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.

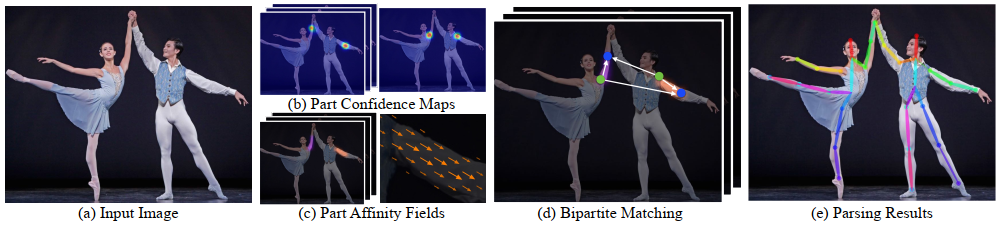
Các phương pháp chung [23, 9, 27, 12, 19] có xu hướng dùng bộ detector 1 cá nhân và ước tính tư thế cho một người duy nhất cho mỗi lần detect. Các phương pháp trên dưới này là đòn bẩy trực tiếp cho những phương pháp có sẵn để ước tính tư thế của một người [17, 31, 18, 28, 29, 7, 30, 5, 6, 20], nhưng 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**



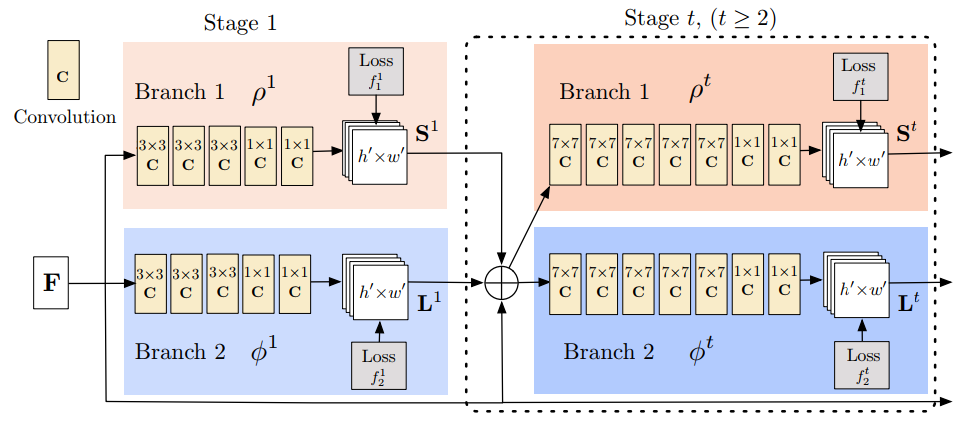
**Hình 1.** Hình trên: sự ước tính tư thế nhiều người. Những phần cơ thể thuộc cùng một người được liên kêt với nhau. Hình phía dưới bên trái: PAFs tương ứng với chi kết nối khủy tay phải và cổ tay phải. Màu sắc mã hóa hướng vector. Hình dưới bên phải: một góc phóng to của PAFs được dự đoán. Tại mỗi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

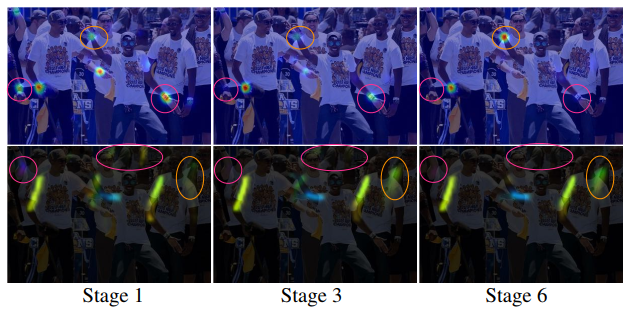
Kiến trúc của chúng tôi, show trong hình 3, dự đoán đồng thời những cfm và affinity fields (af) cái mà mã hóa liên kết giữa các phần. Mạng được tách thành 2 nhánh: Nhánh trên, màu hường da, dự đoán những cfm, và nhánh dưới, màu xanh, dự đoán những af. Mỗi nhánh là một kiến trúc dự đoán lặp lại, theo 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 thực địa, là miền vector liên kết một phần thực địa, 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:



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