Universidade Fernando Pessoa

Mestrado em Computação Móvel

Visão Computacional

Rodrigo Soares - 36824 e Luísa Costa - 37151

Projeto Prático de Visão Computacional

Aquisição e Processamento de imagens com uma câmara estéreo para reconhecimento de atividade humana (human activity recognition)

Porto

Dezembro/2021

Universidade Fernando Pessoa

Praça 9 de Abril, 349

P-4249-004 Porto

Tel. +351-22550.82.70

Fax. +351-22550.82.69

geral@ufp.pt

Índice

1. Introdução

O reconhecimento de atividades humanas com base em vídeo é uma área de estudo que tem vindo a crescer, no entanto, continua a ser uma área complexa devido às variadas nuances existentes, tais como, posição da câmara, luminosidade e escala da cena no processo de captura além das possíveis combinações de ações que os atores podem efetuar para uma mesma atividade [1].

Este trabalho teve como objetivo a criação de um dataset dentro das instalações da UFP onde foram adquiridas várias sequências de vídeo em vários locais envolvendo pessoas em diversas atividades como andar, paradas, sentadas, a ler, e outras.
Seguidamente foram aplicados algoritmos e técnicas de Visão Computacional usando python para a análise e classificação de atividades das pessoas, nomeadamente contagem de pessoas, identificação de máscara, deteção de movimento e deteção de atividade.
Este relatório serve como descrição do processo efetuado e de todas as etapas que foram efetuadas para conseguir atingir os referidos objetivos.

2. Descrição do problema

Em essência, para se poder fazer a classificação de atividades humanas antes é necessário reconhecer a presença de pessoas nas sequências de vídeo capturadas. Uma das metodologias possíveis é usar deteção de objetos em tempo real, para a qual existem diversas abordagens. Neste trabalho foi utilizado o sistema YOLOv3[6] devido à sua velocidade superior e precisão comparável em relação a modelos similares.

Contudo, mesmo sendo capaz de detetar pessoas, antes é preciso compreender o que é o reconhecimento de atividades humanas. A diferença de esta classificação para outra, é que na classificação humana, uma análise estática de uma atividade não nos permite determinar qual é essa atividade, sendo então necessário múltiplos pontos temporais para a determinar.

3. Estado da Arte

A área de reconhecimento de atividade humana tem sido bastante procurada nos últimos anos devido à sua versatilidade. Existem inúmeras maneiras de registar atividade humana, tais como usando sensores de movimento, sons e, como utilizado neste trabalho, capturas de vídeo. Existem, no entanto, desafios relacionados com a previsão da atividade humana com base em vídeos tais como determinar quais os parâmetros e dimensões necessárias, sendo estes fatores altamente dependentes dos recursos utilizados.

Nos dias atuais temos uma panóplia de variantes de redes neuronais convolucionais que nos permitem classificar atividades humanas em conjuntos de imagens/vídeos, contudo a precisão destas depende muito da qualidade dos dados fornecidos, o que depende bastante do pré-processamento efetuado nos mesmos. Este tratamento pode ser por base em filtros nas imagens ou frames de vídeos para melhor extração de características nomeadamente contornos, realces de formas, limpeza de ruído…

4. Descrição do trabalho realizado

4.1 Captura de vídeos stereo

Inicialmente foram obtidas filmagens dentro das instalações da UFP com recurso a uma câmara estereoscópica (Intel Realsense D455). Estas filmagens são compostas de vídeos com canais RGB e respetivos vídeos com indicação da profundidade.

Resolução de gravação	640x480 px
Taxa de gravação	30 fps
Canais de cor	3 (RGB) + 1 (Depth)

Ao todo foram gravadas 16 sequências de vídeo em diversos locais com períodos de gravação entre 10 segundos a 2 minutos.

4.2 Deteção de objetos

Após obter os vídeos, a etapa seguinte foi de analisar os mesmos para se ser capaz de detetar objetos, com foco principal em pessoas. Para tal recorreu-se ao sistema YOLOv3. Para cada vídeo, forneceu-se as suas frames ao sistema e este retornava a mesma e informações sobre os objetos detetados (tipo, posição e dimensão). Um dos tipos obtidos foi “Person” o que é o desejado, como tal restringiu-se os objetos a analisar apenas a esse tipo, tendo então para cada frame dos vídeos a lista das pessoas detetadas e a posição de um retângulo delimitante de cada.

4.3 Extração do dataset

A etapa seguinte foi de extrair de algumas das frames dos vídeos (aproximadamente 25% do total) as sub-imagens das pessoas detetadas. Com isto obteve-se para cada pessoa a sua imagem RGB e imagem de profundidade. Além disso, para facilitar a classificação das mesmas, decidiu-se obter também as imagens transformadas com base em filtros de deteção de bordas (baseado em erosão) e de melhoramento de contraste (baseado em equalização de histogramas).

De seguida estes conjuntos de 4 imagens cada foram catalogados em 12 categorias manualmente com base em diretórios para servirem de dataset de treino para um modelo de rede neuronal de classificação. É de notar que para um conjunto de imagens, este pode se encontrar em múltiplas categorias em simultâneo.

Categoria	Número de conjuntos (4 imagens cada)
walking	1099
with_mask	858
standing	699
sitting	517
walking_up_stairs	134
without_mask	127
walking_down_stairs	117
on_the_phone	71
writing	37
running	33
trotinete	23
wrong_use_mask	4

4.4 Pré-processamento do dataset

Para facilitar a leitura dos dados, criou-se um ficheiro .csv no qual para cada conjunto único de imagens foram catalogadas as categorias onde se encontra. Estas seguem um formato de um array binário onde 1 representa uma classe onde pertence e 0 uma classe onde não pertence, mantendo sempre a ordem correta das classes.

Este ficheiro foi por sua vez usado para importar o dataset para uma dataframe. Para tal, os conjuntos de imagens foram importados, redimensionados para uma escala constante de 64x64 pixels e por sua vez as imagens de cada conjunto foram concatenadas de modo a que ao invés de 4 imagens com 3 canais cada, apenas haver uma imagem com 12 canais (3 canais por imagem) a qual foi armazenada com as respetivas categorias na dataframe.

4.5 Construção e treino do modelo

Para a classificação deste dataset recorreu-se a uma rede neural convolucional com a estrutura indicada na Figura 1 que recebe como entrada estruturas de dados de 64x64x12 e retorna como saída um conjunto de 12 valores probabilísticos os quais indicam qual a probabilidade de, para uma imagem de entrada, ser cada respetiva classe. Este modelo foi treinado durante 30 epochs com batches de 2 a 2 e um ritmo de aprendizagem de 0.001 com base no optimizador “Adam”. Para avaliar a qualidade do modelo usou-se o classificador de

perda “categorical cross-entropy” pois é uma função de perda que é usada em tarefas de classificação de múltiplas classes.

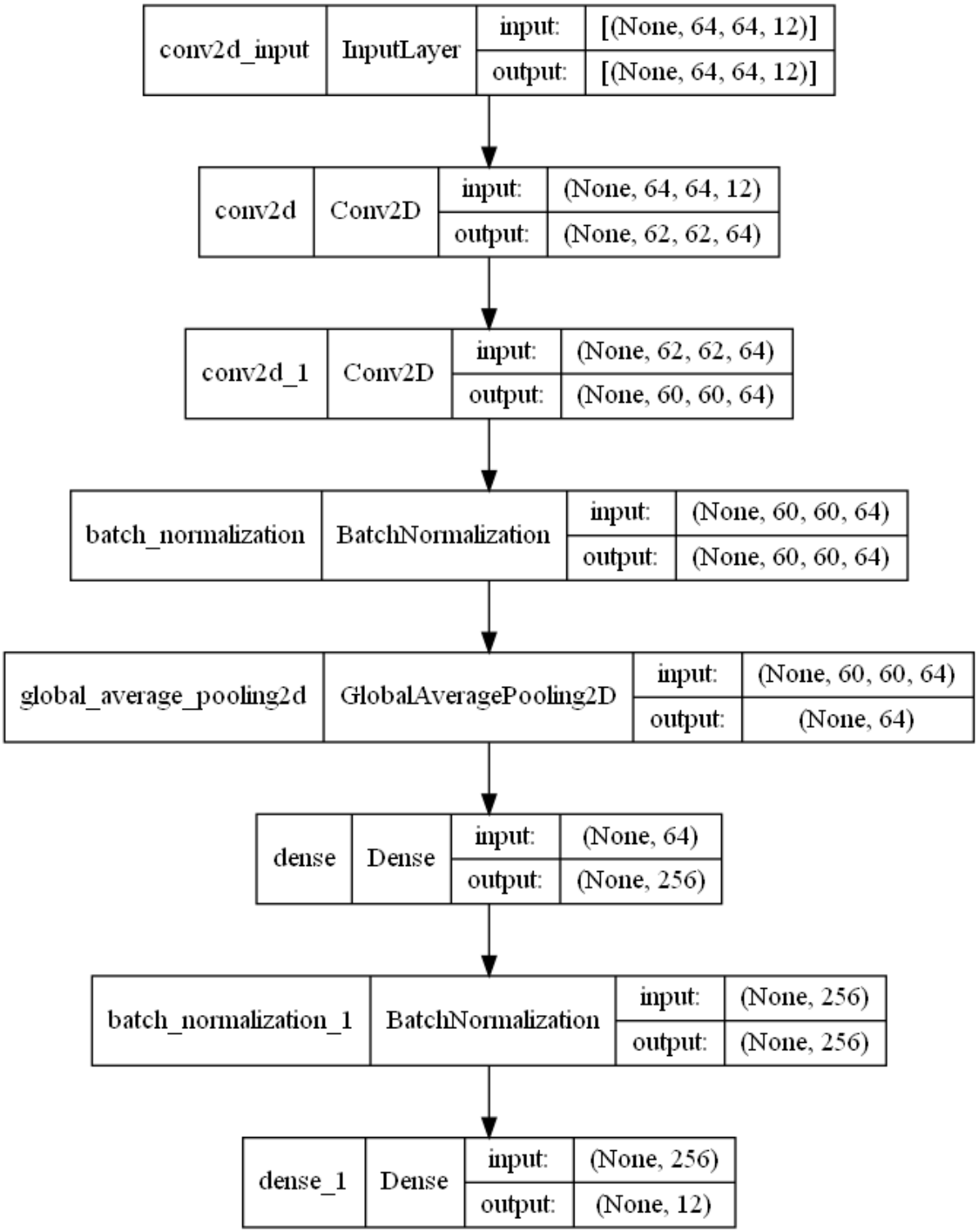

Fig 1 - Estrutura do modelo criado usando CNN

4.6 Análise e processamento dos vídeos

Tendo o modelo treinado, foi então aplicado aos vídeos capturados todo o processo de extração de pessoas e de criação de imagens filtradas às suas frames e foram essas alimentadas ao modelo obtendo assim as probabilidades de cada classe para a respetiva pessoa detetada. Contudo é de notar que algumas classes são mutuamente exclusivas, tais como “sitting” e “running”. Para resolver, em cada conjunto de categorias mutuamente exclusivas, foram apenas escolhidas as que não só possuíam um valor probabilístico acima da média, como também as mais elevadas do conjunto.

Conjuntos de exclusividade
running, sitting, standing, walking, walking_down_stairs, walking_up_stairs, trotinete
with_mask, without_mask, wrong_use_mask
on_the_phone, writing

Finalmente foram reconstruídos os vídeos com informações adicionadas a estes, nomeadamente um contador das pessoas detetadas e uma seleção retangular legendada com as categorias para cada pessoa na frame.

5. Análise de Resultados

Após todo o processo acima descrito foi possível detetar vários tipos de movimentos, uso de máscara, indicação de atividade da pessoa e contagem do número de pessoas presentes nas frames. A deteção dos objetos não é garantida, sendo influenciada por obstruções, luminosidade e escala, contudo dadas estas condicionantes, os resultados foram mais que desejáveis.

A precisão obtida durante o treino do modelo foi de 56%. Esta precisão deve-se ao facto de o tamanho do dataset ser reduzido e de o tamanho dos dados de entrada estar limitado a uma escala pequena, já que as imagens estão a sofrer um redimensionamento para 64x64 pixels. Idealmente esta escala deveria ser superior às imagens originais, contudo devido a limitações computacionais de memória, tal não foi possível. Uma possível alternativa a explorar seria o treino do modelo importando as imagens à medida da sua necessidade, contudo o tempo de processamento seria exponencialmente elevado.

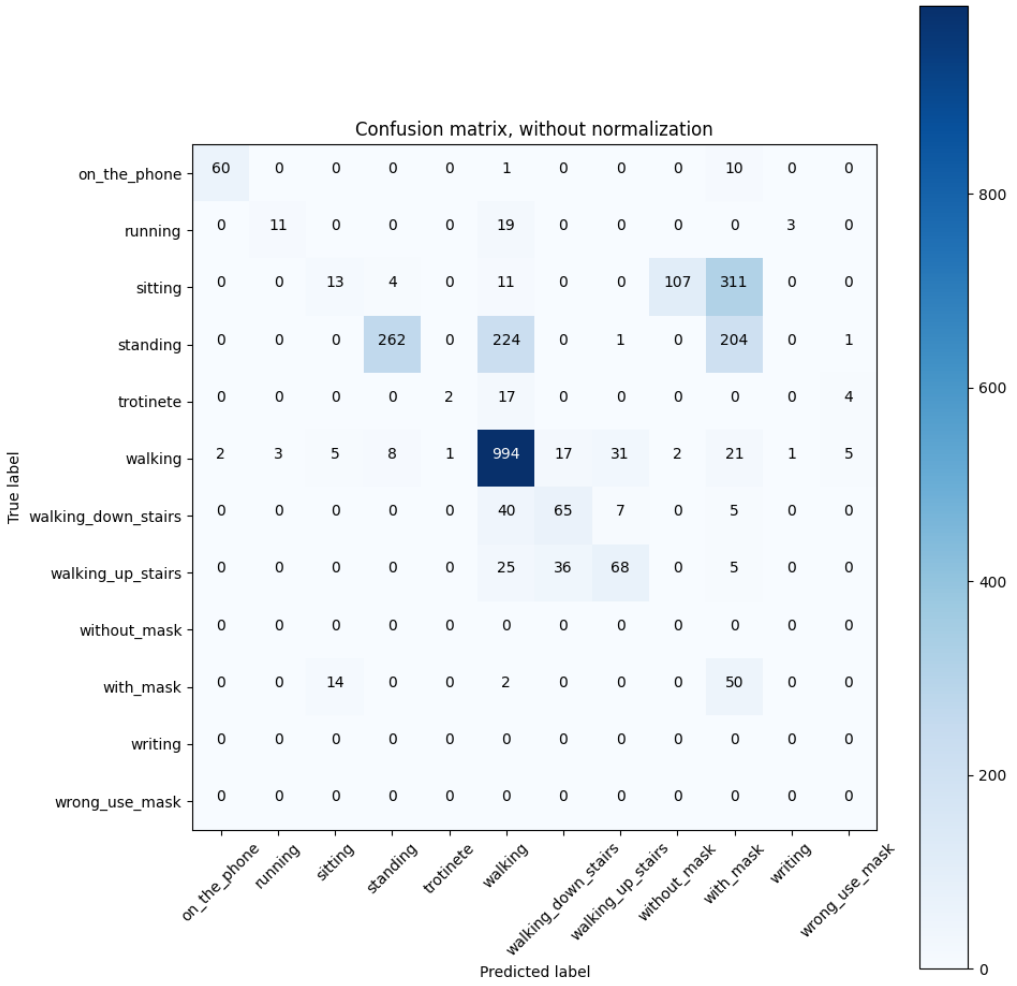

Fig 2 - Matriz de confusão do modelo

6. Conclusões

Em retrospetiva, os resultados obtidos foram positivos. Os filtros utilizados contribuíram para o melhoramento da perceção das imagens pelo modelo.

Numa próxima fase seria ideal acrescentar mais vídeos ao dataset, assim como frames ao dataset de treino e também aumentar a escala de redimensionamento das imagens para que seja possível melhorar o resultado da precisão.

Outra melhoria seria o uso de redes neuronais recorrentes para melhor identificação de ações, pois estas requerem contexto temporal que não está a ser utilizado de momento.

7. Referências

1. A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features," in IEEE Access, vol. 6, pp. 1155-1166, 2018, doi: 10.1109/ACCESS.2017.2778011.

2. Learn OpenCV. (2021, May 3). Introduction to Video Classification and Human Activity Recognition. Learn OpenCV | OpenCV, PyTorch, Keras, Tensorflow Examples and Tutorials. https://learnopencv.com/introduction-to-video-classification-and-human-activity-recognition/

3. Kishore K. Reddy and Mubarak Shah. 2013. Recognizing 50 human action categories of web videos. Mach. Vision Appl. 24, 5 (July 2013), 971–981. DOI:https://doi.org/10.1007/s00138-012-0450-4

4. Multi-output Classification Example with MultiOutputClassifier in Python https://www.datatechnotes.com/2020/03/multi-output-classification-with-multioutputclassifier.html

5. A Gentle Introduction to Object Recognition With Deep Learning https://machinelearningmastery.com/object-recognition-with-deep-learning/

6. YOLO: Real-Time Object Detection https://pjreddie.com/darknet/yolo/

Anexos

- Github do Projeto:
  https://github.com/Infor-Master/VCOP_classes/tree/master/Project
- Yolo Weights:
  https://pjreddie.com/media/files/yolov3.weights