ƯỚC TÍNH TƯ THẾ NHIỀU NGƯỜI TRONG ẢNH 2D THỜI GIAN THỰC DÙNG PAFs

Chữ đỏ: định nghĩa viết tắt lần đầu tiên

Highlight vàng: đọc mà éo hiểu cm gì.

Không highlight: hiểu 69.69%

**ABSTRACT:**

Chúng tôi trình bày một phương pháp hiệu quả để detect tư thế 2D song song của nhiều người trong một ảnh. Phương pháp sử dụng một đại diện không có thông số, PAFs được tham khảo để học cách liên kết các bộ phần cơ thể với mỗi cá nhân trong ảnh. Mô hình mã hóa toàn bộ bối cảnh, cho phép một bước phân tích từ dưới lên trên (bước này có độ chính xác cao, realtime và thực hiện song song nhiều người). Mô hình được thiết kế để kết hợp tìm vị trí các phần và liên kết giữa chúng thông qua 2 nhánh của quá trình dự đoán chuỗi giống nhau. Phương pháp của chúng tối đạt giải nhất cuộc thi COCO 2016 keypoints challenge và vượt trội hơn so với kết quả trước đó trong MPII Multi-Person benchmark về performance và sự hiêu quả. (nên node vô LV)

**1. INTRODUCTION**

Ước tính tư thế người trong ảnh 2D – vấn đề của việc tìm những keypoint hoặc phần – đã và đang tập trung nhiều vào việc tìm phần cơ thể của **một cá nhân.** Vì vậy tư thế của nhiều người trong những ảnh tạo ra một loạt những thử thách đặc biệt. Đầu tiên, mỗi ảnh có thể chưa một lượng không xác định số người, điều này có thể xảy ra ở những vị trí hoặc scale khác nhau. Thứ 2, sự tương tác giữa những người trong ảnh tạo ra sự phức tạp về không gian cũng như liên kết giữa các phần, vd như tiếp xúc, che khuất, khớp nối các chi. Thứ 3, độ phức tạp của thời gian thực thi có xu hướng tăng theo số lượng người trong ảnh, làm cho việc thực thi realtime là một thử thách.

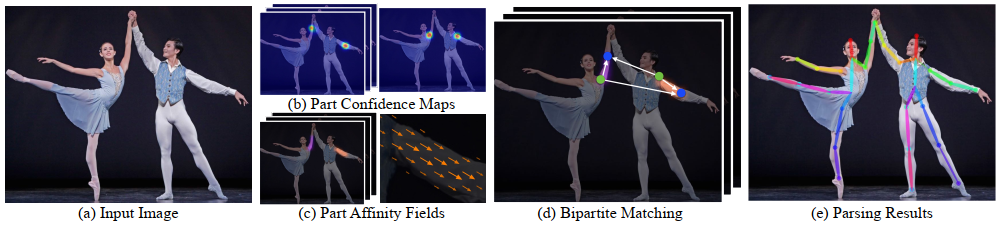
Các phương pháp chung [23, 9, 27, 12, 19] có xu hướng dùng bộ detector 1 cá nhân và ước tính tư thế cho một người duy nhất cho mỗi lần detect. Các phương pháp trên dưới này là đòn bẩy trực tiếp cho những phương pháp có sẵn để ước tính tư thế của một người [17, 31, 18, 28, 29, 7, 30, 5, 6, 20], nhưng theo như những đề cập phía trước: nếu bộ detector 1 cá nhân bị lỗi (bởi vì nó dễ xảy ra khi nhiều người đứng rất gần nhau), không hồi phục được. Hơn thế nữa, thời gian thực thi của những phương pháp trên dưới tỷ lệ với số người: mỗi lần detect, bộ ước tính tư thế một người được chạy, và khi số lượng người càng lớn thì chi phí tính toán càng tăng. Ngược lại, những phương pháp dưới trên được ưa chuộng hơn bởi vì chúng đáp ứng những điều kiện đã nêu ra và có tiềm năng tách rời độ phức tạp của thời gian tính toán so với số người trong ảnh. Chưa hết, những phương pháp dưới trên không sử dụng trực tiếp tín hiệu ngoại cảnh toàn cục từ những phần cơ thể khác và những người khác. Trong thực tế, những phương pháp dưới trên [22, 11] không giữ lại được độ hiệu quả vì những phân tích cuối cùng yêu cầu phân tích toàn cục tốn kém. Ví dụ, ...

Trong bài báo này, chúng tôi trình bày 1 phương pháp hiệu quả cho việc ược tính song song tư thế của nhiều người với số chính xác đỉnh cao trên nhiều chuẩn công khai. Chúng tôi trình bày đại diện dưới trên đầu tiên của điểm liên kết thông qua PAFs, một tập hợp miền vector 2D cái mà mã hoá vị trí và hướng của các chi thông qua miền ảnh. Chúng tôi chứng minh rằng những đại diện dưới trên này của sự detect và liên kết mã hóa nội dung toàn cục đủ tốt để cho phép một phân tích tham lam đạt được những kết quả về chất lượng cao với chi phí tính toán thấp. Chúng tôi đã public code...

**2. METHOD**



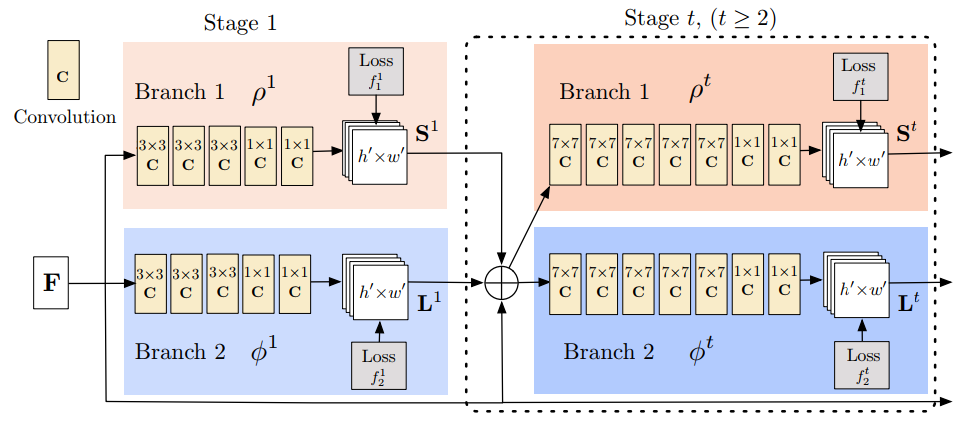
**Hình 1.** Hình trên: sự ước tính tư thế nhiều người. Những phần cơ thể thuộc cùng một người được liên kêt với nhau. Hình phía dưới bên trái: PAFs tương ứng với chi kết nối khủy tay phải và cổ tay phải. Màu sắc mã hóa hướng vector. Hình dưới bên phải: một góc phóng to của PAFs được dự đoán. Tại mỗi pixel trong miền, một vector 2D mã hóa vị trí và hướng của các chi.



**Hình 2.** Toàn bộ pipeline. Phương pháp của chúng tôi đưa toàn bộ ảnh đầu vào qua một mạng CNN 2 nhánh để đồng thời dự đoán những confidence map cho sự detect phần cơ thể, shown trong hình b, và part affinity fields cho sự liên kết các phần, shown trong hình c. Bước phân tích thể hiện một loạt những liên kết giữa hai điểm (liên kết lưỡng cực) để liên kết những phần cơ thể (d). Cuối cùng, chúng tôi lắp ráp chúng lại với nhau tạo thành những tư thế cơ thể hoàn chỉnh cho tất cả những người trong ảnh (e).

Hình 2 mình họa toàn bộ pipeline phương pháp của chúng tôi. Hệ thống lấy đầu vào, một ảnh màu có kích thước *w*x*h* (hình 2a) và tạo ra ngõ ra, tọa độ của những keyponts cho mỗi cá nhân trong ảnh (hình 2e). Đầu tiên, một mạng xxxx đồng thời dự đoán một loạt những confidence maps (cfm) **S** của những vị trí bộ phận cơ thể (hình 2b) và một loạt những miền vector 2D (vf) **L** của part affinities, cái mà mã hóa độ liên kết giữa các phần cơ thể (hình 2C). Tập hợp  có J cfm, một map cho mỗi bộ phận, trong đó . Tập hợp  có C vf, một cho mỗi chi, trong đó , mỗi vị trí ảnh trong Lc mã hóa một vector 2D (được show trong hình 1). Cuối cùng, cfm và affinity fields được phân tích bởi suy luận tham lam (hình 2d) để tạo ra các keypoints 2D cho tất cả người trong ảnh.

**2.1 Phát hiện và liên kết đồng thời:**



**Hình 3.** Kiến trúc của mạng CNN nhiều bước 2 nhánh. Mỗi bước trong nhánh đầu tiên dự đoán những cfm St, và mỗi bước trong nhánh thứ 2 dự đoán PAFs Lt. Sau mỗi bước, những dự đoán từ 2 nhánh, cùng với những đặc trưng ảnh, được nối lại cho bước tiếp theo.

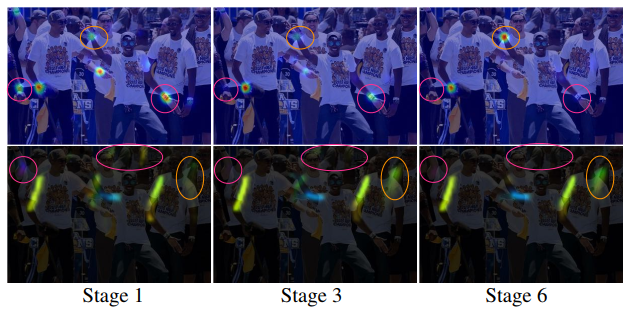
Kiến trúc của chúng tôi, show trong hình 3, dự đoán đồng thời những cfm và affinity fields (af) cái mà mã hóa liên kết giữa các phần. Mạng được tách thành 2 nhánh: Nhánh trên, màu hường da, dự đoán những cfm, và nhánh dưới, màu xanh, dự đoán những af. Mỗi nhánh là một kiến trúc dự đoán lặp lại, theo như Wei et al. [31], nó tinh chỉnh những dự đoán qua các giai đoạn liên tiếp , với giám sát trung gian ở từng giai đoạn.

Đầu tiên, ảnh được phân tích bởi một mạng tích chập (Khởi tạo bởi 10 lớp đầu tiên của VGG-19 [26] và hoàn thiện), việc tạo một tập hợp những feature map F là ngõ vào đến bước đầu tiên của mỗi nhánh. Tại bước đầu tiên, mạng tạo ra một tập những cfm  và một tập những PAFs , trong đó  và  là những mạng CNN cho suy luận tại bước 1. Trong mỗi bước tiếp theo, những dự đoán từ cả những nhánh ở bước trước, cùng với những đặc trưng ảnh gốc F, được nối và sử dụng để tạo ra các dự đoán được tinh chỉnh.





Trong đó  và  là những mạng CNN cho suy luận ở bước t.



**Hình 4.** Những cfm của cổ tay phải (những ảnh thuộc dòng đầu tiên) và những PAF (những ảnh thuộc dòng 2) của cánh tay phải qua các bước. Mặc dù, không có sự khác nhau giữa những phần bên trái và phải của cơ thể và các chi ở các bước đầu nhưng những ước tính được tinh chỉnh tăng cường thông qua suy luận toàn cục ở các bước sau, khu vực được highlight mô tả điều này.

Hình 4 mô tả sự tinh chỉnh của những cfm và những af khi trải qua các bước. Để hướng dẫn mạng dự đoán lặp lại những cfm của các phần cơ thể trong nhánh đầu tiên và PAFs trong nhánh thứ 2, chúng tôi áp dụng 2 hàm loss ở cuối mỗi bước, một ở mỗi nhánh tương ứng. Chúng tôi sử dụng một loss L2 giữa những dự đoán được ước tính và những bản đồ thực địa và các miền. Ở đây, chúng tôi cân nhắc những hàm loss không gian để giải quyết những vấn đề thực tế, cái mà một vài tập dataset không hoàn toàn gắn nhãn tất cả những người. Cụ thể, hàm los trong cả 2 nhánh ở bước *t* được định nghĩa là:





Trong đó  là cfm một phần groundtruth, là miền vector liên kết một phần groundtruth, W là một mặt nạ nhị phân với W(p) = 0 khi chú thích đang biến mất tại một địa phương ảnh p. Mặt nạ được sử dụng để tránh bắt lỗi những dự đoán vị trí chính xác trong quá trình training. Sự giám sát trung gian tại mỗi bước giải quyết vấn đề gradient biến mất bằng cách thay thế gradient định kỳ. Mục tiêu cuối cùng là:

Không hiểu groundtruth là gì



**2.2. Những cfm cho phát hiện bộ phận cơ thể**

Để đánh giá fs trong công thức f = ... trong quá trình training, chúng tối tạo những cfm groundtruth S\* từ những keypoints 2D được chú thích. Mỗi cfm là một đại diện 2D của độ tin cậy cái mà một bộ phận cơ thể cụ thể xảy ra tại mỗi vị trí pixel. Một các lý tưởng, nếu chỉ có một người trong ảnh, một đỉnh đơn sẽ có sẵng trong mỗi cfm nếu bộ phận tương ứng có thể nhìn thấy được; nếu nhiều người có trong ảnh thì sẽ có một đỉnh tương ứng với mỗi bộ phận nhìn thấy *j* cho mỗi người *k*.

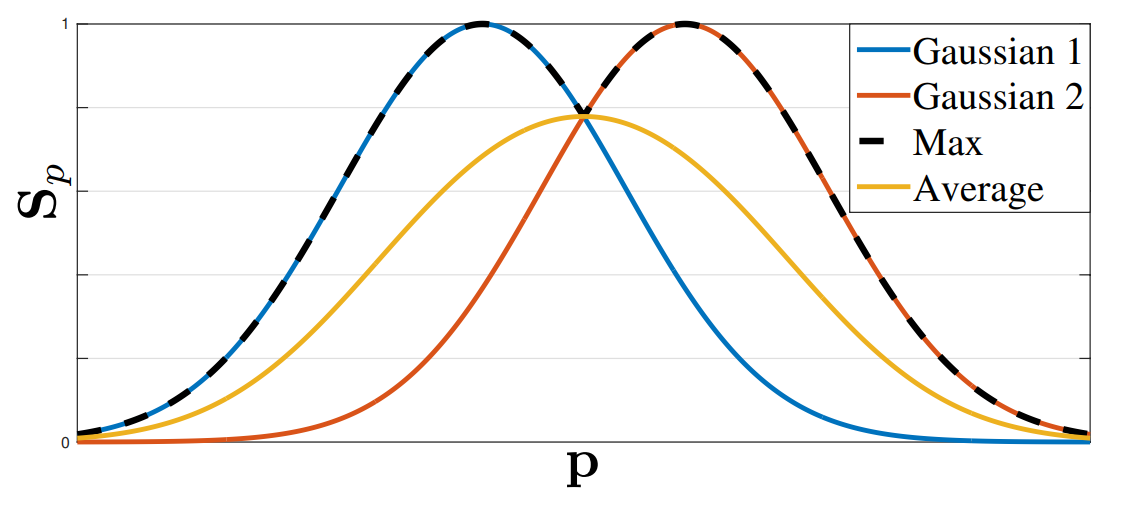
Đầu tiên, chúng tôi tạo những cfm S\*j,k cho mỗi người k. Gọi  là vị trí groundtruth của phần cơ thể *j* cho mỗi người *k* trong ảnh. Giá trị tại tọa độ  trong S\*j,k được định nghĩa là:



Trong đó σ điều khiển độ rộng của đỉnh. cfm groundtruth (được dự đoán bởi mạng) là một tổng hợp của những cfm cá nhân thông qua một hàm max,

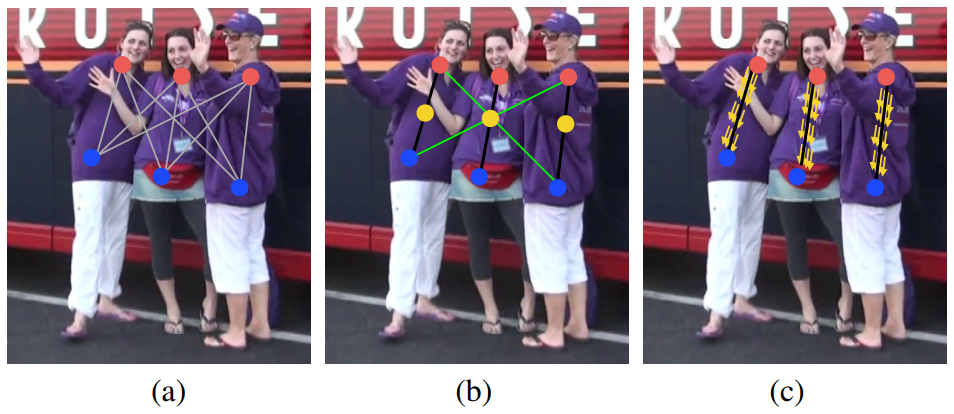


Chúng tôi lấy giá trị max của những cfm thay vì giá trị trung bình, vì vậy độ chính xác gần với những đỉnh nhưng vẫn giữ sự khác biệt, như được minh họa trong hình dưới đây,



Trong thời gian test, chúng tôi dự đoán những cfm (được show trong hàng đầu tiên của hình 4), và có được những candidates bộ phận cơ thể bằng cách thực hiện triệt tiêu không tối đa.

**2.3. PAFs cho liên kết bộ phận**

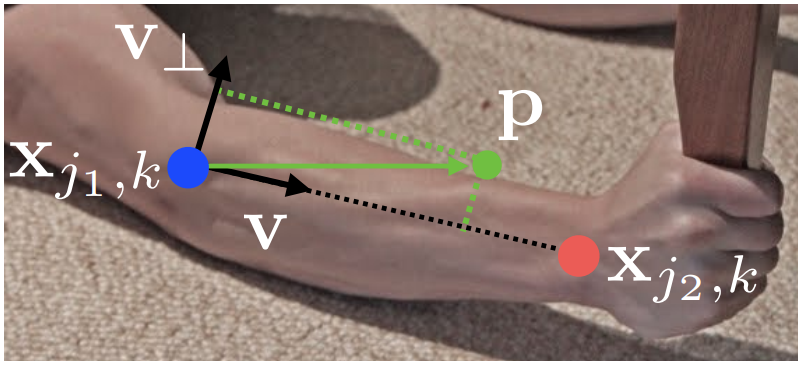


**Hình 5.** Những chiến lược liên kết bộ phận. (a) Những candidates sự detect bộ phận cơ thể (những chấm màu đỏ và xanh) cho 2 laoij bộ phận trên cơ thể và tất cả những liên kết (đường màu xám). (b) Những kết quả của sự kết nối sử dụng trung điểm (những điểm màu vàng): những liên kết chính xác (đường màu đen) và những liên kết không chính xác (đường màu xanh) cái mà cũng thỏa mãn những giới hạn góc tới. (c) Những kết quả sử dụng PAFs (mũi tên màu vàng). Bàng cách mã hóa vị trí và góc dưới sự hỗ trợ của các chi, PAFs loại bỏ những liên kết sai.

Cho một tập những bộ phận cơ thể được detect (show bằng màu đỏ và xanh trong hình 5a), làm cách nào chúng tôi có thể lắp ráp chúng để phù hợp với những tư thế của một số lượng người chưa xác định được? Chúng tôi cần một số đo tin cậy của sự liên kết cho mỗi cặp phương pháp phát hiện bộ phận cơ thể, nói cách khác là chúng thuộc cùng một người. Một cách khả thi để đo sự luên kết là để detect một trung điểm mới giữa mỗi cặp bộ phận trên một chi, và kiểm tra góc tới giữa các sự detect bộ phận ứng viên, được show trong hình 5b. Tuy nhiên, khi nhiều người tụ tập cùng nhau – điều này rất dễ xảy ra – những trung điểm này thì giống như những sự liên kết sai được hỗ trợ (mô tả bởi những đường màu xanh trong hình 5b). Khá nhiều liên kết sai phát sinh bởi vì hai sự giới hạn trong đại diện: (1) nó chỉ mã hóa những vị trí và không mã hóa hướng của mỗi chi; (2) nó giảm khu vực hỗ trợ của một chi thành một điểm đơn.

Để chỉ ra những hạn chế, chúng tôi trình bày một đại diện đặc trưng mới được gọi là *part affinity fields (PAFs)*, nó bao hàm cả thông tin về vị trí cũng như hướng trông qua khu vực hỗ trợ của chi (được thể hiện trong hình 5c). Part affinity là một miền vector 2D của mỗi chi, cũng được mô tả trong hình 1d: cho mỗi pixel trong khu vực thuộc một chi cụ thể, một vector 2D mã hóa co hướng (có gốc từ một phần ủa chi đến chi khác). Mỗi loại chi có một affinity field tương ứng nối hai bộ phận cơ thể được liên kết.

Xem xét một chi đơn được show trong hình dưới đây.



Gọi  và  là những vị trí groundtruth của những bộ phận cơ thể *j1* và *j2* từ chi *c* cho người *k*  trong ảnh. Nếu một điểm p nằm trên chi, giá trị tại  là một vector đơn vị hướng từ j1 đến j2; cho tất cả những điểm khác, giá trị vector là 0.

Để đánh giá fL trng công thức f = ... trong quá trình training, chúng tôi định nghĩa miền vector groundtruth part affinity, , tại một điểm ảnh p như sau:



Ở đây,  là vector đơn vị trong hướng của chi. Tập hợp các điểm trên chi được định nghĩa là những điểm với một mức ngưỡng về khoảng cách của đường phân khúc (segment line), tức là, những điểm p mà:



Trong đó, độ rộng chi σl là một khoảng cách trong những pixels, độ dài chi là , và là một vector vuông góc với v.

PAF groundtruth trung bình những AFs của tất cả những người trong ảnh,



Trong đó, nc(p) là số lượng những vectors khác vector 0 tại điểm p trên tất cả k người (tức là, trung bình những pixels nơi mà những chi của những người khác nhau chồng chéo lên nhau).

Trong suốt quá trình testing, chúng tôi đo sự liên kết giữa những detections bộ phận của ứng cử viên bằng cách tính tích phân đường trên PAF tương ứng, dọc theo đường nối các vị trí bộ phận của ứng cử viên. Ngoài ra, chúng tôi còn đo sự căng chỉnh của PAF được dự đoán với chi của ứng cử viên, cái mà được hình thành bởi kết nối những phần cơ thể được detect. Đặc biệt, cho hai vị trí bộ phận của ứng cử viên  và , chúng tôi lấy mẫu PAF được dự đoán, Lc­ đọc theo đường để đo độ tin cậy trong liên kết của chúng:

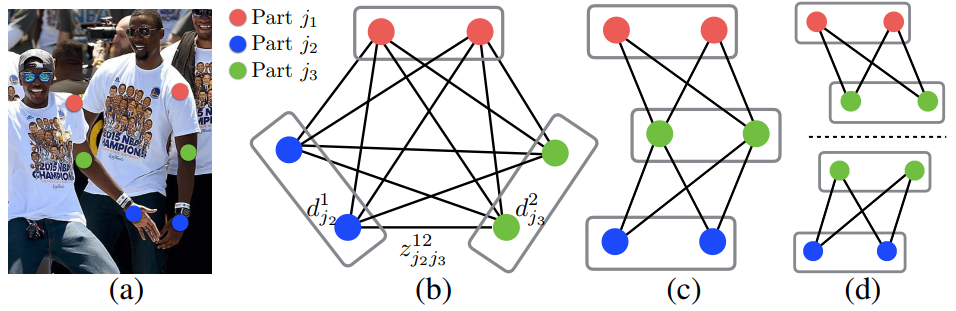


Trong đó p(u) nội suy vị trí của hai bộ phận cơ thể  và ,



Trong thực tế, chúng tôi xấp sỉ tích phân bằng cách lấy mẫu và tính tổng những giá trị cách đều nhau của u.

**2.4. Phân tích nhiều người sử dụng PAFs**



**Hình 6.** Sơ đồ matching. (a) Ảnh gốc với những detection bộ phận (b) Sơ đồ K phần (c) Cấu trúc cây (d) Một tập hợp các sơ đồ lưỡng cực

Chúng tôi trình bày sự ức chế không tối đa trên những cfms được detect để có được một tập rời rạc những vị trí bộ phận. Cho mỗi phần, chúng tôi có thể có nhiều ứng cử viên, bởi vì nhiều người trong ảnh hoặc những tích cực sai (mô tả trong hình 6b). Những bộ phận của ứng cử viên này định nghĩa một tập lớn những chi có khả năng. Chúng tôi cho điểm mỗi chi ứng cử sử dụng phép tính tích phân đường trên PAF, được định nghĩa trong công thức E=... Vấn đề của việc tìm phân tích tối ưu (parse) tương ứng với một vấn đề matching K chiều. Matching K chiều được hiểu là NP-Hard [32] (mô tả trong hình 6c). Trong bài báo này, chúng tôi trình bày một relaxation tham lam, nó đạt được những matches chất lượng cao một cách nhất quán. Chúng tôi dự đoán lý do là điểm số sự liên kết cặp mã hóa nội dung toàn cục, bởi vì the large receptive field của mạng PAF.

Đầu tiên, chúng tôi có một tập những candidates bộ phận cơ thể DJ cho nhiều người, trong đó , với Nj là số lượng những candidates của bộ phận j, và  là vị trí của detection candidate thứ m của bộ phận j. Những candidates bộ phận này vẫn cần được liên kết với những phân khác từ cùng một người – hay nói cách khác, chúng tôi cần tìm những cặp bộ phận được detect mà chúng được kết nối các chi thực sự. Chúng tôi định nghĩa một biến  để biểu thị liệu hai detection candidates  và  có được nối với nhau, và mục đích là để tìm ra sự tối ưu cho tập hợp tất cả những kết nối có khả năng, .

Nếu chúng tôi xét một cặp đơn bộ phận j1 và j2 (ví dụ: cổ và hông phải) cho chi thứ c, việc tìm tối ưu làm giảm một vấn đề matching đồ thị lưỡng cực trọng số tối đa. Trường hợp này được mô tả trong hình 5b. Trong vấn đề matching đồ thị này, những node của đồ thị là những detection candidates bộ phận cơ thể Dj1 và D­j2, và những cạnh là tất cả những kết nối có khả năng giữa các cặp detection candidates. Thêm vào đó, mỗi cạnh được tính bằng công thức E=... – Tổng hợp part affinity. Một matching trong một đồ thị lưỡng cực là một tập hợp con của các cạnh được chọn theo cách không có hai cạnh có cùng một nút. Mục đích của chúng tôi là tìm một matching với trọng số tối đa cho những cạnh được chọn.







Trong đó, Ec là trọng số tổng quan của việc matching từ chi loại c, Zc là tập con của Z cho chi loại c, Emn là part affinity giữa 2 bộ phận  và  được định nghĩa trong công thức E=.... Công thức  và  tuân theo luật hai cạnh không có cùng một nút, nghĩa là, hai chi của cùng một loại (ví dụ, cánh tay trái) không có cùng một bộ phận (part). Chúng tôi có thể sử dụng thuật toán Hungarian [14] để có được matching tối ưu.

Khi tìm kiếm tư thế toàn bộ cơ thể của cùng lúc nhiều người, việc xác định Z là một vấn đề matching K chiều. Vấn đề này là NP Hard [32] và nhiều relaxations có sẵn. Trong bài báo này, chúng tôi thêm hai relaxations vào sự tối ưu. Đầu tiên, chúng tôi chọn một số tối thiểu các cạnh để có một bộ xương cây của tư thế con người thay vì sử dụng đồ thị hoàn chỉnh, được thể hiện trong hình 6c. Tiếp theo, chúng tôi tiếp tục phân tích bài toán matching thành một tập những bài toán con matching lưỡng cực và xác định matching trong những nút cây liền kề một cách đọc lập, mô tả trong hình 6d. Chúng tôi thể hiện những kết quả so sánh chi tiết trong mục 3.1, và chứng minh rằng suy luận tham lam tối thiểu gần đúng với kết quả toàn cục với một phần chi phí tính toán. Lý do là quan hệ giữa các nút cây liền kề được mô hình hóa bởi PAFs, nhưng trong nội bộ, quan hệ giữa các nút cây không liền kề được mô hình hóa ngầm bởi CNN. Đặc tính này xuất hiện là do mạng CNN được đào tạo với miền tiếp nhận lớn, và PAFs từ những nút cây không liền kề cũng ảnh hưởng đến PAF được dự đoán.

Với hai relaxations này, sự tối ưu được phân tích một cách đơn giản bởi công thức:



Do đó, chúng tôi có được các candidates kết nối chi cho mỗi loại chi không phụ thuộc sử dụng công thức 12-14. Với tất cả những candidates kết nối chi, chúng tôi có thể lắp ráp những kết nối cái mà chia sẻ những detection candidates bộ phận giống nhau thành những tư thế toàn cơ thể của nhiều người. Sự tối ưu của chúng tôi trên cấu trúc cây nhanh hơn so với sự tối ưu trên đồ thị kết nối hoàn chỉnh.