ở cuộc thi ICCV'17 PoseTrack Challenge [14] [13].



HÌNH 1.15: Minh họa quá trình của khung theo dõi tư thế. ^[13]

1.5.2 Bottom Up

Phương pháp Bottum Up là một cách tiếp cận khác của ước lượng tư thế người trong không gian hai chiều. Đầu tiên phương pháp ước lượng các tất cả joint của con người trong hình ảnh, sau đó gom nhóm các joint từng người và tiếp theo là các joint kết nối với nhau để tạo thành tư thế người. Để dễ hình dung hơn về

¹³Ảnh được lấy từ bài báo [16]

phương pháp này chúng ta (xem hình 1.16) qui trình theo hướng tiếp cận Bottom Up và (1.17) minh họa chi tiết hơn về qui trình theo hướng tiếp cận Bottom Up. Một số phương pháp theo hướng Bottom Up như: Associative Embedding [9] and PersonLab [11], Deep High-Resolution Representation Learning for Human Pose Estimation [15], DeepCut [], Distribution-Aware Coordinate Representation for Human Pose Estimation [18].



HÌNH 1.16: Qui trình theo hướng tiếp cận Bottom Up¹⁴



HÌNH 1.17: Minh họa về quy trình theo hướng tiếp cận Bottom Up.
 (a) Hình ảnh đầu vào,(b) ước lượng keypoint cho tất cả mọi người trong hình và (c) tất cả các keypoint được phát hiện được kết nối để tạo thành tư thế con người.¹⁵

Openpose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.

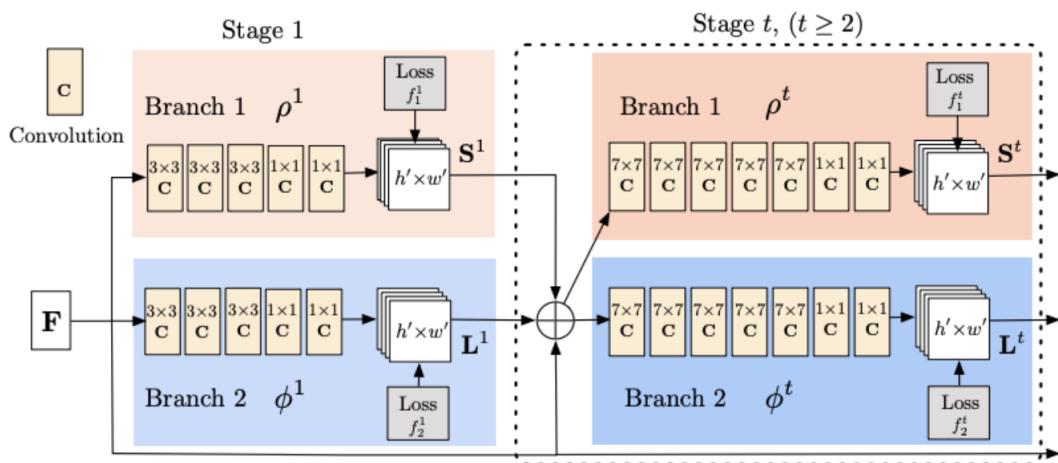
Openpose là nghiên cứu của Zhe Cao và cộng sự của mình, bài báo được đề xuất ở hội nghị [1]

Openpose là một trong những phương pháp Bottum-Up phổ biến nhất cho ước lượng tư thế người trong không gian hai chiều.

¹⁴Ảnh được lấy từ bài báo [3]

¹⁵Ảnh được lấy từ bài báo [3]

Kiến trúc mạng đầu tiên của openpose trích xuất các đặc trưng từ ảnh bằng cách sử dụng một vài lớp đầu tiên của VGG-19. Tiếp đó các đặc trưng này sẽ được đưa vào 2 nhánh song song của các lớp tích chập. Nhánh đầu tiên dự đoán các tập hợp của bản đồ tin cậy, với mỗi bản đồ tin cậy đại diện cho mỗi joint của tư thế người. Nhánh thứ hai dự đoán một tập hợp Part Affinity Fields (PAFs) đại diện cho mức độ liên kết giữa các phần với nhau.



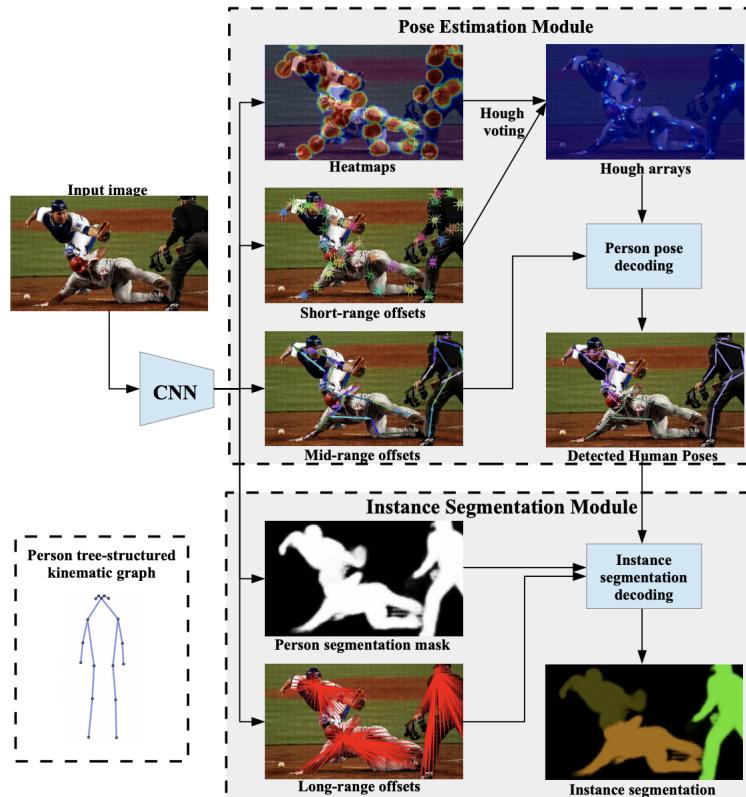
HÌNH 1.18: Kiến trúc openpose¹⁶

PersonLab: Person Pose Estimation and Instance Segmentation with Bottom-up, Part-Based, Geometric Embedding Model.

PersonLab là nghiên cứu của George Papandreou và cộng sự của mình, bài báo được đề xuất ở hội nghị ECCV2018 [11]. Personlab là một phương pháp theo hướng tiếp cận Bottum Up. Tác giả đã đề xuất mô hình PersonLab để giải quyết hai vấn đề là sementic-level reasoning và object-part association bằng cách sử dụng mô hình part-based. Mô hình sử dụng một mạng tích chập để phát hiện tất cả keypoint của người trong hình ảnh và dự đoán mối liên kết giữa các keypoint với nhau, sau đó là gom nhóm những keypoint cùng người với nhau tạo thành tư thế từng người trong hình ảnh. (hình 1.19) Phương pháp này đạt được kết quả

¹⁶Ảnh được lấy từ bài báo [1]

0.665 average precision (AP) trên COCO test-dev keypoint sử dụng trạng thái single-scale và 0.687 với trạng thái multi-scale.



HÌNH 1.19: Qui trình hệ thống Personlab. Hệ thống PersonLab bao gồm một mô hình CNN dự đoán: (1) keypoint heatmaps,(2) short-range offsets ,(3) mid-range pairwise offsets, (4) person segmentation maps và (5) long-range offsets. Ba dự đoán đầu tiên được sử dụng Pose Estimation Module để phát hiện tư thế của con người trong khi hai dự đoán sau cùng với việc phát hiện tư thế con người, được sử dụng Instance Segmentation Module để dự đoán mặt nạ phân đoạn cá thể.¹⁷

¹⁷Ảnh được lấy từ bài báo [11]

1.6 Cấu trúc Khóá luận tốt nghiệp

Phần còn lại của khoá luận được tổ chức như sau. Chương 2, trình bày về cấu trúc và phương pháp ước lượng tư thế người trong không gian hai chiều. Cụ thể là trình bày về phương pháp PifPaf: Composite Fields For Human Pose Estimation. Chương 3, chúng tôi trình bày về ưu điểm và nhược điểm của phương pháp, kết quả thực nghiệm và đánh giá kết quả đạt được. Chương 4, Từ đó chúng tôi đưa ra kết luận và hướng nghiên cứu trong tương lai.

Một trong những thách thức lớn nhất của bài toán hiện nay là sự che khuất giữa người với nhau. Phương pháp hiện nay đạt kết quả tốt nhất về sự che khuất giữa người với nhau, Đó là phương pháp PifPaf, có tên đầy đủ là "PifPaf: Composite Fields For Human Pose Estimation". Đây là phương pháp hiện đại nhất về sự che khuất giữa người với nhau trên độ phân giải thấp. Phương pháp này theo hướng tiếp cận Bottom Up, không theo hướng tiếp cận Top Down bởi vì độ che khuất giữa người với nhau thì hướng tiếp cận Top Down có những bounding box chồng lên nhau dễ gây ra các liên kết sai.

Chương 2

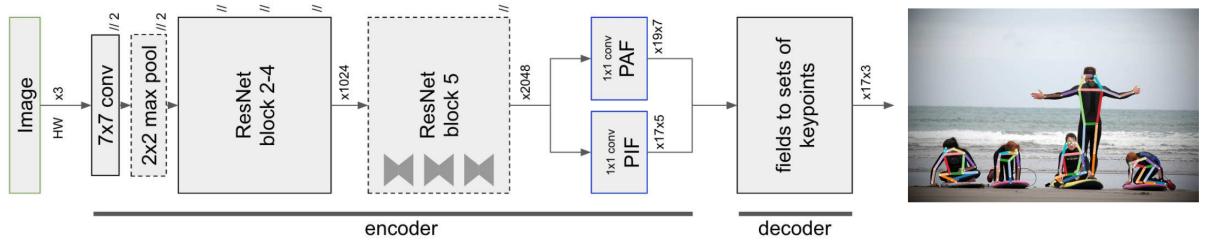
ƯỚC LƯỢNG TƯ THẾ NGƯỜI TRONG KHÔNG GIAN HAI CHIỀU SỬ DỤNG PifPaf

Trong chương này chúng tôi sẽ giới thiệu về phương pháp PifPaf và bài báo "PifPaf: Composite Fields for human pose estimation" đã đề xuất tại hội nghị CVPR2019 [8]. Cụ thể là chúng tôi trình bày chi tiết về hai phần chính của PifPaf là: Part Intensity Fields và Part Association Fields.

2.1 Tổng quan phương pháp

Những công trình nghiên cứu gần đây tập trung và giải quyết các thách thức khác nhau của bài toán ước lượng tư thế người trong không gian hai chiều. Những phương pháp hiện đại gần đây dựa trên CNN [[5], [1], [10], [11]] để giải quyết các thách thức trên. Thách thức về che khuất ở đám đông là một thách thức lớn của bài toán ước lượng tư thế người trong không gian hai chiều, Cũng giống như những phương pháp trên phương pháp PifPaf cũng dựa trên CNN để tiếp cận bài toán, Mục tiêu của phương pháp PifPaf [8] là ước lượng tư thế người trong môi trường đám đông và phương pháp giải quyết những thách thức liên

quan đến độ phân giải thấp của hình ảnh và che khuất giữa người với người trong đám đông. Để giải quyết thách thức về che khuất trong đám đông thì hướng tiếp cận Top down gặp khó khăn khi trong môi trường đông người thì nhiều người đi bộ che khuất nhau, và khi đó các bounding box sẽ chồng lên nhau, điều đó sẽ gây trở ngại lớn cho việc ước lượng các tư thế người. Còn về hướng tiếp cận Bottom up thì có kết quả khả thi hơn trong thách thức về che khuất giữa người với người trong đám đông, Những phương pháp bottom up trước đây không sử dụng bounding box nhưng kết quả về vị trí của các joint chưa tốt. Phương pháp PifPaf, tác giả không sử dụng ô lưới cố định để xác định vị trí các joint và thêm vào đó là tác giả đề cập đến phương pháp PifPaf có khả năng ước lượng các tư thế che khuất nhau. Hình 2.1 trình bày tổng quan về hệ thống được đề xuất. Hệ thống bao gồm kiến trúc ResNet50 kết hợp với hai thành phần : Part intensity fields (PIF) và Part association fields (PAF).



HÌNH 2.1: Kiến trúc PifPaf. Đầu vào là một hình ảnh có kích thước (H,W) với ba kênh màu được biểu thị bằng cách $x3$. encoder dựa trên mạng nơ-ron tạo ra các trường PIF và PAF với các kênh 17×5 và 19×7 . Một hoạt động với bước nhảy là 2 được biểu thị bằng $x2$. Decoder là chương trình chuyển đổi các trường PIF và PAF thành các ước tính tư thế chứa 17 joints trên mỗi người. Mỗi joint được biểu diễn bằng toạ độ x,y và độ tin cậy.¹

Đầu vào: hình ảnh gồm 3 kênh màu Red, Green, Blue.

Đầu ra: Ước lượng khung xương cho người trong hình ảnh.

Phương pháp PifPaf gồm 2 thành phần là encoder và decoder. Đầu tiên Đầu vào sẽ qua mô hình ResNet50 để trích xuất đặc trưng từ hình ảnh, sau đó qua hai

¹Ảnh được lấy từ bài báo [8]

phần chính, đây cũng là hai phần mà tác giả đề xuất để giải quyết thách thức độ che khuất giữa người với nhau trên độ phân giải thấp. Đó là Pif và Paf, Pif có tên đầy đủ là Part intensity fields có nhiệm vụ là ước lượng tất cả joint trong hình (phát hiện vị trí, kích thước joint, ngoài ra còn dự đoán độ tin cậy của joint), Paf có tên đầy đủ là Part association field và có nhiệm vụ là học các liên kết giữa các joint với nhau. Tiếp theo decoder này sẽ tổng hợp kết quả của hai phần Pif và Paf lại với nhau để chuyển đổi sang 17 joint tạo thành tư thế cho người trong hình.

Trong phần tiếp theo chúng tôi sẽ mô tả chi tiết hơn về cấu trúc của Part Intensity Fields và Part Association Fields.

2.2 HeatMaps

Một trong những thành phần quan trọng nhất đối với bài toán ước lượng tư thế người trong không gian hai chiều là vị trí của keypoint, là điểm mà được xem xét đầu tiên trong bài toán ước lượng tư thế người trong không gian hai chiều bởi vì vị trí keypoint ảnh hưởng đáng kể đến hiệu suất của các thuật toán.

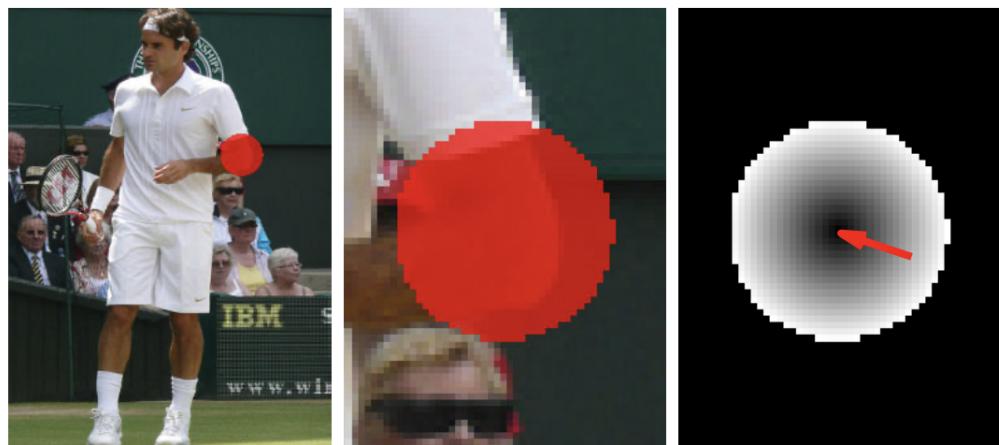
Hiện nay, Có ba cách để tạo ra bản đồ nhiệt:

- Cách đầu tiên là tạo ra bản đồ nhiệt với 2D Gaussian activation trên mỗi vị trí keypoint.
- Cách thứ hai là các pixel nằm trong bán kính bản đồ nhiệt thì sẽ có giá trị là 1 với tâm là vị trí keypoint và bán kính R, còn lại các giá trị pixel ngoài bán đồ nhiệt sẽ có giá trị là 0. Khi bản đồ nhiệt được thiết lập, thì các vị trí offset được dự đoán để xác định vị trí các keypoint chính xác hơn.
- Cách thứ 3 để tạo ra bản đồ nhiệt là one-hot binary mask, trong đó chỉ có một pixel được gán nhãn là chứa thông tin quan trọng nhất.

2.3 Part Intensity Fields

Part intensity fields dùng để phát hiện vị trí và kích thước của một joint, ngoài ra còn dự đoán độ tin cậy của joint. Với sự kết hợp của confidence map với tính năng hồi qui cho việc phát hiện keypoint được giới thiệu ở bài báo [12].

Bài báo [12], tác giả đã sử dụng phương pháp kết hợp giữa phân loại và hồi quy. Đối với mỗi vị trí không gian, Đầu tiên là kiểm tra xem tại vị trí này có nằm trong vùng lân cận của mỗi keypoint K hay không (được gọi là bản đồ nhiệt), sau đó dự đoán vị trí offset để giúp cho việc ước tính các keypoint chính xác hơn về vị trí của mỗi keypoint. Hình 2.2 minh họa ba kênh đầu ra này trên mỗi keypoint.



HÌNH 2.2: Kết quả đầu ra kiến trúc mạng. Bên Trái: Ở Giữa: Mục tiêu bản đồ nhiệt cho keypoint ở khuỷu tay trái (màu đỏ cho biết heatmaps là 1). Bên Phải: Độ lớn offset fields L2 (hiển thị ở thang độ xám) và 2D offset vector được hiển thị bằng màu đỏ).²

Tác giả sử dụng mô hình ResNet 101 đã được huấn luyện trước có sẵn từ Imagenet và chỉ thay thế lớp cuối cùng của mô hình bằng 1x1 convolution và đầu ra là 3K (K là Keypoint). Khởi tạo mô hình này trên ảnh đã được xử lý trước đó để tạo ra bản đồ nhiệt (một kênh là cho keypoint) và (hai kênh còn lại là trên mỗi keypoint cho vị trí x,y) với đầu ra là 3K (K=17 keypoint)

²Ảnh được lấy từ bài báo [12]

Có thể hiểu là tác giả đã sử dụng cách thứ hai ở trên để tạo ra bản đồ nhiệt là các pixel nằm trong vòng tròn thì sẽ có giá trị là 1 với tâm là vị trí keypoint và bán kính R, còn lại các giá trị pixel ngoài vòng tròn sẽ có giá trị là 0. Khi bản đồ nhiệt được thiết lập, thì các vị trí offset được dự đoán để xác định vị trí các keypoint chính xác hơn.

Ngoài bản đồ nhiệt, Tại mỗi vị trí i và mỗi keypoint k , tác giả dự đoán một vector offset $F_k(x_i) = l_k - x_i$ từ mỗi vị trí không gian đến keypoint tương ứng. Qua đó, tác giả tạo ra K như vector fields, để giải quyết vấn đề hồi quy trên từng vị trí và keypoint một cách độc lập.

Sau khi tạo ra bản đồ nhiệt và vị trí offset, tiếp đó tác giả kết hợp cả hai lại để tạo ra bản đồ kích hoạt vị trí $f_k(x_i)$:

$$f_k(x_i) = \sum_j \frac{1}{\pi R^2} G(x_j + F_k(x_j) - x_i) h_k(x_j) \quad (2.1)$$

Trong đó:

R : bán kính bản đồ nhiệt

G : bilinear interpolation kernel

h_k : xác suất của x_j để xác định x_j nằm trong hay ngoài bán kính R

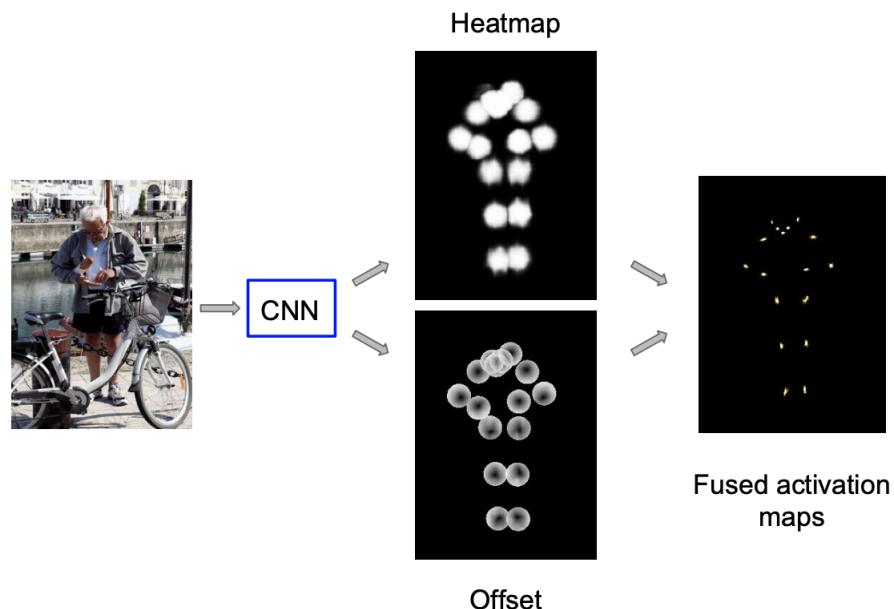
x_j : vị trí trong không gian

$F_k(x_j)$: 2-D offset vector

$G(\cdot)$ là bilinear interpolation kernel. Đây là một dạng bỏ phiếu Hough: với mỗi điểm j trong phần lưới cắt đại diện một phiếu cho sự ước lượng vị trí mỗi keypoint của nó, biết phiếu đó đã được đo lường bằng xác suất mà nó xuất hiện trong tầm ảnh hưởng của keypoint tương ứng. Hệ số chuẩn hóa bằng diện tích của vùng ảnh hưởng và đảm bảo rằng nếu các bản đồ nhiệt và các độ dời là hoàn hảo (chuẩn), thì hàm $f_k(x_i)$ sẽ là một hàm delta unit-mass có tâm tại vị trí của keypoint thứ k

Quá trình được minh họa trong Hình 2.3. Chúng ta thấy rằng việc dự đoán

bản đồ nhiệt và offsets, kết hợp cả hai lại theo quy trình biểu quyết được đề xuất thành các bản đồ kích hoạt có độ chính xác cao giúp xác định chính xác vị trí của các keypoints.



HÌNH 2.3: Mạng tích chập hoàn toàn của bài báo [12] dự đoán hai mục tiêu: (1) Các bản đồ nhiệt hình xung quanh mỗi keypoint và (2) cường độ của các trường bù đối với vị trí điểm chính xác trong đĩa. Tổng hợp chúng trong một quy trình bỏ phiếu có trọng số dẫn đến các bản đồ kích hoạt được bản địa hóa cao. Hình vẽ cho thấy các bản đồ nhiệt và thông số chính xác của trường bù trên ảnh xác thực. Lưu ý rằng trong hình minh họa này, chúng tôi áp đặt các kênh từ các keypoint khác nhau³

Tác giả đã tổng hợp lại kỹ thuật trên theo hướng trường tổng hợp và thêm vào một scale tạo thành Part intensity fields.

Pif là một cấu trúc tổng hợp, bao gồm: một thành phần scalar cho chúng ta biết độ tin cậy của từng joint, ngoài ra thành phần scalar cho biết được kích thước của joint. Với mỗi vị trí toạ độ (x, y) thì sau khi qua thành phần Pif, mỗi joint là

³Ảnh được lấy từ bài báo [12]

một vector gồm 5 phần tử, được thể hiện qua công thức bên dưới:

$$P^{ij} = \{p_c^{ij}, p_x^{ij}, p_y^{ij}, p_b^{ij}, p_\sigma^{ij}\} \quad (2.2)$$

Trong đó:

p_c^{ij} : độ tin cậy của joint.

p_x^{ij} : toạ độ x của joint.

p_y^{ij} : toạ độ y của joint.

p_b^{ij} : độ loang của joint.

p_σ^{ij} : kiểm soát độ dịch chuyển của joint, được sử dụng trong đánh giá OKS.

Confidence map ở trên thì chưa chính xác ở việc xác định vị trí các joint. Hình 2.4 hiển thị confidence map của vai trai trên người. Để cải thiện vị trí của confidence map, tác giả đã hợp nhất confidence map ở trên với vertical part (hình 2.5) tạo thành high resolution confidence map (hình 2.6). Tác giả tạo ra high resolution confidence map $f(x,y)$ với unnormalized Gaussian kernel.

$$f(x,y) = \sum_{ij} p_c^{ij} N(x,y|p_x^{ij}, p_y^{ij}, p_\sigma^{ij}) \quad (2.3)$$

Trong đó:

$f(x,y)$: high resolution Confidence map

p_c : độ tin cậy của joint

p_x : toạ độ x của joint

p_y : toạ độ y của joint

p_σ : độ dịch chuyển của joint

N : unnormalized Gaussian kernel

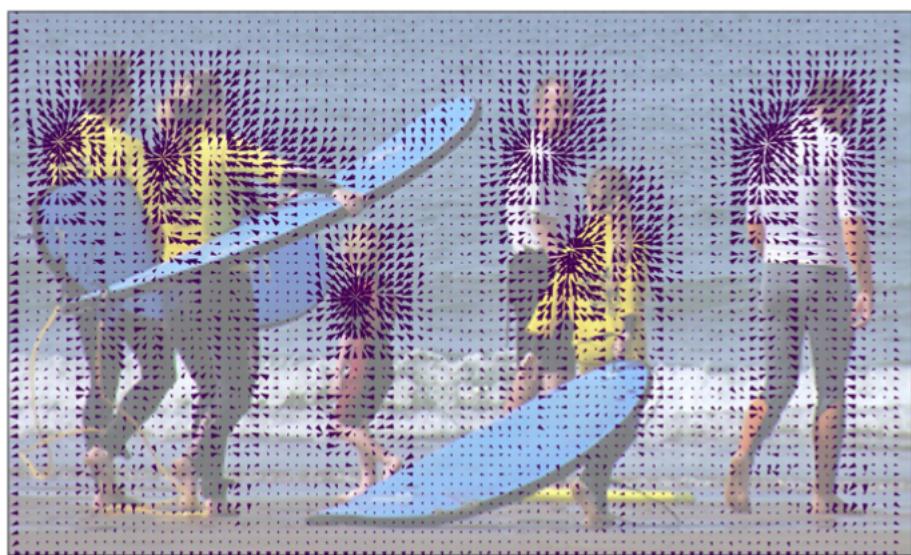
⁴Ảnh được lấy từ bài báo [8]

⁵Ảnh được lấy từ bài báo [8]

⁶Ảnh được lấy từ bài báo [8]



HÌNH 2.4: Minh họa Confidence map của vai trái trên mỗi người trong hình ảnh⁴



HÌNH 2.5: Minh họa Vector fields của joint (vai trái) trong hình ảnh kết hợp với confidence map để tạo nên high resolution confidence map⁵



HÌNH 2.6: Kết quả của joint (vai trái) sau khi qua thành phần Pif⁶

2.4 Part Association Fields

Part Association Fields (PAF), dùng để học các liên kết của các joint với nhau. Liên kết joint này với một joint khác là một thách thức trong những cảnh đông đúc, nơi mà người này che người làm cho việc liên kết các joint của một người gấp khó khăn. Với hướng tiếp cận như phương pháp Top Down, thường gặp khó trong vấn đề trên khi mà các bounding box chồng lên nhau: đầu tiên là phát hiện các bounding box chứa người và sau đó ước lượng các khớp người trong mỗi bounding box đó. Phương pháp từ Bottom Up thì không sử dụng bounding box nên không bị các vấn đề về bounding box chồng lên với nhau. Từ khó khăn trên nên tác giả đã đề xuất phương pháp Bottom Up Part Association Fields (PAF) để kết nối các vị trí joint với nhau tạo thành tư thế của người. Minh họa cho sơ đồ PAF trong hình 2.7.

Ở mỗi vị trí thì đầu ra sau khi qua thành phần PAF thì mỗi liên kết giữa hai joint là một vector gồm 7 phần tử. Có thể được viết như sau:

$$a^{ij} = \{a_c^{ij}, a_{x1}^{ij}, a_{y1}^{ij}, a_{b1}^{ij}, a_{x2}^{ij}, a_{y2}^{ij}, a_{b2}^{ij}\} \quad (2.4)$$

Trong đó:

a_c^{ij} : độ tin cậy của liên kết.

a_{x1}^{ij} : toạ độ x1 của joint thứ nhất.

a_{y1}^{ij} : toạ độ y1 của joint thứ nhất.

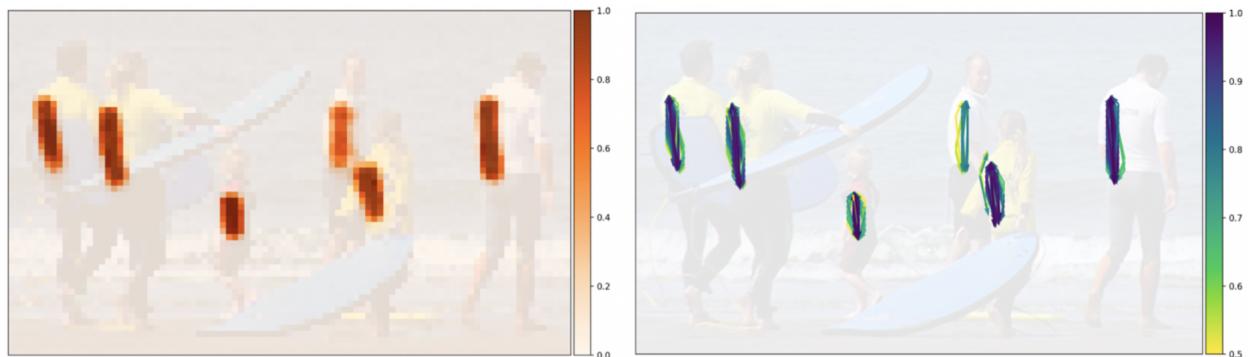
a_{b1}^{ij} : độ loang của joint thứ nhất

a_{x2}^{ij} : toạ độ x1 của joint thứ hai.

a_{y2}^{ij} : toạ độ y1 của joint thứ hai.

a_{b2}^{ij} : độ loang của joint thứ hai.

Để cho chúng ta có thể hình dung về sự liên kết của joint với nhau thì thông qua hình 2.7 minh họa sự liên kết giữa vai trái và hông phải.



HÌNH 2.7: Minh họa các thành phần PAF liên kết vai trái và hông phải ⁷

Thuật toán Paf gồm hai bước:

- Đầu tiên, Bắt đầu từ một joint và từ đó tìm joint gần nhất. Sau đó thì xác định một thành phần của vector.

⁷Ảnh được lấy từ bài báo [8]

- Tiếp đó, Ground truth xác định các thành phần còn lại của vector để đại diện cho sự kết nối đó.

Trong quá trình huấn luyện, cho chúng ta biết các joint nào nên được liên kết với nhau.

Theo tiêu chí của COCO thì có tổng cộng là 19 kết nối giữa hai joint với nhau cho một người (Hình 2.8) đó là:

- liên kết mũi - mắt trái.
- liên kết mũi - mắt phải.
- liên kết mắt trái - mắt phải.
- liên kết mắt trái - tai trái.
- liên kết mắt phải - tai phải.
- liên kết tai trái - vai trái.
- liên kết tai phải - vai phải.
- liên kết vai phải - vai trái.
- liên kết vai trái - khuỷu tay trái.
- liên kết khuỷu tay trái - cổ tay trái.
- liên kết vai phải - khuỷu tay phải.
- liên kết khuỷu tay phải - cổ tay phải.
- liên kết vai trái - hông trái.
- liên kết vai phải - hông phải.
- liên kết hông trái - hông phải.

- liên kết hông trái - đầu gối trái.
- liên kết đầu gối trái - mắt cá trái.
- liên kết hông phải - đầu gối phải.
- liên kết đầu gối phải - mắt cá phải.



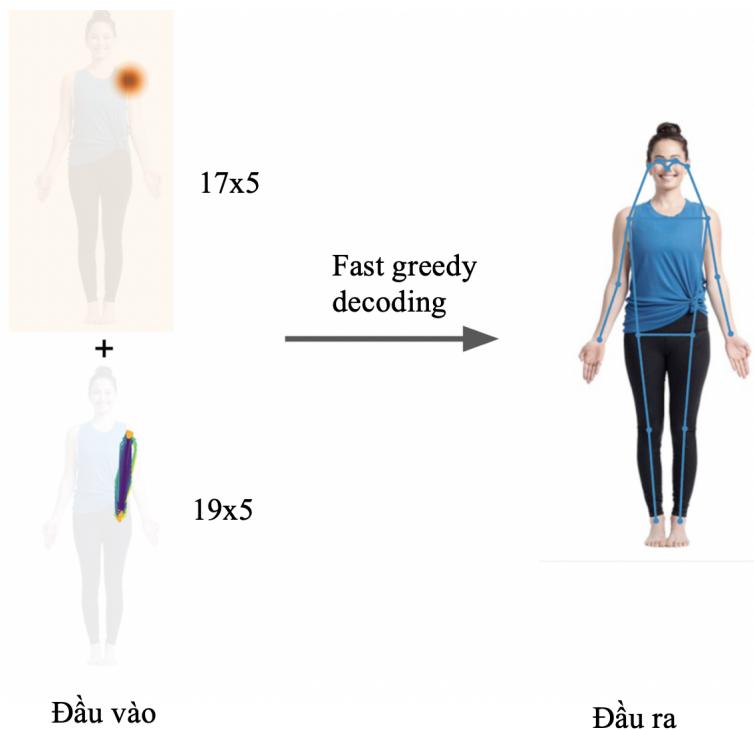
HÌNH 2.8: Minh họa 19 kết nối cho một người⁸

2.5 Greedy Coding

Phần Decoder là quá trình kết hợp kết quả từ hai trường Pif và Paf và chuyển đổi thành các ước tính tư thế chứa 17 joints trên mỗi người để tạo thành một tư thế người. Quá trình này tác giả sử dụng giống như fast greedy decoding. [11]

⁸Ảnh được lấy từ COCO Keypoint Challenge

Một tư thế mới được hình thành bởi các vector Pif với giá trị trong confidence map cao nhất. Bắt đầu từ joint đầu tiên, kết nối đến các joint khác được thêm vào với sự trợ giúp của thành phần Paf. Thuật toán này được gọi là fast and greedy, Sau khi kết nối đến một joint mới đã được đưa ra thì đó là quyết định cuối cùng.



HÌNH 2.9: Minh họa sự kết hợp kết quả của Pif và Paf chuyển đổi thành các ước tính tư thế chứa 17 joints để tạo thành tư thế người. Đầu vào là Pif và Paf, Tại Pif sẽ là 17 joint và mỗi joint là một vector gồm 5 phần tử (minh họa vai trái), Tại Paf là một liên kết của hai joint với một vector là 7 phần tử (minh họa kết nối vai trái và hông trái), gồm 19 kết nối. Đầu ra là chuyển đổi thành 17 joint tạo thành một khung xương thông qua thuật toán fasst greedy decoing⁹

Nhiều kết nối Paf có thể hình thành từ các kết nối giữa hai joint(joint hiện tại và joint tiếp theo). Để biết chính xác joint nào sẽ được kết nối với joint hiện tại để hình thành kết nối hoặc kết nối nào bị loại bỏ thì chúng ta dựa vào độ tin cậy của

⁹Ảnh tự thiết kế

liên kết đó thì tại vị trí của joint bắt đầu là x, và độ tin cậy liên kết giữa hai joint là s của Paf được tính dựa trên công thức sau:

$$s(a, \vec{x}) = a_c \exp\left(-\frac{||\vec{x} - \vec{a}_1||_2}{b_1}\right) f_2(a_{x2}, a_{y2}) \quad (2.5)$$

Trong đó :

$s(a, \vec{x})$: độ tin cậy thể hiện sự liên kết của joint x và liên kết a

\vec{x} : vector của joint bắt đầu liên kết

$a1$: vector của joint 1 trong liên kết a

$b1$: độ loang của joint bắt đầu

a_{x2} : vị trí x joint 2 của liên kết a

a_{y2} : vị trí y joint 2 của liên kết a

a_c : độ tin cậy của liên kết a

$f_2(a_{x2}, a_{y2})$: high resolution confidence map joint 2 của liên kết a.

Để xác nhận vị trí joint mới được đề xuất là chính xác. Tác giả nới ngược lại. Quá trình này được lặp lại cho đến khi một tư thế hoàn chỉnh. Ngoài ra tác giả áp dụng non maximum suppression ở mỗi keypoint [11] với bán kính dựa vào dự đoán thành phần scale của Pif.

Chương 3

THỰC NGHIỆM VÀ ĐÁNH GIÁ

Trong chương này, chúng tôi giới thiệu về bộ dữ liệu COCO và nói về các kết quả mà phương pháp đạt được trên bộ dữ liệu COCO khi so sánh với các phương pháp khác như OpenPose và Mask-RCNN. Ngoài ra, chúng tôi phân tích và thông kê chi tiết phương pháp sẽ sai trên những lỗi nào và độ phân giải nào thì phương pháp sẽ làm tốt nhất trên 3 bộ dữ liệu khác nhau: 4000 hình ảnh COCO test-dev, 2000 hình ảnh về thể thao do chúng tôi thu thập, 1000 hình ảnh bộ dữ liệu MPII trên 3 mô hình khác nhau: ResNet50, Shufflenetv2k16w và Shufflenetv2k30w trên 3 độ phân giải: 256 pixel, 321 pixel và 641 pixel.Thêm vào đó là giới thiệu các lỗi chúng tôi thống kê, Và từ kết quả chúng tôi đưa ra những nhận xét và hướng tiếp cận để cải thiện phương pháp PifPaf. Ngoài ra chúng tôi trình bày về độ đo mà tác giả sử dụng để đánh giá.

3.1 Dữ liệu

Tác giả đã đánh giá phương pháp PifPaf trên Common Object in Context (COCO) ở độ phân giải thấp. Ban đầu, tác giả sử dụng dữ liệu gốc COCO, sau đó tác giả giới hạn chiều dài cạnh còn 321 pixel .

Phương pháp PifPaf tập trung chủ yếu vào độ phân giải thấp, không tập trung nhiều vào độ phân giải cao. Tuy nhiên, tác giả vẫn đánh giá phương pháp của

mình trên độ phân giải cao cùng với một số phương pháp khác tối ưu cho độ phân giải cao ở bộ dữ liệu COCO.

Ngoài việc đánh giá kết quả trên bộ dữ liệu COCO thông qua các độ đo, chúng tôi còn phân tích chi tiết kết quả của phương pháp bằng cách thủ công (quan sát bằng mắt) để biết phương pháp sai ở những lỗi nào phổ biến. Chúng tôi thực hiện điều đó dựa trên 4000 hình ảnh COCO test-dev, 1000 hình ảnh MPII, 2000 hình ảnh về chủ đề thể thao (bóng đá, bóng chuyền, ...).

3.2 COCO Dataset

COCO là một trong những bộ dữ liệu hình ảnh phổ biến hiện nay được thiết kế để thúc đẩy việc nghiên cứu phát hiện đối tượng với trọng tâm là phát hiện các đối tượng trong bối cảnh cụ thể. Một số đặc trưng của COCO: bao gồm các instance segmentations cho đối tượng gồm 80 danh mục, stuff segmentations gồm 91 danh mục, keypoint annotations cho person instances ...

Đặc biệt những thách thức COCO đặt ra chủ yếu cho các bài toán là panoptic segmentations (hình 3.1), object detection with segmentation masks (hình 3.2), person keypoint estimation (hình 3.3), DensPose (hình 3.4).



HÌNH 3.1: Panoptic segmentations¹

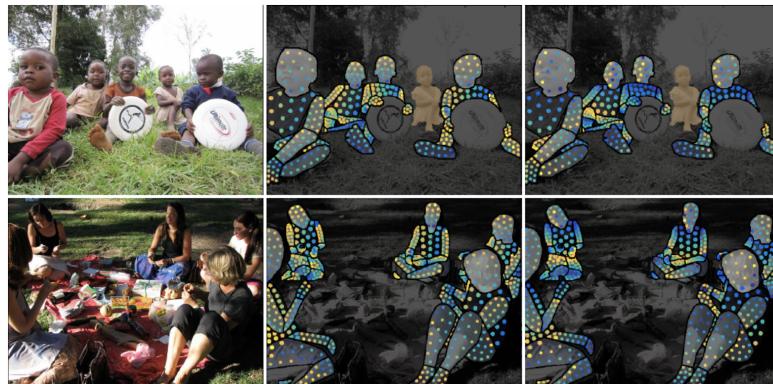
¹Ảnh được lấy từ Website COCO Dataset



HÌNH 3.2: Object detection with segmentation masks²



HÌNH 3.3: Person keypoint estimation³



HÌNH 3.4: DensePose⁴

²Ảnh được lấy từ Website COCO Dataset

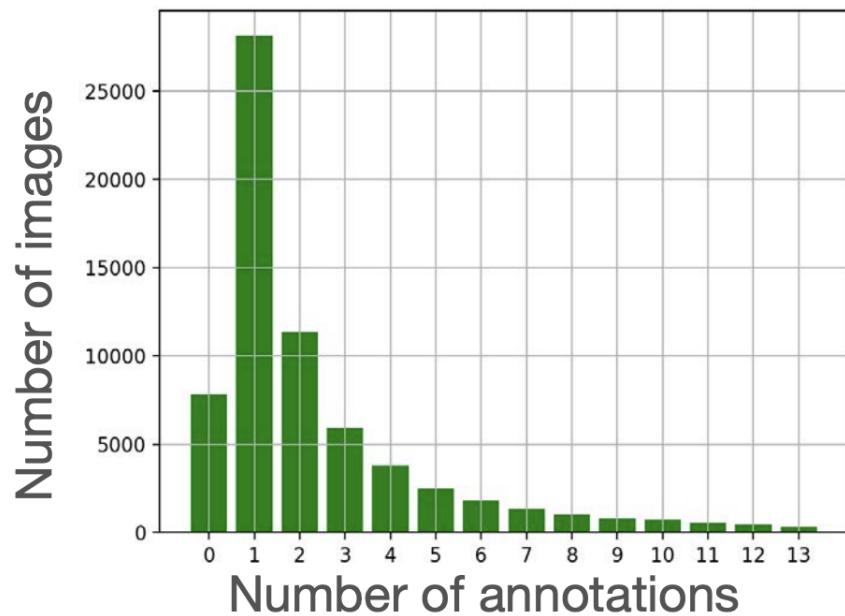
³Ảnh được lấy từ Website COCO Dataset

⁴Ảnh được lấy từ Website COCO Dataset

Bộ dữ liệu COCO 2017 bao gồm 118000 hình ảnh cho việc huấn luyện, 5000 hình ảnh cho đánh giá và 41000 hình ảnh cho thử nghiệm.

Đối với bài toán person keypoint estimation bộ dữ liệu COCO bao gồm:

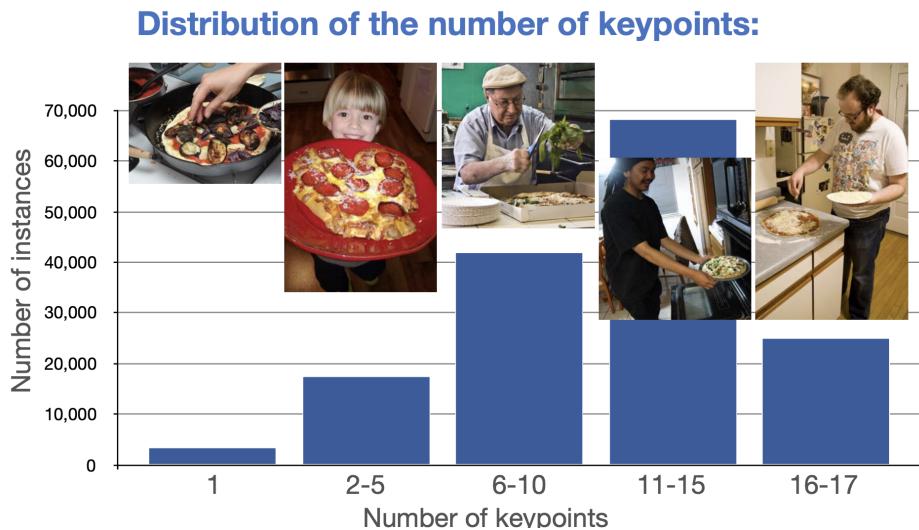
- 17 loại keypoint.
- 156165 người được gán nhãn.
- Số keypoint tổng cộng là : 1710498.
- Số người trung bình gán nhãn trong một ảnh là 2.
- Số lượng người được gán nhãn tối đa trong ảnh là 13 (hình 3.5).



HÌNH 3.5: Thống kê số người trong một ảnh trên bộ dữ liệu COCO⁵

⁵Ảnh được lấy từ 2017 Keypoint Challenge - COCO Dataset

⁶Ảnh được lấy từ 2017 Keypoint Challenge - COCO Dataset



HÌNH 3.6: Thông kê số lượng keypoint trên một người trong một ảnh trên bộ dữ liệu COCO⁶

3.3 Dữ liệu đánh giá hệ thống

Ngoài việc hệ thống được tác giả đánh giá trên bộ dữ liệu COCO validation dựa trên các độ đo, Để hiểu hơn kết quả đó chúng tôi phân tích chi tiết hơn các lỗi của phương pháp PifPaf bằng phương pháp thủ công (quan sát bằng mắt) dựa trên 4000 hình ảnh COCO test-dev, 2000 hình ảnh về thể thao (bóng đá, bóng bầu dục và bóng chuyền) do chúng tôi tự thu thập và 1000 hình ảnh trên bộ dữ liệu MPII ở các độ phân giải khác nhau như 256 pixel, 321 pixel, 641 pixel và ở mô hình khác nhau như: ResNet50, Shufflenet2k16w, Shufflenet2k30w. Từ đó có những thông kê khách quan hơn về chi tiết các lỗi mà hệ thống gặp phải và đưa ra các nhận xét cụ thể hơn về phương pháp Pifpaf.

Dưới đây là một số hình ảnh về bộ dữ liệu 4000 hình ảnh COCO test-dev:

⁷Ảnh được lấy từ COCO dataset

⁸Ảnh được lấy từ COCO dataset



HÌNH 3.7: Minh họa về bộ dữ liệu COCO test-dev⁷



HÌNH 3.8: Minh họa về bộ dữ liệu COCO test-dev⁸

Dưới đây là một số hình ảnh về bộ dữ liệu 2000 hình ảnh thể thao:

⁹Ảnh được lấy từ fanpage Manchester United trên facebook

¹⁰Ảnh được lấy từ fanpage Chelsea trên facebook



HÌNH 3.9: Minh họa về bộ dữ liệu thể thao⁹



HÌNH 3.10: Minh họa về bộ dữ liệu thể thao¹⁰

Dưới đây là một số hình ảnh từ bộ dữ liệu 1000 hình ảnh MPII:



HÌNH 3.11: Minh họa về bộ dữ liệu MPII¹¹



HÌNH 3.12: Minh họa về bộ dữ liệu MPII¹²

¹¹Ảnh được lấy từ MPII Dataset

¹²Ảnh được lấy từ MPII Dataset

3.4 Tiêu Chí đánh giá

Việc đánh giá keypoint detection có phần giống với đánh giá được sử dụng để phát hiện đối tượng, cụ thể là average precision (AP) và average recall (AR) và các biến thể của chúng tại ngưỡng Object keypoint Similarity.

3.4.1 Precision và Recall

Precision của một lớp cho biết thực chất có bao nhiêu dự đoán trong tổng số các dự đoán của mô hình là chính xác. Công thức tính:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.1)$$

Tương tự Recall được sử dụng để xác định tỉ lệ dự đoán đúng của mô hình so với tổng số đối tượng có trong dữ liệu. Công thức tính:

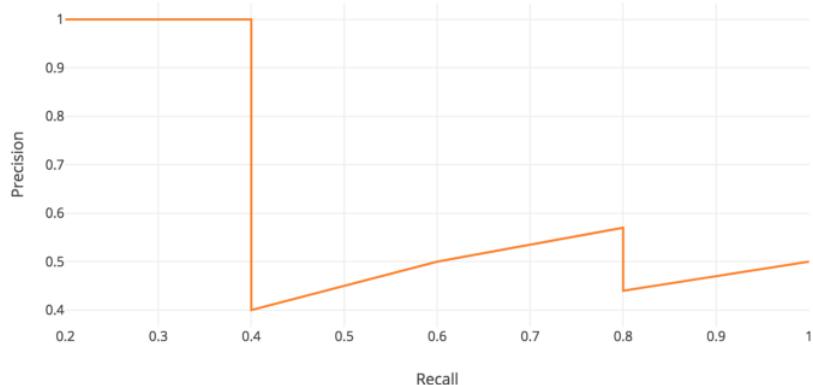
$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.2)$$

Trong đó:

- TP (True Positive): Phát hiện chính xác. Phát hiện với OKS \geq threshold.
- FP (False Positive): Phát hiện không chính xác. Phát hiện với OKS $<$ threshold.
- FN (False Negative): Các ground truth không được phát hiện.
- TN (True Negative): Các trường hợp mô hình dự đoán chính xác là không phải đối tượng.
- Threshold (Ngưỡng giới hạn): tùy vào loại metric, thường sẽ chọn ngưỡng là 50%, 75% , 95%.

3.4.2 Average Precision

Sau khi tính toán được Precision và Recall ta có thể dễ dàng xây dựng được biểu đồ quan hệ giữa precision và recall.



HÌNH 3.13: Biểu đồ quan hệ Precision và Recall¹³

Average Precision: là diện tích khu vực bên dưới đường precision-recall trong biểu đồ trên.

Công thức:

$$\text{Precision} = \int_0^1 p(r)dr \quad (3.3)$$

Precision và Recall nằm trong khoảng [0,1] nên AP (Average Precision) có giá trị trong khoảng [0,1].

3.4.3 Object Keypoint Similarity (OKS)

Trọng tâm của việc đánh giá là sự tương đồng giữa ground truth object và predicted object. Trong trường hợp phát hiện đối tượng, IoU đóng vai trò trong việc đánh giá. Để áp dụng AP và AR cho keypoint detection, chúng tôi chỉ cần xác

¹³Ảnh được lấy từ Internet

định độ đo tương tự bằng cách xác định độ đo object keypoint similarity (OKS) đóng vai trò như IoU.

Cho mỗi đối tượng, ground truth keypoint có dạng $x_1, y_1, v_1, \dots, x_k, y_k, v_k$, trong đó, x, y là tọa độ của keypoint, v là cờ để cho biết joint có bị che khuất hay không nếu $v = 0$ không có gán nhãn, $v=1$ có gán nhãn nhưng bị che khuất, $v=2$ là gán nhãn và không bị che khuất. Mỗi ground truth object có scale s , được định nghĩa là diện tích vùng của đối tượng.

Cho mỗi đối tượng, keypoint detector có đầu ra vị trí keypoint và độ tin cậy của object level. Dự đoán keypoint cho mỗi đối tượng nên có cấu trúc như ground truth: $x_1, y_1, v_1, \dots, x_k, y_k, v_k$

$$OKS = \frac{\sum_i [exp(-d_i^2 / 2s^2 k_i^2) \delta(v_i > 0)]}{\sum_i [\delta(v_i > 0)]} \quad (3.4)$$

Trong đó:

d_i : Khoảng cách Euclidean giữa ground truth và detected keypoint

v_i : cờ để đánh dấu keypoint có được nhìn thấy hay không trong ground truth

s : diện tích đối tượng trên diện tích ảnh

k_i : hệ số của từng keypoint được định nghĩa bởi COCO

σ : Kiểm soát độ dịch chuyển của keypoint giữa ground truth và predicted.

Dưới đây là 10 metrics được sử dụng cho việc đánh giá hiệu suất của keypoint detector trên COCO.

Ngoài ra chúng tôi phân tích chi tiết phương pháp dựa trên một số lỗi như sau: wrong connection (joint người này nối qua joint người khác tạo nên một liên kết sai), missing part (nhầm joint trên cùng một người), other (có nhiều người không được ước lượng joint), lack joint (người vẫn được ước lượng joint nhưng không đủ số lượng joint cần thiết), no detection (không ai trong hình được ước lượng joint) (Minh họa hình 3.14).

¹⁴Ảnh được lấy từ COCO dataset

Average Precision (AP)	
AP	AP tại ngưỡng OKS=.50: .05: .95
$AP^{OKS=.50}$	AP tại ngưỡng OKS=.50
$AP^{OKS=.75}$	AP tại ngưỡng OKS=.75
AP Across Scales	
AP^{medium}	AP cho những đối tượng trung bình: $32^2 < area < 96^2$
AP^{large}	AP cho những đối tượng lớn: $area > 96^2$
Average Recall (AR)	
AR	AR tại ngưỡng OKS=.50: .05: .95
$AR^{OKS=.50}$	AR tại ngưỡng OKS=.50
$AR^{OKS=.75}$	AR tại ngưỡng OKS=.75
AR Across Scales	
AR^{medium}	AR cho những đối tượng trung bình: $32^2 < area < 96^2$
AR^{large}	AR cho những đối tượng lớn: $area > 96^2$

BẢNG 3.1: 10 metrics đánh giá phương pháp tại các ngưỡng khác nhau



HÌNH 3.14: Các lỗi sử dụng cho phân tích và đánh giá phương pháp PifPaf bằng phương pháp thủ công(quan sát bằng mắt như wrong connection, missing part, other, lack joint, no detection.)¹⁴

3.5 Đánh giá hệ thống

Tất cả mô hình được dựa trên Imagenet pretrained sau đó tác giả chỉnh sửa lại để phù hợp với phương pháp PifPaf. Trong quá trình đào tạo, tác giả đã chỉnh sửa Batch Normalization operation [7] cho phù hợp với giá trị pretrained của tác giả. Tác giả sử dụng SGD optimizer với learning rate là 10^{-3} , momentum là 0.95, batch size là 8 và no weight decay. Ở mỗi bước tối ưu, tác giả đã cập nhật trọng số của mô hình. Decay constant là 10^{-3} . Thời gian huấn luyện cho 75 epoch trên mô hình ResNet101 trên 2 máy GTX1080Ti là xấp xỉ 95 giờ.

Tác giác so sánh PifPaf với một phương pháp hướng tiếp cận Bottom Up là Openpose, một phương pháp hướng tiếp cận Top Down là Mask RCNN, Table 3.2 hiển thị kết quả đánh giá trên bộ dữ liệu COCO-val. Phương pháp PifPaf vượt trội hơn so với các phương pháp bottum-up là Openpose và top-down là Mask RCNN, và chỉ có ở average recall tại ngưỡng OKS=0.5 là PifPaf bằng với Mask RCNN là 76%.

	AP	AP ^{0.50}	AP ^{0.75}	AP ^{medium}	AP ^{large}	AR	AR ^{0.50}	AR ^{0.75}	AR ^{medium}	AR ^{large}
Mask RCNN	41.6	68.1	42.5	28.2	59.8	49.0	76.0	50.0	35.6	67.5
OpenPose	37.6	62.5	37.2	25.0	55.3	43.9	65.3	44.9	26.7	67.5
PifPaf (ours)	50.0	73.5	52.9	35.9	69.7	55.0	76.0	57.9	39.4	76.4

BẢNG 3.2: Áp dụng ước lượng tư thế ở hình ảnh độ phân giải thấp với long side edge 321px cho top-down (top-part) và bottum-up (bottum-part) phương pháp. Mask-RCNN và Openpose, tác giả thử nghiệm với độ dài cạnh ảnh tối đa. Mask-RCNN được huấn luyện lại cho ảnh độ phân giải thấp. PifPaf dựa trên Resnet50 backbone.

Ngoài ra những phương pháp khác tối ưu cho hình ảnh độ phân giải cao, mặc dù phương pháp PifPaf tập trung chủ yếu vào độ phân giải thấp nhưng tác giả cũng đánh giá phương pháp PifPaf trên độ phân giải cao và so sánh với một số phương pháp khác ở độ phân giải cao trên bộ dữ liệu COCO 2017 test-dev ở bảng 3.3. Kết quả cho thấy Phương pháp PifPaf có kết quả tốt bằng các phương pháp bottum-up tốt nhất hiện có.

	AP	AP ^M	AP ^L
Mask RCNN	63.1	58.0	70.4
OpenPose	61.8	57.1	68.2
PersonLab single-scale	66.5	62.4	72.3
PifPaf single-scale	66.7	62.4	72.9

BẢNG 3.3: Metric tính theo phần trăm được đánh giá trên bộ thử nghiệm COCO test-dev 2017 ở độ phân giải tối ưu cho các phương pháp top down và bottom up.

Ngoài ra, Để đạt được kết quả tốt hơn tác giả đề xuất có thể điều chỉnh hiệu suất đối với các đối tượng nhỏ hoặc lớn hơn bằng cách thay đổi scale r^{smooth} khác nhau. Tuy nhiên, phương pháp PifPaf thực sự đạt kết quả tốt hơn lại dựa trên hàm Laplace-based loss (bảng 3.4).

	AP	AP ^M	AP ^L
Vanila L1	41.7	26.5	62.5
SmoothL1 $r = 0.2\sqrt{A_i}\delta_k$	61.8	57.1	68.2
SmoothL1 $r = 0.5\sqrt{A_i}\delta_k$	41.9	27.0	62.5
SmoothL1 $r = 1.0\sqrt{A_i}\delta_k$	41.6	26.5	62.3
Laplace	45.1	31.4	64.0
Laplace (using b in decoder)	45.5	31.4	64.9

BẢNG 3.4: Nghiên cứu sự phụ thuộc vào L1 loss. metric tính theo phần trăm, dựa trên model backbone Resnet50 đã được huấn luyện trong 20 epoch

Thêm vào đó tác giả còn thử nghiệm một số hướng tiếp cận khác để kết quả tốt hơn như thêm scale component σ vào Pif đã cải thiện được AP của ResNet 101 từ 64.5% lên 65.7%. (bảng 3.5)

Ngoài ra, tác giả cũng mới cập các kết quả thử nghiệm ở các mô hình khác nhau như Shufflenet và thêm vào đó là thời gian của từng mô hình ở bảng 3.6, thông qua đó thì chúng ta có thể xem xét việc đánh đội độ chính xác và thời gian để chọn ra kiến trúc phù hợp nhất.

	AP[%]	t[ms]	$t^{dec}[ms]$
Resnet50	62.6	222	178
Resnet101	65.7(60.0)	240	175
Resnet152	67.4	263	173

BẢNG 3.5: Kết quả thực nghiệm trên các model backbone như resnet50, resnet101, resnet152

Backbone	AP	AP^M	AP^L	$t_{total}[ms]$	$t_{dec}[ms]$	size
Resnet50	67.8	65.3	72.6	70	28	105MB
Shufflenetv2k16w	67.3	62.2	75.3	54	25	43.9MB
Shufflenetv2k30w	71.1	66.0	79.0	94	22	122.3MB

BẢNG 3.6: Kết quả thực nghiệm trên các model backbone như ResNet50, Shufflenetv2k16w, Shufflenetv2k30w

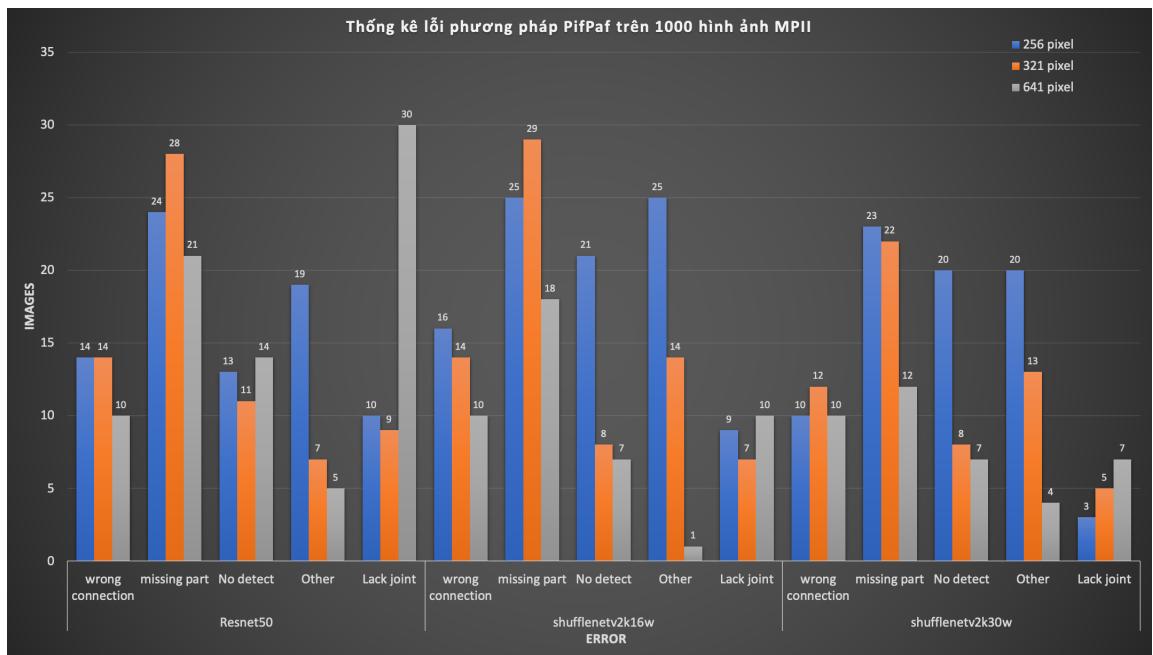
Như đã đề cập thì phương pháp PifPaf giải quyết tốt thách thức che khuất ở độ phân giải thấp. Dựa vào các kết quả ở trên thì chúng ta cũng nhận thấy được kết quả đánh giá chỉ đạt mức trung bình cho đến khá (Bảng 3.2) thì chứng tỏ phương pháp PifPaf vẫn còn một số vấn đề chưa giải quyết được, chúng ta có thể tập trung vào điều đó để tìm cách cải thiện phương pháp Pifpaf, để đơn giản hơn cho việc đưa ra các hướng đi tiếp theo để cải thiện phương pháp Pifpaf thì chúng tôi phân tích chi tiết hơn về các lỗi của PifPaf gấp phải, Phương pháp đề xuất là giải quyết ở độ phân giải thấp thì ở độ phân giải nào thì PifPaf làm tốt nhất. Và dưới đây là các bảng đánh giá và phân tích chi tiết các lỗi mà phương pháp gấp phải dựa trên 2000 hình ảnh về thể thao (bóng đá, bóng chuyền, bóng bầu dục) do chúng tôi tự thu thập, 4000 hình ảnh COCO tes-tdev 2017 , 1000 hình ảnh MPII trên các backbone model khác nhau.

Bảng 3.7 thống kê số lỗi mà pifpaf gấp phải dựa trên thực nghiệm 1000 hình ảnh MPII. Hình 3.15 là minh họa cụ thể hơn, thay vì dùng bảng thì chúng tôi sử dụng biểu đồ cột để dễ hình dung hơn.

¹⁵Ảnh được tự thiết kế

Backbone	long edge	wrong connection	missing part	no detection	other	lack joint	total error
Resnet50	256px	14	24	13	19	10	80
Resnet50	321px	14	28	11	7	9	69
Resnet50	641px	10	21	14	5	30	81
Shufflenetv2k16w	256px	16	25	21	25	9	96
Shufflenetv2k16w	321px	14	29	8	14	7	72
Shufflenetv2k16w	641px	10	18	7	1	10	52
Shufflenetv2k30w	256px	10	23	20	20	3	42
Shufflenetv2k30w	321px	12	22	8	13	5	60
Shufflenetv2k30w	641px	10	12	7	4	7	40

BẢNG 3.7: Phân tích thủ công các lỗi PifPaf trên 1000 hình ảnh MPII ở các độ phân giải: 256px, 321px, 641px dựa trên các model backbone resnet50, Shufflenetv2k16w, Shufflenetv2k30w



HÌNH 3.15: Minh họa kết quả phân tích 1000 hình ảnh MPII trên biểu đồ cột 15

Nhận xét trên 1000 hình ảnh MPII:

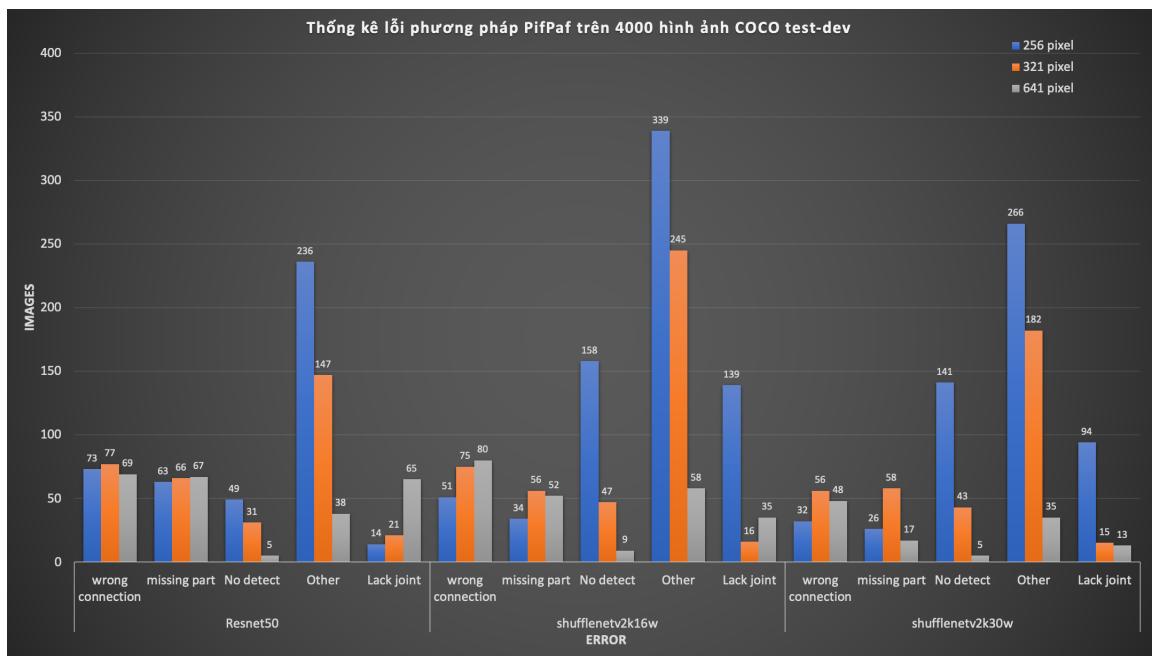
- Tổng số lỗi thì ở mô hình Shufflenetv2k16w độ phân giải cao cho kết quả tốt hơn và ở ResNet50 thì tổng lỗi 321 pixel lại nhỏ hơn 641 pixel.
- Phương pháp PifPaf tập trung lỗi vào missing part và other. Mô hình ResNet50 và shufflenetv2k16w thì lỗi missing part tại 256 pixel lại nhỏ hơn 321 pixel.

- Ở lỗi wrong connection và lack joint thì 256 pixel lại có lỗi thấp hơn 321 pixel (shufflenetv2k30w)

Bảng 3.8 thống kê số lỗi mà pifpaf gặp phải dựa trên thực nghiệm 4000 hình ảnh COCO test-dev. Hình 3.16 là minh hoa cụ thể hơn, thay vì dùng bảng thì chúng tôi sử dụng biểu đồ cột để thống kê các lỗi.

Backbone	long edge	wrong connection	missing part	no detection	other	lack joint	total error
Resnet50	256px	73	63	49	236	14	435
Resnet50	321px	77	66	31	147	21	342
Resnet50	641px	69	67	5	38	65	244
Shufflenetv2k16w	256px	51	34	158	339	139	721
Shufflenetv2k16w	321px	75	56	47	245	16	439
Shufflenetv2k16w	641px	80	52	9	58	35	234
Shufflenetv2k30w	256px	32	26	141	266	94	559
Shufflenetv2k30w	321px	56	58	43	182	15	354
Shufflenetv2k30w	641px	48	17	5	35	13	118

BẢNG 3.8: Phân tích thủ công các lỗi PifPaf trên 4000 hình ảnh COCO testdev ở các độ phân giải: 256px, 321px, 641px dựa trên các model backbone resnet50, Shufflenetv2k16w, Shufflenetv2k30w



HÌNH 3.16: Minh họa kết quả phân tích 4000 hình ảnh COCO test-dev trên biểu đồ cột^[16]

Nhận xét trên 4000 hình ảnh COCO test-dev:

- Về tổng lỗi của 3 mô hình thì độ phân giải càng thấp thì lỗi càng cao.
- Lỗi missing part và wrong connection ngược lại, ở 256 pixel lại có lỗi thấp hơn 321 pixel và thậm chí là 641 pixel (shufflenetv2k16w, shufflenetv2k16w).
- PifPaf tập trung lỗi nhiều vào other vì ở độ phân giải thấp hình ảnh thiếu chi tiết nên gây khó khăn cho việc ước lượng joint.

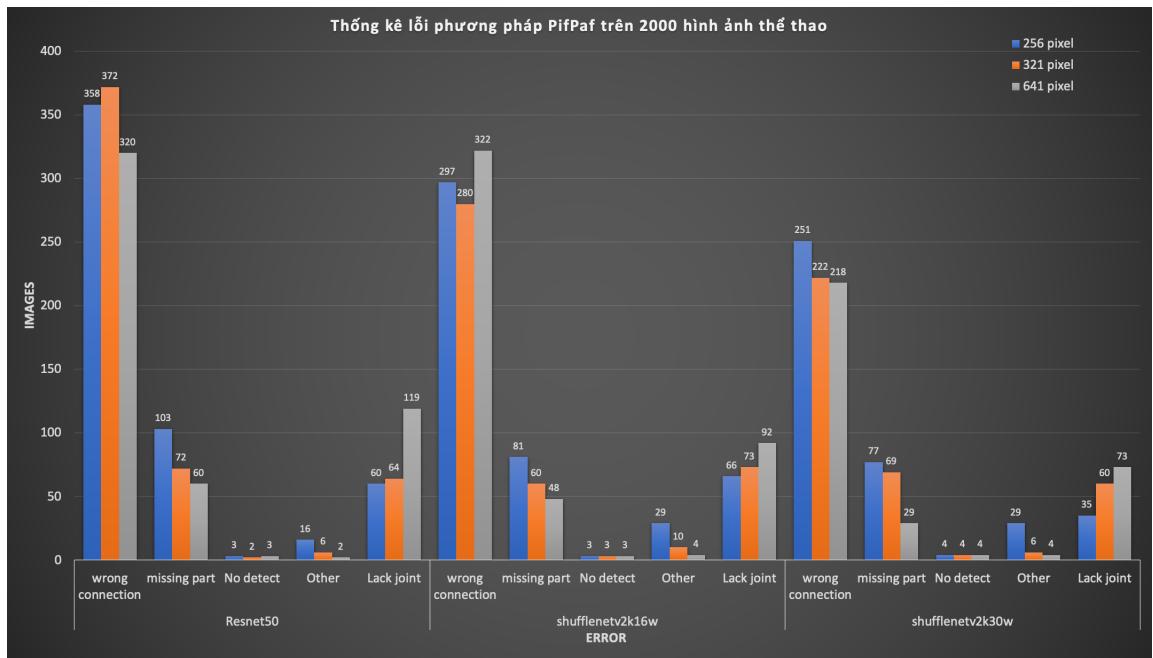
Bảng 3.9 thống kê số lỗi mà pifpaf gặp phải dựa trên thực nghiệm 2000 hình ảnh do chúng tôi thu nhập về chủ đề thể thao. Hình 3.17 là minh họa cụ thể hơn, thay vì dùng bảng thì chúng tôi sử dụng biểu đồ cột để thống kê các lỗi.

Backbone	long edge	wrong connection	missing part	no detection	other	lack joint	total error
Resnet50	256px	358	103	3	16	60	540
Resnet50	321px	372	72	2	6	64	516
Resnet50	641px	320	60	3	2	119	504
Shufflenetv2k16w	256px	297	81	3	29	66	476
Shufflenetv2k16w	321px	280	60	3	10	73	426
Shufflenetv2k16w	641px	322	48	3	4	92	496
Shufflenetv2k30w	256px	251	77	4	29	35	396
Shufflenetv2k30w	321px	222	69	4	6	60	361
Shufflenetv2k30w	641px	218	29	4	4	73	328

BẢNG 3.9: Phân tích thủ công các lỗi PifPaf trên 2000 hình ảnh thể thao tự thu thập ở các độ phân giải: 256px, 321px, 641px dựa trên các model backbone resnet50, Shufflenetv2k16w, Shufflenetv2k30w

¹⁶Ảnh được tự thiết kế

¹⁷Ảnh được tự thiết kế



HÌNH 3.17: Minh họa kết quả phân tích 2000 hình ảnh thể thao trên biểu đồ cột^[17]

Nhận xét 2000 hình ảnh về thể thao:

- Lỗi tập trung nhiều vào wrong connection và missing part, vì tác giả đề xuất phương pháp để giải quyết thách thức sự che khuất giữa những người với nhau mà một trong những môi trường thể hiện sự che khuất nhiều nhất là thể thao: bóng đá, bóng bầu dục, bóng chuyền...
- Tại lỗi lack joint thì ở độ phân giải 256 pixel có độ lỗi thấp hơn so với 321 pixel thậm chí là 641 pixel.
- Lỗi wrong connection thì 256 pixel nhỏ hơn 321 pixel ở mô hình ResNet50.

Ngoài ra, chúng tôi còn so sánh thời gian chạy trên 3 model ResNet50, Shufflenet2k16w, Shufflenet2k30w và 3 độ phân giải 256 pixel, 321 pixel, 641 pixel. (hình [3.10])

Backbone	256px	321px	641px
Resnet50	119ms	150ms	351ms
Shufflenetv2k16w	46.3ms	90ms	187ms
Shufflenetv2k30w	50ms	111ms	140ms

BẢNG 3.10: So sánh thời gian chạy trên một ảnh trên các model backbone: resnet50, shufflenetv2k16w, shufflenetv2k30w

Nhận xét chung:

- PifPaf có lỗi tập trung nhiều vào lỗi wrong connection và missing part. Ở bộ dữ liệu 1000 hình ảnh MPII và 4000 hình ảnh COCO test-dev thì lỗi tập trung vào other vì ở độ phân giải thấp thì hình ảnh thiếu chi tiết gây khó khăn trong việc ước lượng joint.
- Chúng ta có thể xem giữa hai độ phân giải 256 pixel và 321 pixel, Ở tổng số lỗi thì 321 pixel thấp hơn so với 256 pixel nhưng ở lỗi wrong connection và lack joint thì ngược lại, đôi khi là lỗi missing part và thời gian chạy của 256 pixel cũng nhanh hơn gần gấp đôi so với 321 pixel.
- Ở lỗi lack joint thì 256 pixel lại có số lỗi nhỏ hơn so với 321 pixel thậm chí là 641 pixel. Nếu chúng ta tập trung vào độ chính joint thì có thể xem xét chọn độ phân giải thấp thay vì chọn ảnh có độ phân giải cao để tiết kiệm chi phí.
- Như kết quả thống kê 2000 hình ảnh thể thao như bóng đá, bóng bầu dục thì lỗi tập trung nhiều vào wrong connection (chiếm 50-60% tổng số lỗi), thì chúng ta nên cân nhắc có nên áp dụng phương pháp này vào môi trường có tính che khuất cao như thể thao hay không.

Chương 4

KẾT LUẬN

Trong chương này chúng tôi nêu ra một số kết luận từ việc nghiên cứu phương pháp mà chúng tôi nghiên cứu và những kết quả đạt được qua quá trình thực nghiệm.

4.1 Kết luận

Kết thúc khoá luận, chúng tôi đã đạt được một số kết quả như mục tiêu đề ra: tìm hiểu các kỹ thuật liên quan ước tính tư thế người trong không gian hai chiều. Cụ thể như sau:

- Tìm hiểu các công trình nghiên cứu liên quan đến ước tính tư thế người trong không gian hai chiều.
- Tìm hiểu và phân tích phương pháp Bottum Up “PifPaf: Composite Fields for human pose estimation”.
- Phân tích và thống kê chi tiết các lỗi phương pháp PifPaf gấp phải dựa trên các hình khác nhau và ở những độ phân giải khác nhau trên 4000 hình ảnh COCO test-dev, 1000 hình ảnh MPII, 2000 hình ảnh về thể thao do chúng tôi thu thập. Từ đó đưa các nhận xét và hướng tiếp cận tổng quan hơn về phương pháp PifPaf.

Tài liệu tham khảo

- [1] Zhe Cao et al. *Openpose: Realtime Multi-Person 2D Person Estimation using Part Affinity Fields*. In *CVPR*, 2017, p. 7.
- [2] Y. Chen et al. *Cascaded pyramid network for multi-person pose estimation*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 7103–7112.
- [3] Qi Dang et al. *Deep Learnig Based 2D Human Pose Estimation: A Survey*. In *Tsinghua Science and Technology*, 2019.
- [4] Hao-Shu Fang et al. *RMPE: Regional Multi-Person Pose Estimation*. In *ICCV*, 2017.
- [5] K. He et al. *Mask r-cnn*. In *Computer Vison (ICCV) 2017 IEEE International Conference on*, IEEE, 2017, pp. 2980–2988.
- [6] S. Huang et al. *A coarse-fine network for keypoint localization*. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [7] S. Ioffe and C. Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. *arXiv preprint arXiv:1502.03167*, 2015.
- [8] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. *Pifpaf: Composite for Human Pose Estimation*. In *CVPR*, 2019.
- [9] A. Newell, Z. Huang, and J. Deng. *Associative embedding: End-to-end learning for joint detection and grouping*. In *Advances in Neural Information Processing Systems*, 2017, 2277–2287.

- [10] A. Newell et al. *Associative embedding: End-to-end learning for joint detection and grouping*. In *Advances in Neural Information Processing Systems*, 2017.
- [11] George Papandreou et al. *PersonLab: Person Pose Estimation and Instance Segmentation with Bottom-up, Part-Based, Geometric Embedding Model*. In *ECCV*, 2018.
- [12] George Papandreou et al. *Towards Accurate Multi person Pose Estimation in the Wild*. In *CVPR*, 2017.
- [13] *PoseTrack: PoseTrack Leader Board*, d. <https://posetrack.net/leaderboard.php>.
- [14] Girdhar R. et al. *Detect-and-track: Efficient pose estimation in videos*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 350–359.
- [15] Ke Sun et al. *Deep High-Resolution Representation Learning for Human Pose Estimation*. *CVPR*, 2019.
- [16] Bin Xiao, Haiping Wu, and Yichen Wei. *Simple Baselines for Human Pose Estimation and Tracking*. In *CVPR*, 2018.
- [17] Sen Yang, Wankou Yang, and Zhen Cui. *Pose Nerual Fabrics Search*. [Online]. Available: <https://arxiv.org/abs/1909.07068>, 2020, 34–50.
- [18] Feng Zhang et al. *Distribution-Aware Coordinate Representation for Human Pose Estimation*. In *CVPR*, 2020, 34–50.