Relatório de utilização do OpenPose para identificação de pedestres

Saulo Caliman Gomes
Programa de Pós-Graduação em Informática – (PPGI)
Universidade Federal do Espírito Santo – (UFES)
Vitória, Espírito Santo 29075–910
Email: saulocalimangomes@gmail.com

***Abstract*—Foi se analisado o artigo de Wei et al. 1 que propõe utilizar uma abordagem de representação não-paramétrica, a que é referido como *Part Affinity Fields* (PAFs), para aprender a associar partes do corpo a indivíduos na imagem. A arquitetura codifica o contexto global, permitindo uma análise *bottom-up* de alta precisão assim como identificação em tempo real, independentemente do número de pessoas na imagem. A arquitetura é projetada para aprender PAF e sua associação através de dois ramos do mesmo processo de previsão sequencial. É explicado como se dá o treinamento da rede para a geração do PAF e também é ressaltado sua aplicação no experimento em cima do material do LCAD.**

1. Introdução

O presente trabalho segue como princípio utilizar a ferramenta desenvolvida por Cao et al. [2] para realizar a detecção de pessoas a fim de proporcionar a carros autônomos a posição exata de uma determinada pessoa ou um conjunto de pessoas na faixa de pedestres.

A inferência de pose de múltiplas pessoas em imagens, representa um conjunto de desafios. Pode acontecer de cada imagem conter um número desconhecido de pessoas que pode ocorrer em qualquer posição ou escala. As interações entre pessoas também podem induzir complexas interferências espaciais, pelo contato, oclusão, e articulação de membros, fazendo a associação de partes difícil. E a complexidade de execução tende a crescer de acordo com o número de pessoas na imagem, fazendo com que o desempenho da execução em tempo real seja desafiadora. Pode-se destacar que os principais tipos de abordagem para realizar a identificação de pessoas, são os modelos *bottom-up* e *top-down*. Cao et al. [2]

No que tange pequenos níveis de características, cor, intensidade, orientação, textura, e movimentação podem renderizar a medida de saliência para cada pixel na imagem. Esse tipo de modelo é definido como modelo *bottom-up* de saliência ou atenção visual como representado na figura 2. A maioria das abordagens de saliência *bottom-up* usam características locais ao invés de características globais. Nas abordagens *top-down*, o mapa de saliência é determinado por características globais como informação contextual de um objeto ou uma cena. Um contexto de um objeto ou cena pode ser definido pela distribuição da probabilidade estatística do mesmo. A informação global provê informação contextual da categoria de um objeto ou uma cena. Essas abordagens são mais rápidas do que as *bottom-up* devido ao seu mecanismo de busca e sua estabilidade. Patel et al. [3]

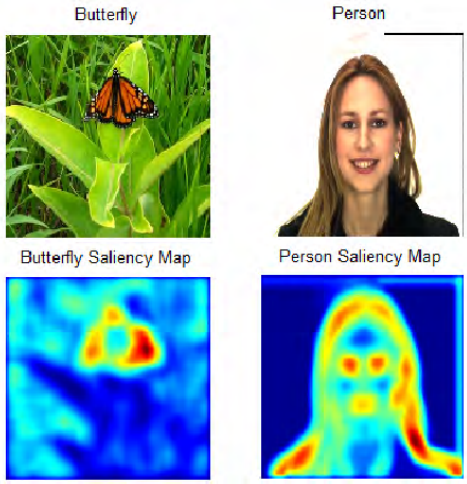
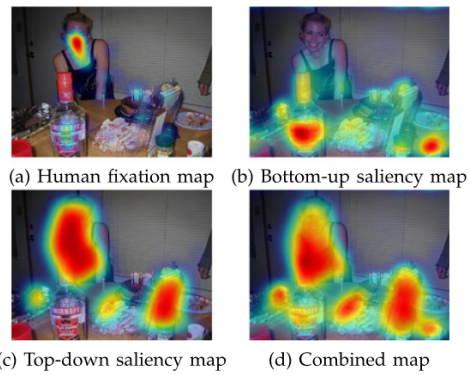


Figure 1. Duas imagens e seus mapas de saliência usando a abordagem *bottom-up*, qual indica características interessantes na imagem. A cor vermelha representa valores com saliência mais alta e a cor azul com a saliência mais baixa. Kanan and Cottrell [4]

Também vale ressaltar o ponto de fixação humano, em inglês *human fixation point*, é um modelo que prevê o ponto em que o humano fixa seu olhar. Judd et al. [6]

A saliência, no sistema de visão humano, é um termo bem abrangente que tem como principal significado o de destacar regiões sobressalentes de uma imagem. É possível utilizar a saliência em diversos sistemas de identificação de imagens. Kadir and Brady [7] Zhao et al. [8]

Para Cao et al. [2] a abordagem *top-down* alavancam técnicas já existentes para a estimação de uma única pessoa, porém sofrem com a má detecção em baixas proximidades– onde não existem recursos de recuperação. Também destacam que o tempo de execução de tal abordagem é proporcional ao número de pessoas, para cada detecção, o identificador da pose de uma única pessoa é executado,



(a) Human fixation map (b) Bottom-up saliency map

(c) Top-down saliency map (d) Combined map

Figure 2. Comparação do modelo do mapa de saliência *bottom-up* e *top-down* para predição de fixação humana. Yang and Yang [5]

e se houver mais pessoas, maior o custo computacional. Em contraste, a abordagem *bottom-up* possui maior atrativo pela sua robustez e sua complexidade de execução baixa para cada pessoa na imagem. Mas a abordagem *bottom-up* não utiliza diretamente o contexto global de outras partes do corpo de outras pessoas. E ainda existem técnicas para atribuir o contexto global porém a resolução do problema acontece em questão de horas.

Em seu artigo, Cao et al. [2], apresenta um método para estimar as poses com eficácia e bom desempenho através de uma representação *bottom-up* de associação de resultados por *Part Affinity Fields* (PAFs), que são um conjunto de campos vetoriais de duas dimensões, 2D, que codificam a orientação e localização dos membros de uma pessoa sobre o domínio de imagem. É demonstrado que inferindo simultaneamente a representação *bottom-up* da detecção e associação da codificação do contexto global é possível atingir bons resultados em uma fração do custo computacional. Foi também disponibilizado o código para a execução e reprodução de seu trabalho. CMU Perceptual Computing Laboratory [9]

2. Trabalhos Correlatos

Cao et al. em seu artigo utiliza a arquitetura da rede neural proposta por Wei et al. que é especializada em identificar os *keypoints* das pessoas e é baseada no trabalho de Ramakrishna et al..

A técnica de Ramakrishna et al. cita que a abordagem mais popular para se estimar poses de imagens é utilizar estruturas pictoriais. Os modelos de estrutura pictorial expressam o corpo humano como um modelo gráfico de estrutura de árvore, hierárquico, que conectam os membros do corpo de uma pessoa seguindo a hierarquia de seus membros. Porém a representação dessa estrutura faz com que o reconhecimento seja equivocado, uma vez que só reconhece aquilo que aparece na imagem, não realizando a inferência correta do membro da pessoa. Por conta desses equivocos ele propõe utilizar a técnica de Munoz et al., que treina a rede neural para realizar a decomposição hierárquica da imagem usando suas características e as predições passadas, fazendo com que o contexto de cada variável venha das variáveis vizinhas em relação ao seu espaço e ao seu tamanho.

3. Metodologia

3.1. Treinamento

No que tange o treinamento da rede para realizar o reconhecimento dos membros. É a mesma técnica utilizada por Wei et al. [1] e Ramakrishna et al. [10]. A rede recebe como entrada uma imagem, como representados na figura 3 e figura 4, e extrai características importantes da imagens, as posições exatas dos *keypoints*, é produzido os mapas de confiança, que por sua vez destacam o centro de um *keypoint* como representado na figura 5, e através dos mapas de confiança é gerada a estrutura de árvore e identificado os *PAFS*

Figure 3. Exemplo dos *labels* de *keypoints* da *MSCOCO* Lin et al. [12]

Figure 4. Exemplo dos *labels* de *keypoints* da *MPII* Andriluka et al. [13]

3.2. Identificação e Associação dos PAFs

Cao et al. propós uma arquitetura de rede neural aonde *h x w* é respectivamente o tamanho e a largura da imagem, na

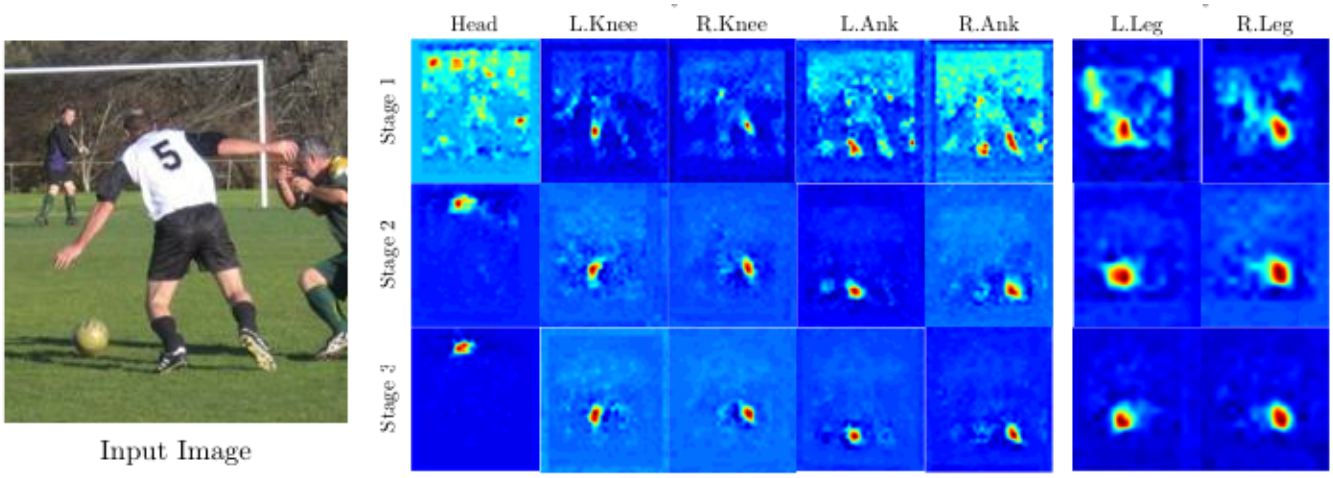


Figure 5. A técnica de Munoz et al. [11] produz iterativamente estimativas mais refinadas de confiança para a localização de cada parte. As confianças de da esquerda para a direita são para cada estágio de refinamento *head, left-knee, right-knee, left-ankle, right-ankle, left-leg, right-leg* Ramakrishna et al. [10]

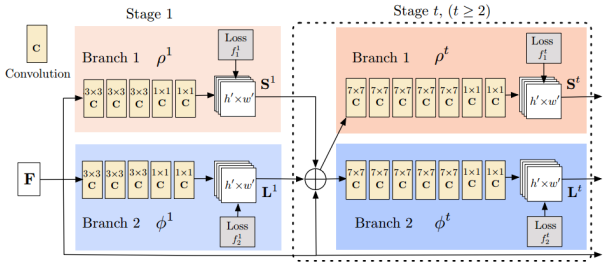


Figure 6. Arquitetura de dois nós multi-estágio CNN. Cada estágio no primeiro ramo prediz os mapas de confiança S^t, e cada estágio no segundo ramo prediz o PAFs L^t. Após cada estágio, as predições dos dois ramos, juntamente com as características das imagens, são concatenadas para o próximo estágio.

Figure 7. Identificação do OpenPose juntamente com *bounding box* do ciclista.

documentação do OpenPose é sugerido que o tipo de mídia a ser identificado contenha a resolução que seja de múltiplos de 16, levando em consideração também se a resolução é aumentada, a acurácia aumenta. Se for decrementada a velocidade aumenta. O melhor balanceamento de treinamento é a resolução 656x368 para videos com o aspecto 16:9, 1280 x 720 para HD e 1980 x 1080 para Full HD.

Primeiramente é utilizado uma rede *feed foward* que prediz um conjunto 2D de mapas de confiança de localizações das partes do corpo. E um conjunto de campos de vetores 2D para a *PAF* que codifica o grau de associação entre as partes e como último passo acontece a fusão dos mapas de confiança e os PAFs através de uma inferência gulosa tendo como saída os *keypoints* 2D de todas as pessoas da imagem.

4. Experimentos

No experimento foi utilizado a maquina CARMEN 2 do Laboratório de Computação de Alto Desempenho (LCAD), a máquina é uma Dell Precision 5500 com processador Intel Xeon E5606 2.3GHz x8, 32GB de memória RAM, possui duas GPUs, GPU 0- Quadro 600 sendo apenas para display e GPU 1- Titan X (Pascal) com 12GB de memória que é utilizada apenas para processamento e como sistema operacional utiliza o Ubuntu 14.04. As bibliotecas usadas no experimento foram Cuda 8.0 com cudnn v6, caffe e a Opencv 3.1.

Utilizando a OpenPose CMU Perceptual Computing Laboratory [9], com sua base já treinada, foi possível, utilizando o material do próprio LCAD, detectar pedestres na rua.

O experimento consiste em utilizar o material do LCAD e executar a detecção do OpenPose para a identificação dos pedestres na rua. Assim como fazer com que a *framework* OpenCV renderize ao redor de cada pessoa identificada um *bounding box* como pode ser visto na figura 7.

5. Resultados

Utilizando a OpenPose CMU Perceptual Computing Laboratory [9], com sua base já treinada, foi possível, utilizando o material do próprio LCAD, o reconhecimento de 12 pedestres sendo o total 13 pedestres que atravessaram na faixa, descartando os falsos positivos que apareceram no decorrer do vídeo. Sendo os falsos positivos; Arvores, carros e semáforos. Vale ressaltar que a identificação também ocorreu em pessoas fora da faixa de pedestre, foi identificado pessoas andando na calçada, paradas no ponto e esperando para atravessar.

Acredito que seja possível realizar a detecção de pedestres utilizando a arquitetura de rede proposta por Wei et al. [1] e usando como ferramenta de detecção feita por Cao et al. [2]. O cenário ideal no que tange a identificação de pedestres para o carro autônomo e para aumentar a eficacia da rede, seria realizar o seu treinamento com imagens de pedestres, uma vez que a base nativa da rede foi treinada com o COCO Lin et al. [12] e MPII human multi-person dataset Andriluka et al. [13] que são imagens, em sua maioria, especializadas no cotidiano. Especializar a rede para a detecção de pedestres com uma base própria para tal é possível.

References

[1] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[3] C. I. Patel, S. Garg, T. Zaveri, and A. Banerjee, "Top-down and bottom-up cues based moving object detection for varied background video sequences," *Advances in Multimedia*, vol. 2014, p. 13, 2014.

[4] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2472–2479.

[5] J. Yang and M.-H. Yang, "Top-down visual saliency via joint crf and dictionary learning," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2296–2303.

[6] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th international conference on*. IEEE, 2009, pp. 2106–2113.

[7] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[8] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[9] . CMU Perceptual Computing Laboratory. Openpose. [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose

[10] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *European Conference on Computer Vision*. Springer, 2014, pp. 33–47.

[11] D. Munoz, J. A. Bagnell, and M. Hebert, "Stacked hierarchical labeling," in *European Conference on Computer Vision*. Springer, 2010, pp. 57–70.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[13] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.