Estimativa de Pose 2D no Corpo Humano (Cabeça e pé) em tempo real de 5fr/s

Issufi Badji

Instituto de Computação – Universidade Federal Fluminense (UFF)
CEP 24.210-346 – Niterói – RJ – Brazil

@issufi.badji@aluno.ufr.edu.br

1. Introdução

Nas últimas décadas, houve um grande número de problemas a serem tratados computacionalmente e espacialmente no campo da visão computacional, devido ao rápido avanço digital e de técnicas de visão computacional que possibilita capturar, reconhecer, reconhecer e processar objetos visualmente perceptíveis. A visão computacional deriva das duas subáreas da inteligência artificial (IA): o aprendizado de máquina *(machine learning)* e, mais especificamente, as redes neurais artificiais *(Deep learning)* e são os maiores responsáveis pela grande evolução desta área nos últimos anos.

Na computação, o termo de *Machine learning (ML)* significa criar modelos lógico onde é possível simular uma inteligência computacional, de forma que o software possa ser alimentado com dados de entrada e a partir destes, são realizadas diversas tarefas para extrair informações relevantes ou desencadear tomadas de decisões a partir dos mesmos, com intuito de ter um algoritmo funcional que que capaz de aprender a atingir um objetivo com base nos dados ao qual foi alimentado (Feltrin, 2019).

A visão computacional, foca-se no desenvolvimento das teorias, métodos e técnicas voltados à extração automática de informações úteis contidas em imagens. Tem como objetivo principal programar o computador/máquina para que conseguem "enxergar" ou entender o mundo visual, interpretando e extraindo informações das imagens, que são capturados por meio de câmeras e sensores. A visão computacional vem sendo utilizada em várias áreas distintas, como por exemplo, navegação móvel por robô, tarefas complexas de manufatura, análise de imagens de satélites, processamento de imagens médicas e segurança organizacional.

Existem muitos campos na visão computacional, mas neste relatório, falaremos da estimativa de pose humano com cinco pessoas em tempo real. A estimativa de pose humano é um assunto de *Deep learning*, especialmente da visão computacional, que visa detectar e mapear os conjuntos dos pontos-chave do corpo humano. Os principais problemas relacionados à estimativa de pose, estão relacionados com corpo humano ou com uma da sua filiação como, das mãos (dedos), do rosto (boca e olhos) e dos pés. A estimativa de pose humano, tem atraído a atenção da investigação devido às suas potenciais aplicações em vários níveis de compreensão de alto nível tarefas, tais como vigilância de câmara, detecção de da placa de carros, reconhecimento de movimento humano, gesto e ações.

O relatório está dividido em seis grandes seções e da seguinte forma: Na primeira seção, introduzimos um breve relato de que será desenvolvido posteriormente. Na segunda seção, abordamos a descrição teórica da temática sobre a visão computacional. Na terceira seção, trabalhos relacionado sobre estimativa de pose humano em 2D. Na quarta seção, a técnica de captura e mapeamento de esqueleto humano. Na quinta seção, implementação (descrição de problemas e objetivo) e resultado obtidos. Na sexta e último, considerações finais.

2. Descrição Teórica do Tema

2.1 Machine learning e Deep learning

O *Machine learning (ML)* surgiu como uma evolução da área de inteligência artificial (IA). Inicialmente os problemas de IA eram mais simples e poderiam ser resolvidos através da aquisição de conhecimento através dos especialistas de um domínio. Com o passar dos anos, os problemas computacionais se tornaram mais complexos, assim como o crescente volume de dados disponível para a análise desses problemas (FACELI, 2017).

Para solução desses problemas mais complexos exige-se a descoberta de uma hipótese ou função que seja capaz de resolver esse problema através da experiência passada e de forma autônoma, sem a necessidade da interferência humana e de especialistas (FACELI, 2017).

Através da indução, que é uma inferência lógica, é possível obter conclusões genéricas baseadas em um conjunto particular de exemplos de um problema complexo. A inferência indutiva pode ou não apresentar resultados verdadeiros, mesmo assim, é um dos principais métodos para classificação de eventos passados e predição de eventos futuros.

O *Deep learnin (DL)* é um subconjunto de um campo de aprendizado de máquina, que visa aprender a partir de exemplos. Na DL, em vez de dar instruções para o computador resolver uma lista de problema, damos a ele um modelo com o qual ele pode avaliar exemplos e um pequeno conjunto de instruções para modificar o modelo quando cometer um erro. Assim, esperamos que, ao longo do tempo, um modelo adequado seria capaz de resolver o problema com extrema precisão (BUDUMA et al, 2017).

O DL surgiu a muito tempo, tornou-se mais útil conforme a quantidade de treinamento disponível os dados aumentaram e seus modelos cresceram em tamanho ao longo do tempo como infraestrutura de computador para aprendizado profundo melhorou. Graças a essa evolução hoje, o aprendizado profundo consegue resolver problemas mais complexos com menos precisão ao longo do tempo (GOODFELLOW et al, 2016). Portanto, DL é feito para aprender e otimizar conceitos e previsões a partir de conjunto de dados ou modelo treinado de um determinado assunto, com intuito de conseguir prever os possíveis erros e problemas, se ajustando e adaptando para consertá-los.

A visão Computacional é uma área da computação que estuda formas de extrair informação de dentro de uma imagem como, métodos de aquisição de imagens, pré-processamento, segmentação, extração de atributos ou características e reconhecimento de padrões. As soluções computacionais da visão computacional são baseadas em técnicas de processamento e análise de imagens, as quais permitem extrair informações a partir de imagens (FORSYTH et al., 2012).

3. Trabalho relacionado

Atualmente, existem vários métodos já foram desenvolvidos a partir das redes neurais convolucionais (CNNs) profundas, que as suas camadas convolucionais consistem em neurônios agrupados em filtros que convoluem os dados de entrada para produzir saídas ativadas. DeepPose é uma das primeiras aprendizagens profunda baseada na abordagem para a estimativa da pose humana, através de uma abordagem apoiada na regressão de pontos-chave do corpo usando redes neurais convolucionais, ou seja, são métodos baseados em mapas térmicos melhor aproveitam melhor o propulsor distribuído, e supera o desempenho de abordagens clássicas dos métodos antecessores. Diferente dos métodos mais recentes que são baseados nas previsões dos principais mapas que caracterizam as probabilidades de cada ponto-chave em diferentes locais de corpo. (WANG, RUI, et al, 2019).

Uma abordagem tradicional da estimativa de pose humana articulada em 2D, consiste em realizar inferência sobre uma combinação de observações locais em partes do corpo e as dependências espaciais entre eles (CAO, ZHE, et al, 2019). A estimativa de pose humano, é um problema que detecta e localiza os conjuntos dos pontos-chave do corpo humano, das mãos, do rosto e dos pés. Isso geralmente significa detectar o esqueleto de uma pessoa através da identificação das partes de articulação de um corpo humano.

A estimativa de pose humano com várias pessoas é um problema de detectar conjuntamente os pontos-chave do corpo humano múltiplas pessoas uma só vez, ou seja, é capaz de identificar as articulações de várias pessoas em uma única cena com algoritmos de detecção de esqueleto.

A detecção de esqueleto humano visa a localização e mapeamento de partes chave de corpo humano para gerar os mapas térmicos das articulações dos esqueletos. Os modelos precisam de uma imagem de entrada colorido, precisando passar a largura e comprimento da imagem para uma CNN, para previsão dos mapas que caracterizam as probabilidades, assim a saída será a marcação em 2D dos pontos-chave de cada pessoa na imagem. (CAO, ZHE, et al, 2019).

4. Método de captura de esqueleto humano

O modelo usado neste trabalho é baseado em um artigo intitulado *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields* (CAO, ZHE, et al, 2019). Os autores do artigo treinam redes neurais muito profundas para essa tarefa.

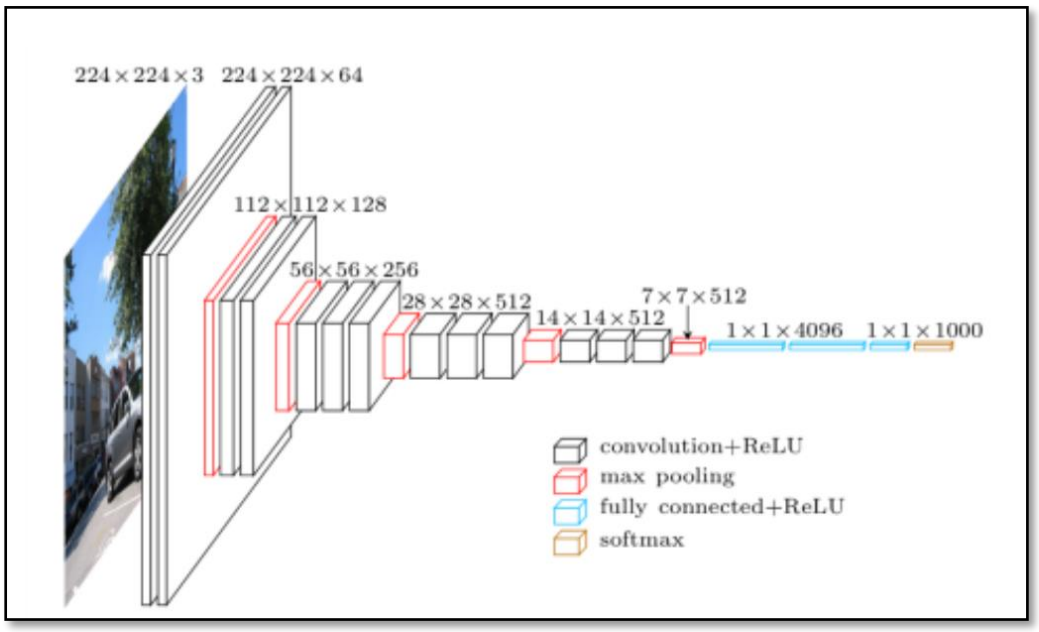
Para isso, usamos a biblioteca openpose para realizar estimativa de pose humano com cinco pessoas onde é necessário detectar e localizar as principais partes de articulações do corpo (por exemplo, ombros, tornozelo, joelho, pulso etc.). A biblioteca openpose, ferramenta para detecção de filiação das partes do corpo humano. É um método que se baseia em medidas de partes do esqueleto humano, como braços e pernas, e em mapas térmicos das articulações dos esqueletos, obtidos de poses 2D estimadas nas imagens.

Como a detecção de esqueleto humano requer a localização e mapeamento de parte/ponto-chave de corpo humano para mapas térmicos das articulações dos esqueletos. Tendo essas informações, possibilita ligação das partes filiais do corpo para obter a estimativa de pose humano. Essa saída consiste em mapas de confiança e mapas de afinidade são analisados por inferência gananciosa para produzir os pontos-chave 2D para todas as pessoas na imagem.

4.1 Arquitetura da Rede Neural Convolucional (CNNs) -VGGNet

Arquitetura VGG é uma CNN básica que consiste em uma camada convolucional e uma camada de agrupamento. No entanto, conforme mostrado na Figura 1 abaixo, caracteriza-se pelo aprofundamento de todas as 16 camadas. Ele passa por uma série de camadas convolucionais com um pequeno filtro de 3x3 conforme mostrado na figura 1, a camada convolucional é repetida 2 a 4 vezes seguidas para reduzir o tamanho pela metade, colocando a camada de agrupamento. E, finalmente, ele passa pela camada totalmente conectada e gera o resultado. (Will et al, 2018)

Figura 1: Arquitetura VGGNet



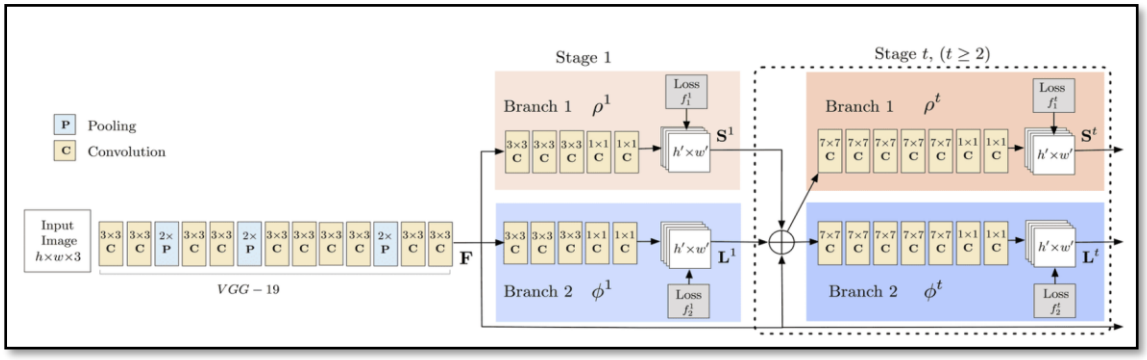
Will et al, (2018)

A rede funciona da seguinte forma: primeiramente passa a imagem inteira como entrada para uma CNN, depois prevê conjuntamente mapas de confiança para detecção de partes do corpo e gera os campos de afinidade para associação de partes, por último, na etapa de análise/previsão executa duas partes: encontrar partes do corpo e construir a pose de corpo inteiro.

Segundo Cao, et al (2017), o modelo recebe como entrada uma imagem colorida de tamanho w × h e produz, como saída, as localizações 2D dos pontos-chave para cada pessoa na imagem. A detecção ocorre em três etapas:

- **Etapa 0:** As 10 primeiras camadas VGGNet são usadas para detectar todos os pontos do modelo;
- **Etapa 1:** Uma CNN de múltiplos estágios de dois ramos é utilizada para prever um conjunto de mapas de confiança 2D (S) de localizações de pontos específicos ((por exemplo, cotovelo, joelho, etc.);
    - No primeiro ramo *Stage 1 – Branch 1*, prevê os mapas de confiança
    - Em *Stage 1 – Branch 2*, é gerado o mapa de afinidades de partes que codificam o grau de associação entre as partes;
    - Depois de cada estágio, as previsões de cada ramo são concatenadas com as características da imagem e repassadas para a próxima fase.

Figura 2: A arquitetura da CNN de duas filiais em várias fases



Cao, Zhe, et al, (2019)

Figura 3: Mostrando mapas de confiança para o nariz da imagem fornecida

Fonte: Elaborado pelo autor

A segunda ramificação prevê um conjunto de campos vetoriais 2D (L) de afinidades de peças, que codificam o grau de associação entre as peças. Na Figura 4 abaixo é mostrada a afinidade entre o pescoço e o quadril direito.

Figura 4: Mostrando mapas de afinidade de peças entre o pescoço e o quadril direito

Fonte: Elaborado pelo autor

Etapa 2: Os mapas de confiança e afinidade são analisados por inferência gananciosa para produzir os pontos chave 2D para todas as pessoas na imagem.

4.2 Modelos pré-treinados para estimativa de pose humana

Os autores (Cao, Zhe, et al,2019) compartilharam dois modelos pré-treinado modelo treinado no *Multi-Person Dataset (MPII)* e o outro é treinado no dataset COCO. O modelo caffe (COCO) produz 18 pontos, enquanto o modelo MPII produz 15 pontos. Apesar de os autores usarem dois modelos, mas no nosso caso usaremos o modelo COCO que produz 18 pontos chave como vimos anteriormente.

O formato de saída para o modelo COCO (18 pontos chave): Nariz - 0, Pescoço - 1, Ombro Direito - 2, Cotovelo Direito - 3, Pulso Direito - 4, Ombro Esquerdo - 5, Cotovelo Esquerdo - 6, Pulso Esquerdo - 7, Quadril Direito - 8, Joelho Direito - 9, Tornozelo direito- 10, quadril esquerdo - 11, joelho esquerdo - 12, tornozelo - 13, olho direito - 14, olho esquerdo - 15, orelha direita - 16, orelha esquerda - 17, fundo – 18. As saídas plotadas em uma pessoa são mostradas na imagem abaixo.

Figura 5: Saída de 18 pontos chave do modelo COCO com uma pessoa

Fonte: Elaborado pelo autor

Figura 6: Saída de 18 pontos chave do modelo COCO com várias pessoas

Fonte: Elaborado pelo autor

5. Implementação e resultado

5.1 Objetivo

Este relatório tem por objetivo apresentar um estudo de caso sobre aplicação de openpose para estimativa de pose 2D corpo humano em tempo com cinco pessoas.

5.2 Descrição de problema:

Trata-se de um problema da área de visão computacional onde o objetivo é implementar o openpose básico para estimativa de Pose 2D no Corpo humano que seja capaz de capturar várias pessoas (entre cinco a vinte pessoas que reconhecem cabeça e pé) em tempo real de 5fr/s.

5.3 Dados

Na implementação deste projeto, usamos os conjuntos de dados pré-treinado de datasetCOCO, que possui mais de 200.000 imagens e instâncias de 250.000 pessoas rotuladas com pontos-chave (a maioria das pessoas no COCO em escalas médias e grandes). Anotações em train e val (com mais de 150.000 pessoas e 1,7 milhão de pontos-chave rotulados) estão disponíveis publicamente. e ganhou o desafio de pontos chave COCO em 2016, o seu site oficial é o http://cocodataset.org/#keypoints-2018

5.4 Linguagem e as Biblioteca

Escolhemos a linguagem C++ para desenvolver este projeto, porque além de ser uma linguagem de programação rápida, também é altamente dinâmica e em evolução, ficou perfeito para nosso cenário. Com Biblioteca OpenCV significa (Open Source Computer Vision Library), é uma biblioteca de programação, de código aberto, que visa oferecer um vasto ferramental para tratamento de imagens e vídeos, visão computacional e reconhecimento de padrões. O OpenCV possui uma estrutura modular, o que significa que o pacote inclui várias bibliotecas compartilhadas ou estáticas. O seu site oficial é o https://opencv.org. Utilizamos a imagem do sistema operacional Linux/Ubuntu para implementação.

O link para acessar código fonte de projeto: < https://github.com/issufibadji/PoseHumano >. As orientações de pré-requisitos de como executar projeto estão no arquivo README.md.

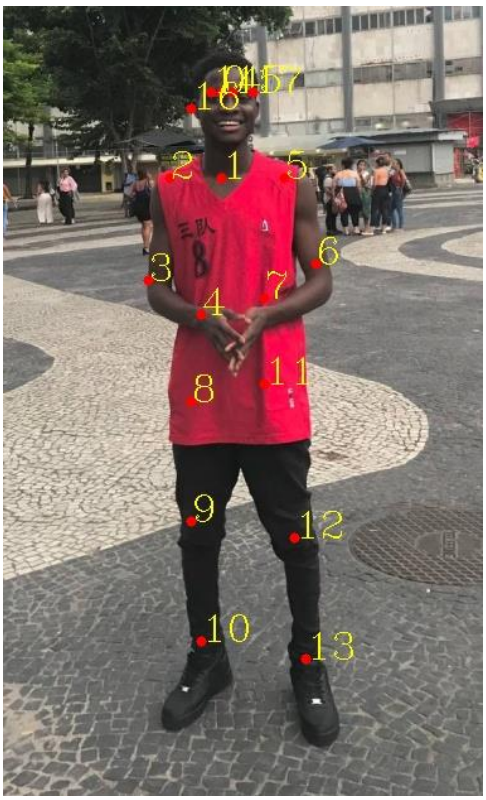
5.5 Detalhes da implementação e resultados

A detecção de esqueleto humano requer a localização e mapeamento de parte/ponto-chave de corpo humano para mapas térmicos das articulações dos esqueletos. Tendo essas informações, possibilita ligação das partes filiais do corpo para obter a estimativa de pose humano. Essa saída consiste em mapas de confiança e mapas de afinidade são analisados por inferência gananciosa para produzir os pontos-chave 2D para todas as pessoas na imagem. Para mais detalhes explicaremos a implementação e resultado obtido:

Etapa 1: primeiramente entramos como uma imagem colorida de tamanho (h x w), mostrada na Figura 1 e 2 à esquerda. E em seguida, produzimos, como saída, as localizações 2D dos pontos-chave para cada pessoa na imagem plotada na Figura 2 e 2 à direita. As saídas são plotadas abaixo:

Fig. 7. Entrada da imagem de uma pessoa e saída de pontos-chave em 2D

Imagem original



Pontos-chave de esqueleto

Fonte: Elaborado pelo autor

Fig. 8. Entrada da imagem de várias pessoas e saída de pontos-chave em 2D

Imagem original

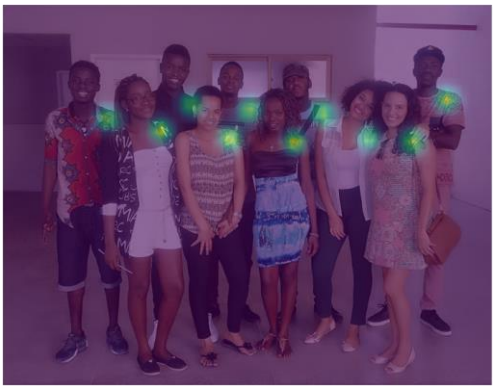
Pontos-chave de esqueletos

Fonte: Elaborado pelo autor

Etapa 2: Prevemos conjunto de mapas de confiança em 2D (S) para detecção de partes específicas do corpo, por exemplo, na Figura 9 a imagem à esquerda apresenta o resultado dos mapas de confiança para o **nariz** para imagem fornecida, ao passo que na imagem à direita mostra os mapas de confiança para o ombro esquerdo para a imagem fornecida.

Figura 9: Mapas de confiança para o nariz e ombro esquerdo

Mapas de confiança de nariz



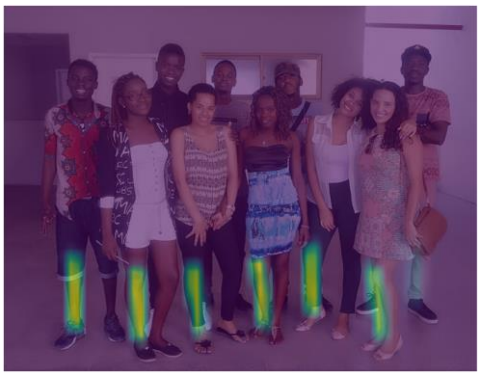
Mapas de confiança de ombro esquerdo

Fonte: Elaborado pelo autor

Etapa 3: Depois da previsão do mapa de confiança, geramos os campos de afinidade, que foi produzido através de campos vetoriais 2D (L) de afinidades de peças, que codificam o grau de associação entre as partes. Na Figura abaixo no lado esquerdo é mostrada a afinidade entre o **pescoço e o quadril direito** e no lado direito da mostrada a afinidade entre **joelho direito e o tornozelo direito**

Figura 10: Mapas de afinidade entre o pescoço e o quadril direito e joelho direito e o tornozelo direito

Pescoço e o quadril direito



Joelho direito e o tornozelo direito

Fonte: Elaborado pelo autor

Etapa 4: Por último, fizemos a análise e previsão em duas partes:

- Primeiramente, procuramos encontrar partes do corpo;
- Em seguida, desenhamos os esquilos inteira, ou seja, construímos a pose de corpo inteiro. Na Figura 11 abaixo é mostrado o resumo dos resultados obtidos durante o experimento.

Fig. 11. Resumo da pose humana em 2D com uma/várias pessoa (s)

6. Considerações finais

Neste trabalho, foi relatado o procedimento da implementação estimativa de pose de corpo humano em 2D com cinco pessoas em tempo real, juntamente com a utilização do método openpose. Portanto, compreendemos que o reconhecimento de gestos e ações é um fenômeno que merece atenção, tendo a compreensão de ação de uma pessoa, é possível prever o que ele está fazendo.

Com essas informações é possível construir muitas aplicações, que pode ajudar minimizar vários problemas sociais, por exemplo: Vigilância visual para monitoramento de crime e ação anormais, vigilância visual para motorista ao dirigir, detecção de mão para semântica da linguagem de libra (tradução da linguem libra), monitoramento do ambiente com cenas dinâmicas e desordenados, entre outra atividade ações ilegais.

Por fim, com base nos detalhes dos trabalhos supracitados e do método utilizado para captura de esqueleto, ficou fácil de entender o funcionamento e a implementação de estimativa de pose humana. Os resultados realmente demonstram o poder do método openpose por meio de diferentes soluções de estimativa de pose humana apontadas acima. Garças isso, torna-se fácil e eficiente descobrir padrões e informações que podem ser úteis para previsões futuras em análise de comportamento de ação humano.

Referências Bibliográficas

BUDUMA, L. et al. **Fundamentals of Deep Learning.** ed.Shiny KalapuraKel. 2017

CAO, Zhe et al. **Realtime multi-person 2d pose estimation using part affinity fields**. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2019. p. 7291-7299.

FACELI, K. et al. "**Inteligência Artificial:** Uma Abordagem de Aprendizado de Máquina", *LTC, Rio de Janeiro.*2011

FELTRIN, Fernando Belomé. **Ciência de Dados e Aprendizado de Máquina**: Uma abordagem prática as redes neurais artificiais. 2019. 296 p.

FORSYTH, D. A.; PONCE, J. **Computer Vision: A Modern Approach**. 2nd Edition. Prentice Hall, 2012

GOODFELLOW, B. et al. **Deep Learning.** ed.MIT Press. 2016

WANG, Rui et al. **Human pose estimation with deeply learned multi-scale compositional models**. **IEEE Access**, v. 7, p. 71158-71166, 2019.

NASH, Will; DRUMMOND, Tom; BIRBILIS, Nick. **A review of deep learning in the study of materials degradation**. npj Materials Degradation, v. 2, n. 1, p. 1-12, 2018.