Instituto Universitário de Lisboa

Departamento de Ciências e Tecnologias da Informação

Reconhecimento de Interações Cliente-Produto em Espaços de Vendas

Francisco Marques Gracias

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Informática e Gestão

Orientador(a):
Doutor Tomás Gomes da Silva Serpa Brandão, Professor Auxiliar,
ISCTE-IUL

Coorientador(a):
Doutor Luís Miguel Martins Nunes, Professor Auxiliar,
ISCTE-IUL

Outubro, 2018

Agradecimentos

A realização da presente Dissertação de Mestrado revelou ser a experiência mais marcante do meu percurso Universitário, acompanhada do apoio imprescindível de várias pessoas, as quais é dedicado este sincero agradecimento.

Ao meu orientador, Professor Doutor Tomás Gomes da Silva Serpa Brandão, um especial obrigado por partilhar o seu entusiasmo pela investigação junto dos alunos, assim como pela oportunidade de trabalhar sobre a sua orientação, onde o gratifico pela sua consistente boa disposição, competência, disponibilidade e incentivo.

Ao meu coorientador, Professor Doutor Luís Miguel Martins Nunes, muito obrigado pela sua predisposição em integrar este desafio, partilha de conhecimentos e eficácia na orientação.

A ambos os orientadores reforço o meu agradecimento pela forma como articularam proactivamente a sua disponibilidade em prol da orientação, ajuda prestada na angariação de todos os recursos necessários à concretização da tese e esclarecimento contínuo de dúvidas e incertezas.

À família, e em especial aos meus pais, Teresa Marques e Mário Gracias pelo apoio incondicional ao nível económico e do meu bem-estar, assim como a oportunidade concedida de formação académica e investimento nos meus interesses pessoais.

Ao meu avô, José Cruz Marques, e, recentemente falecida, avó Maria Lucília, pelo vosso eterno contributo para a meu desenvolvimento pessoal, na forma de inúmeros ensinamentos, que procuro refletir no meu trabalho.

Ao meu irmão e amigos próximos, pela motivação e apoio dado nos momentos maior dificuldade.

À entidade Vitruvius FabLab, nomeadamente ao Sr. João Pedro Sousa, fico grato pela cedência de materiais para montar o cenário para recolha de dados.

À entidade ISTAR, destacando a Sra. Fátima Estevens, pela ajuda no processo de aluguer do sensor *Microsoft Kinect v2*, o qual foi fundamental para a concretização da tese.

Por fim, ao grupo de investigadores VRAI e empresa Axians, pelo interesse no trabalho desenvolvido e partilha de recursos e conhecimentos.

Resumo

O reconhecimento de atividades humanas baseado em visão por computadores é uma área de investigação desafiante com crescente interesse entre os investigadores e empresas. Com a introdução de sensores RGB-D, que adiciona a dimensão de profundidade às câmeras convencionais, é possível gerar modelos de esqueletos em tempo real. Com base em atributos extraídos do esqueleto e em modelos de aprendizagem automática treinados é possível reconhecer as atividades humanas.

Nesta dissertação, propõe-se um modelo para reconhecer interações de clientes com produtos em prateleiras de lojas com base em informação do esqueleto e RGB-D, assim como algoritmos existentes para deteção de objetos e gestos. Estes algoritmos são interligados num único sistema e testados num ambiente de loja simulado, caracterizado por interações humano-objeto, necessidade de acompanhar simultaneamente diferentes atividades de clientes em tempo real e um ângulo de visão típico de câmeras em lojas (vista superior) que potencia oclusões entre sujeitos ou partes do corpo deste.

As principais contribuições deste estudo são a introdução de um novo modelo que combina reconhecimento de objetos e gestos e a análise detalhada dos resultados sobre diversas perspetivas consideradas pertinentes.

Acresce o conjunto de dados recolhido que está disponível para fins de investigação, como o desenvolvimento, melhoria e comparação de desempenho de modelos destinados a este contexto aplicacional. Três cenários com quatro tipos de produto e graus de complexidade distintos são avaliados - um único cliente a interagir com duas prateleiras, dois clientes e uma prateleira para cada e dois clientes disputando duas prateleiras.

No modelo desenvolvido, o reconhecimento de interações com a prateleira passa pela deteção de extensões e flexões do braço trama-a-trama, que posteriormente são generalizadas em gestos e interações para um intervalo de tramas. O modelo desenvolvido apresenta um *f1*-score médio de 69,78% para deteção da extensão/flexão do braço e 66,46% para deteção do tipo de produto. Com base na agregação de informações de deteção de objetos e gestos, são reconhecidas 53.97% das interações de prateleira testadas (*recall*) e detetadas corretamente 30.47% das vezes (*precision*).

Palavras-chave: *Deteção de Ações de Clientes; Classificação de interações com prateleiras; Deteção de Produtos; Reconhecimento de gestos; Seguimento de esqueletos.*

Abstract

Computer vision-based human activities recognition is a challenging research area with increasing interest amongst researchers and companies. The introduction of RGB-D sensors. which add the depth dimension to the conventional colored 2D cameras, allows real-time skeleton model generation of humans. This skeleton data provides meaningful information that enabled researchers to model human activities by training machine learning models and later utilize them to recognize activities.

In this dissertation, we propose a model to recognize customer interactions with products in store's shelves based on RGB-D and skeleton data, as well as existing algorithms for gesture and object detection. We demonstrate how those existing algorithms perform in an integrated system tested in a simulated retail store context, particularly characterized by human-object interactions, the capacity to simultaneously track in real-time different customer's activities and a field of view captured by the sensor that is typical in retail environments (top view), which makes it prone to occlusions between subjects and body parts.

The main contributions of our study are the introduction of a novel model that combines object and gesture recognition as well as detailed performance metrics regarding different analytical perspectives.

The collected dataset is available for researching purposes, namely to allow different model's development, improvement and performance comparison in this specific research area. Three scenarios with four types of products and different recognition complexities are evaluated – a single customer interacting with two shelves, two customers interacting with a one shelf each and two customers disputing two shelves.

In the developed model, recognizing shelf interactions is done through the generalization of frame by frame arm extension/flexion detections in gestures and interactions regarding specific frame intervals. The developed model has a *f1-score* of 69.78% for arm extension/flexion detection and 66.46% for product type detection. Based on the aggregation of gesture and object detection information we recognize 53.97% of the existing shelf interactions (*recall*) with a *precision* of 30.47%.

Keywords: *Customer Action Detection; Shelf Interaction Classification; Skeleton Tracking; Product Detection; Gesture Recognition.*

Índice

Índice de Quadros

Índice de Figuras

Capítulo 1 – Introdução

1.1 Apresentação do tema

O reconhecimento de atividades humanas é uma importante temática na investigação em torno da visão por computadores. O seu objetivo mais alargado prende-se com a capacidade de automaticamente detetar e analisar atividades humanas com base em informação captada por sensores.

No caso do atual projeto de dissertação, pretende-se desenvolver um modelo de extração e classificação de ações de clientes enquadrado no contexto de uma loja, recorrendo ao sensor *Microsoft Kinect* e a algoritmos de aprendizagem automática.

A viabilidade económica da aquisição deste tipo de sensores, assim como a relevância da informação visual obtida e a conveniência da sua utilização, veio promover um rápido crescimento na investigação em torno da aplicabilidade de dados visuais 3D para a deteção e seguimento do corpo humano, bem como o reconhecimento e análise de ações humanas [1].

O desenvolvimento de um modelo com esse objetivo e aplicável a um contexto de loja é particularmente interessante na inovação do setor de retalho e comércio onde a compreensão dos comportamentos dos clientes poderá ser de máxima utilidade para o desenvolvimento adequado de modelos estratégicos de gestão.

O estudo comportamental dos clientes em loja é também relevante para investigar o processo de tomada de decisão destes e contribui para a avaliação das estratégias de gestão implementadas. Adicionalmente, a forma como os clientes interagem com o espaço, na forma de ações, permite às lojas oferecer serviços não solicitados, mas baseados na inteligência dos sistemas, ou otimizar a organização do espaço em loja.

Importa assim averiguar com que precisão se conseguem classificar as ações típicas neste contexto, assim como compreender até que ponto as limitações atuais no reconhecimento de ações poderão inviabilizar ou não a sua aplicabilidade num contexto real.

O recurso à aprendizagem automática aparenta ser uma abordagem com enorme potencial para desempenhar tarefas como o reconhecimento de ações humanas. Os modelos usados nesta subárea da inteligência artificial podem adaptar-se a novos dados e reconhecer padrões cuja complexidade muitas vezes ultrapassa a capacidade humana.

No entanto, problemáticas como o alcance limitado do sensor ou a falta de robustez perante oclusão de partes do corpo constituem-se como alguns dos desafios na atual investigação que utiliza este tipo de sensores para reconhecimento de ações humanas.

Acresce também que a generalidade dos estudos é feita com imagens geradas sem oclusões, através do posicionamento das câmeras ao nível do sujeito e não em pontos altos como, por exemplo, os das câmeras de vigilância.

Como tal, a construção do modelo proposto na presente dissertação irá procurar definir a abordagem mais robusta a esta problemática potenciando o seu funcionamento em cenários reais.

1.2 Questões e objetivos de investigação

O objetivo alargado do projeto de dissertação é averiguar se, com base em algoritmos conhecidos para deteção e classificação, é possível criar um modelo adaptado ao reconhecimento de interações de clientes com prateleiras em loja (i.e., tirar/colocar produtos em prateleiras).

A validação do modelo proposto deverá ser feita atingindo as métricas de desempenho encontradas no atual estado-de-arte para o reconhecimento de atividades humanas típicas no contexto de uma loja, como a interação com produtos.

Para tal, é necessário que existam conjuntos de dados específicos deste contexto aplicacional já testados com outros modelos de forma a possibilitar a comparação direta de resultados.

Alternativamente, a hipótese de investigação será desenvolvida, testada e validada com um conjunto de dados extraído em ambiente simulado através de um sensor *Microsoft Kinect v2*.

Adicionalmente, pretende-se testar a capacidade de fazer o seguimento (*tracking*) de mais do que um cliente, reconhecendo as suas ações separadamente.

O trabalho intrínseco ao desenvolvimento do modelo consiste principalmente na seleção, treino e teste de algoritmos para deteção e classificação de gestos e objetos, assim como a sua interligação num único sistema que permita reconhecer as interações dos clientes com os produtos em prateleiras de lojas.

O teste do sistema desenvolvido será feito por meio da análise das métricas de desempenho que melhor representem o comportamento do modelo.

1.3 Abordagem metodológica

Para garantir que uma pesquisa seja reconhecida como sólida e potencialmente relevante, tanto pelo campo académico quanto pela sociedade em geral, ela deve demonstrar que foi desenvolvida com rigor e que é passível de debate e verificação [2].

A partir da revisão da literatura, verificou-se que os conceitos da proposta metodológica *Design Science Research*, são pertinentes e aplicáveis no contexto da presente dissertação.

Design Science Research define-se em [3] como o conjunto de técnicas e perspetivas sintéticas e analíticas para desenvolver investigação em Sistemas de Informação. Envolve duas atividades principais para compreender e melhorar o comportamento de vários aspetos dos Sistemas de Informação: (1) criação de novo conhecimento através de artefactos novos ou inovadores e (2) análise do uso e/ou desempenho do artefacto com reflexão e abstração. Os artefactos criados segundo o processo intrínseco a esta metodologia incluem, por exemplo, algoritmos, linguagens, metodologias para desenho de sistemas, entre outros.

Em linhas gerais, esta metodologia caracteriza-se pelo seu foco na produção de conhecimento novo, verdadeiro e interessante para uma determinada comunidade.

Na Figura 1 apresenta-se o modelo cíclico dos processos gerais que compõem esta metodologia, pelo que ao longo da dissertação ir-se-á estabelecer o paralelismo entre o trabalho desenvolvido e a sua correspondência com os processos da abordagem metodológica.

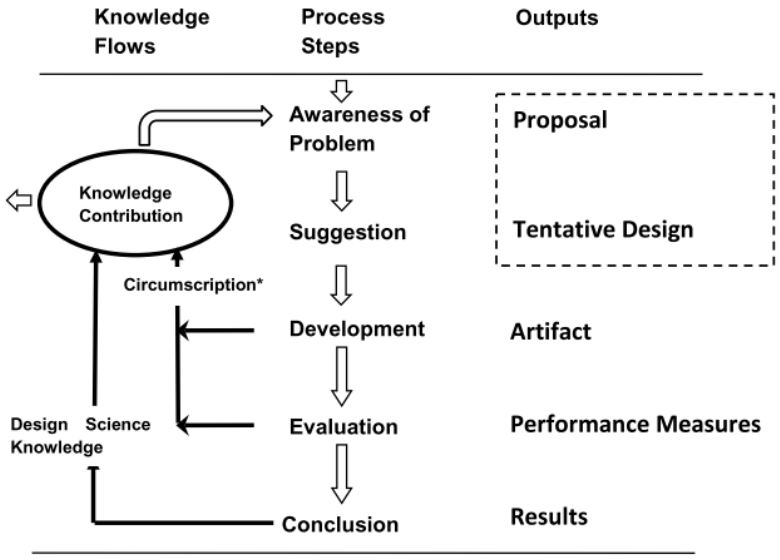


Figura 1- Modelo Cíclico da abordagem metodológica Design Science Research [3].

1.1 Estrutura da tese

Com o intuito de apresentar uma visão clara dos conteúdos da presente dissertação, na Tabela 1 apresenta-se o mapeamento dos vários capítulos face ao fluxo de trabalho associado à metodologia de investigação selecionada (*Design Science Research)*.

Capítulo	**Fase em *Design Science Research***
1 – Introdução	Identificação do problema
2 – Revisão da Literatura	Identificação do problema Relevância do tema
3 – Desenvolvimento da Solução	Sugestão (Proposta de artefacto) Desenvolvimento do artefacto
4 – Extração e Análise de Resultados	Desenvolvimento do artefacto (Extração de resultados) Avaliação
5 – Conclusões e Recomendações	Conclusões

Tabela 1- Mapeamento de capítulos face a etapas do processo associado à metodologia DSR.

No Capítulo 2 respeitante à revisão literária aprofunda-se o problema previamente identificado no capítulo introdutório (sec. 1.1), nomeadamente, analisando as tecnologias, algoritmos e abordagens mais populares no meio científico. Em resultado disso, o Capítulo 3 foca-se no desenvolvimento da solução, começando por propor, em linhas conceptuais, um sistema enquadrado nos objetivos e questões de investigação (sec. 1.2).

Após conceptualização do modelo, comparam-se as diversas potenciais abordagens tecnológicas que resultam da análise do estado-de-arte, fundamenta-se as escolhas tomadas e detalha-se a arquitetura e implementação do artefacto proposto. O treino do modelo, assim como a extração e avaliação de resultados estão contemplados no Capítulo 4, acompanhados de métricas de desempenho sobre múltiplas perspetivas de análise consideradas pertinentes.

Por fim, no Capítulo 5 faz-se uma apreciação geral do esforço de investigação conduzido, salientam-se as principais conclusões, contributos científicos, limitações do estudo e propostas de investigação futura.

Capítulo 2 – Revisão da Literatura

O primeiro passo da abordagem metodológica adotada consiste na identificação de um problema de investigação interessante para a comunidade. Tal ocorreu antes de se iniciar o presente projeto, por sugestão do orientador e motivada pelo interesse de empresas na concretização de um projeto similar. Assim, ao longo deste capítulo é realizada uma revisão literária do trabalho relacionado que servirá não só para aprofundar a compreensão do problema de investigação e destacar as diversas abordagens mais populares no meio científico para o solucionar, como também para delinear e comprovar a relevância do problema de investigação.

O capítulo encontra-se organizado segundo quatro secções principais:

- Tecnologias para captação de informação 3D (sec. 2.1);
- Modelos para reconhecimento de ações humanas (sec. 2.2);
- Avaliação de modelos para reconhecimento de ações humanas (sec. 2.3);
- Modelos para reconhecimento de objetos (sec. 2.4).

A primeira secção procura descrever as características principais dos sensores usados para captar informação visual 3D, compará-los relativamente ao seu desempenho e adequabilidade contextual e fundamentar a escolha do sensor a utilizar na vertente experimental da dissertação.

A secção 2.2 visa retratar o estado de arte dos modelos para reconhecer ações humanas, estando organizada em subsecções referentes às principais etapas associadas à construção destes modelos. Em cada subsecção reúnem-se as técnicas mais usadas no meio científico para criar modelos destinados a problemas semelhantes, mesmo que num contexto de aplicação diferente. No final desta secção, a revisão literária é particularizada para o contexto de loja, a fim de descrever as abordagens concorrentes, compreender as particularidades deste contexto e delinear as suas limitações atuais.

Seguidamente, a secção 2.3 foca-se na avaliação do desempenho de modelos deste tipo, destacando as métricas de desempenho utilizadas e os conjuntos de dados disponíveis que poderão permitir validação e comparação de resultados.

Por fim, na secção 2.4 aborda-se o estado-de-arte relativo aos modelos para deteção de objetos assim como as métricas de desempenho tipicamente usadas na avaliação do seu desempenho.

2.1 Tecnologias para captação de informação 3D

2.1.1 Sensores

Os sensores utilizados para obter dados 3D nos últimos 20 anos podem-se categorizar segundo três grupos [4]: sistemas de captura de movimento, sistemas 3D via stereo e sistemas 3D via profundidade. Para cada um dos grupos, são em seguida descritos os seus aspetos mais relevantes:

I - Sistemas de captura de movimento (*MOCAP*)

Este tipo de sistemas permite uma captura e análise de movimentos com base em marcadores posicionados em pontos estratégicos do objeto a analisar, como as articulações do corpo humano. Procedendo-se à triangulação de múltiplas câmeras é possível estimar a posição 3D de cada marcador. Existem também abordagens que detetam movimentos sem marcadores. Para tal, utilizam-se algoritmos especialmente desenhados que permitem analisar múltiplos fluxos óticos dos vídeo de entrada e identificam a forma humana através da deteção das suas partes constituintes.

Estes sistemas, também denominados como *MOCAP* (sigla oriunda do inglês, *Motion Capture Systems*), tem sido amplamente utilizado para animações gráficas (Figura 2), como as dos videojogos ou filmes, para analisar e aperfeiçoar a sequencialidade de mecânicas de atletas de alta competição ou monitorizar o progresso de recuperação em terapias físicas.

Neste tipo de sistemas apenas é capturada a posição 3D dos pontos selecionados e, como tal, os algoritmos que usam este tipo de informação tipicamente são construídos com base na sequência de posições e ângulos das articulações. Em termos gerais, a informação do esqueleto de articulações capturada em sistemas *MOCAP* é mais credível e menos ruidosa do que a capturada com sensores de profundidade, descritos mais à frente. No entanto, requer *hardware* (Figura 2) e *software* específicos, que tipicamente têm custos elevados e são complexos de operar, exigindo, adicionalmente, um espaço físico com requisitos específicos (Figura 2) para operar nas condições desejadas.

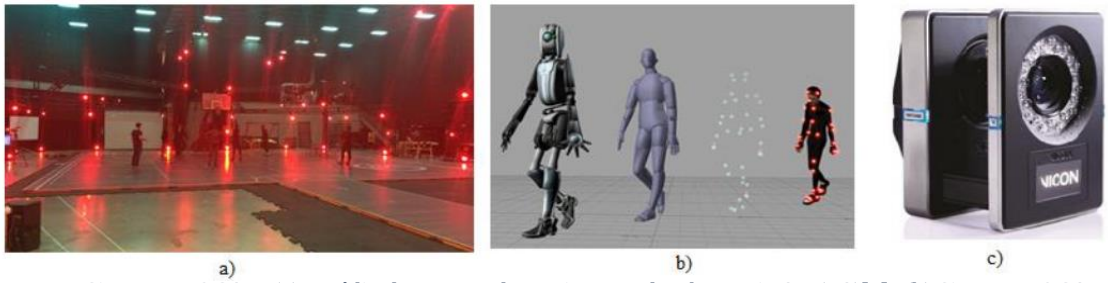


Figura 2 - Sistemas MOCAP: (a) Estúdio de captura de movimentos da Electronic Art (EA)[5]*; (b) Sistema MOCAP de marcadores ativos* [6]*; (c) Sensor para sistemas MOCAP (VICON Vantage)*[7].

II - 3D via *stereo*

Esta abordagem caracteriza-se pela reconstrução de informação 3D a partir das imagens a cores captadas por câmeras *stereo* constituídas por duas ou mais lentes com sensores de imagem/vídeo separados que podem estar localizados em diferentes posições e como tal orientados sobre diferentes ângulos (Figura 3).

O interesse do uso desta técnica em sistemas baseados em vídeo surge como alternativa aos primeiros sensores de profundidade, que tinham custos elevados e são difíceis de operar, e tem uma presença significativa em áreas como a robótica, entretenimento e sistemas automatizados.

Devido à complexidade geométrica intrínseca à reconstrução de informação 3D, o seu uso permanece uma tarefa desafiante. Reflexões, transparências, descontinuidades na profundidade e falta de textura em imagens confundem o processo de correspondência e resultam em ambiguidades na determinação da profundidade por existirem múltiplas boas correspondências.

Ao mesmo tempo, a reconstrução da informação 3D baseia-se em informação extraída no espaço de cores 2D, pelo que, tal como referido anteriormente, é sensível a variações na intensidade da luz e a separação de planos é dificultada quando existem similaridades na textura e cores. Acrescendo a necessidade das múltiplas câmeras estarem calibradas e sincronizadas, a aplicabilidade da visão *stereo* a um espectro mais alargado de contextos é limitada.

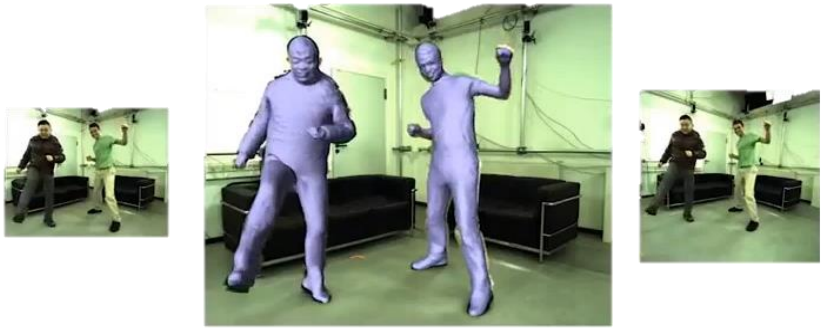


Figura 3 - Captação de vídeo com câmera stereo: (a) Vista da lente esquerda (b) Geometria 3D resultante (c) Vista da lente direita [8].

III - 3D via sensores de profundidade

A aquisição de dados 3D a partir de uma única câmera constitui-se um objetivo desejado ao longo do tempo. Dispositivos que capturam características de superfícies 3D já existem há cerca de três décadas. Estes dispositivos, conhecidos por *scanners* de alcance, oferecem como *output* uma matriz de duas dimensões com as distâncias correspondentes a cada ponto na imagem.

Os primeiros sensores deste tipo caracterizavam-se por serem demasiado caros, lentos a adquirir dados e oferecem uma pobre estimação de distâncias. Nesta altura, a investigação em torno do reconhecimento de atividades humanas era inexistente, sendo que o autor em [4] sugere que a causa intrínseca estava relacionada com o facto da aquisição de imagens de profundidade em tempo real ser uma tarefa difícil.

Avanços recentes em tecnologias para captação de dados visuais 3D permitem capturar informação de profundidade em tempo real a um custo significativamente inferior, o que potenciou a investigação em torno do reconhecimento de atividades humanas e a sua adoção em contextos reais.

Os sensores amplamente utilizados para captação de imagens de profundidade incluem câmeras *time-of-flight* (*ToF*) e câmeras de luz estruturada. A título exemplificativo, na Figura 4 e Tabela 2 apresentam-se os sensores *Kinect* desenvolvidos pela *Microsoft*, onde a primeira versão, *Kinect v1*, incorpora um sensor de luz estruturada e a segunda, *Kinect v2*, adota o método *ToF*.

As câmeras *ToF* calculam a profundidade medindo o tempo de voo de um sinal de luz emitido para cada ponto da imagem. Por outro lado, as câmeras de luz estruturada determinam a profundidade projetando uma luz estruturada num cenário e comparando o padrão refletido com o padrão original guardado.

Em geral, numa imagem de profundidade o valor de cada pixel corresponde à distância entre o sensor e um ponto no mundo real.

As imagens obtidas são também denominadas imagens 2.5D, simplesmente porque a estrutura 3D de pontos visíveis ao sensor é a contida na imagem, logo nada é conhecido acerca do que está atrás do sujeito, objeto ou cena.

No entanto, em [9] afirma-se estar provado que a extração de informação de profundidade oferece um complemento importante às imagens de cor (*RGB*), visto aumentar significativamente a precisão dos modelos para classificação de imagens/vídeo, e que esta forma emergente de representação de dados visuais ajuda a resolver problemas fundamentais na área da visão por computadores.

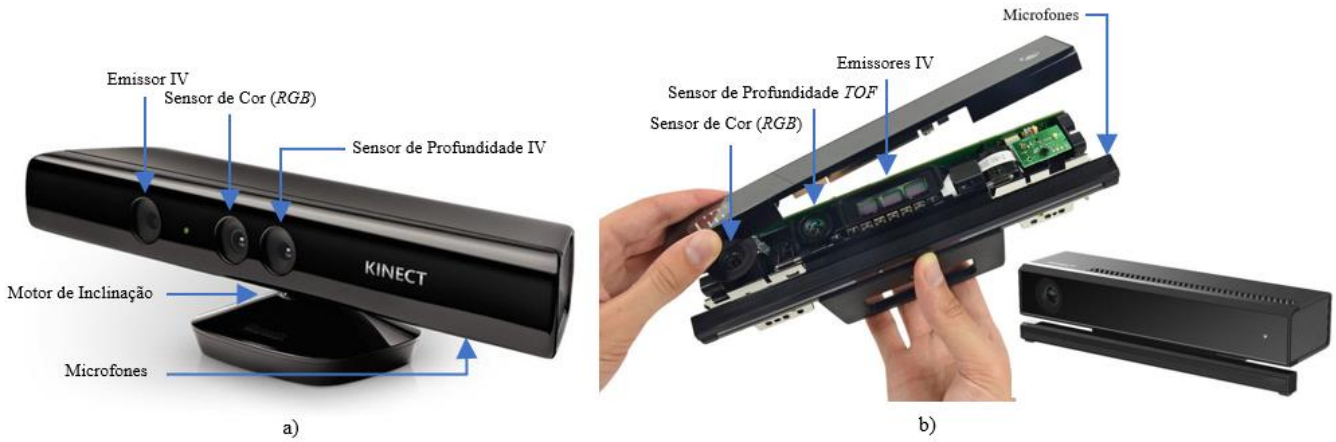


Figura 4 - Sensor Microsoft Kinect for Windows v1 (a) e v2 (b).

Componente	***Kinect v1***	***Kinect v2***
Técnica	Luz estruturada	*Time-of-flight*
Câmera de Cor	640 x 480 @ 30 *fps*	1920 x 1080 @ 30 *fps*
Câmera de Profundidade	320 x 240	512 x 424
Alcance Profundidade	0.8 - 4.0 M	0.5~4.5 M
Campo de Visão Horizontal	57º	70º
Campo de Visão Vertical	43º	60º
Motor de Inclinação	✓	×
Articulações do Esqueleto	20 articulações	26 articulações
Máx. Esqueletos Detetados	2	6
Preço	~80€	~407€

Tabela 2 - Especificações técnicas dos sensores Kinect v1 e v2 [10].

2.1.2 Análise comparativa de sensores

Nos estudos [11] e [12] foi demonstrada superioridade da *Kinect v1* comparativamente às câmeras *SLR Stereo* e sensores *TOF* (*SR-4000* e *Fotonic B70*), em termos de precisão, resolução, propriedades de erro e capacidade de medir 3D.

Os resultados experimentais determinam que a performance do sensor *Kinect v1* é muito próxima do sensor laser usado como referência (*ground truth*) para ambientes de alcance curto (até

3.5 metros). Conclui ainda que nenhuma das duas modalidades de sensores em teste alcançam as métricas de desempenho do sensor a laser para ambientes com distâncias de médio e longo alcance.

Os resultados dos estudos referidos sugerem implicitamente que a *Kinect v1* deverá ser uma melhor escolha comparativamente às câmeras *ToF (SR-4000* e *Fotonic B70)* e *SLR Stereo* se a intenção da sua aplicabilidade for em ambientes de curto alcance.

Refira-se também que as câmeras *ToF* são tipicamente mais caras do que o sensor *Kinect v1*, no entanto, oferecem velocidades maiores e mapas de profundidade mais densos que cobrem todos os pixéis [13].

A vantagem inerente ao dispositivo de luz estruturada *Kinect v1* reside no facto de ser mais acessível e permitir o seguimento de pessoas (*tracking*), bem como reconhecer as suas poses e gestos. Apesar da forte versatilidade e amplo domínio aplicacional que os sensores de luz estruturada possibilitam, o sinal de profundidade é afetado por uma quantidade significativa de ruído que degrada a qualidade da reconstrução 3D e o desempenho de algoritmos que processam informação de profundidade [14].

Partindo das premissas anteriores, a segunda versão do sensor, *Kinect v2*, passa a adotar a tecnologia *ToF*, reduzindo a quantidade de ruído e melhorando a precisão das medições. Apresenta ainda alguns elementos inovadores que ultrapassam limitações conhecidas dos sensores *ToF*, nomeadamente, problemas derivados da distorção harmónica e limitações na distância máxima mensurável, através, respetivamente, da modulação da luz emitida com uma onda quadrada e não sinusoidal e do recurso a múltiplas frequências de modulação [14].

Em [14] avalia-se a precisão da informação de profundidade dos dois sensores *Kinect* (*v1* e *v2*) na reconstrução 3D e seguimento/deteção de pessoas. Para tal, calcula-se a distância média e desvio padrão das nuvens de pontos captadas pelos sensores *Kinect* relativamente aos valores de referência (*ground truth*) extraídos com o *scanner* a laser *NextEngine 2020i*. Na Figura 5 o código de cores permite identificar as partes onde esta distância é alta (vermelho) ou baixa (azul).

Os resultados dos vários testes elaborados sumarizam-se nas seguintes conclusões:

- Ambos os sensores demonstram invariância a alterações na iluminação verificando-se, no entanto, uma maior robustez por parte *Kinect v2* em cenários de luz artificial e solar.

- A *Kinect v2* prova ser duas vezes mais precisa em ambientes de curto alcance e dez vezes mais precisa a partir dos 6 metros de distância.
- Os dois sensores apresentam um desvio padrão superior no contorno dos objetos, onde é mais difícil estimar corretamente a profundidade.
- A precisão da *Kinect v2* é superior na qualidade da reconstrução 3D e ambos os sensores apresentam artefactos concentrados nos contornos dos objetos (Figura 6).
- Na deteção de pessoas a distâncias superiores a 3.5m, os resultados qualitativos demonstram a superioridade da segunda geração da *Kinect* (Figura 7) e os resultados quantitativos destacam ganhos em 20% na precisão para deteção de pessoas.

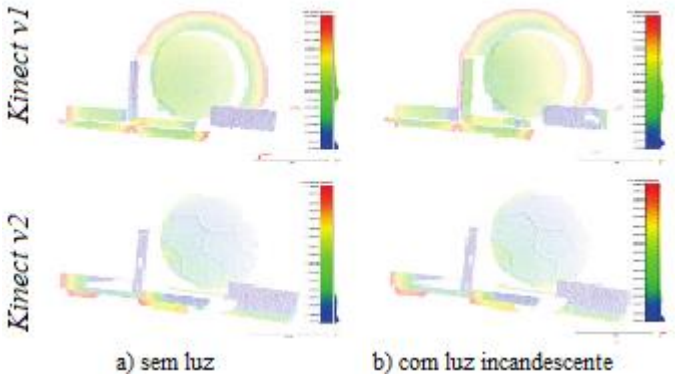


Figura 5 - Comparação com os valores de referência em ambientes sem luz (a) e com luz incandescente (b) entre as gerações v1 e v2 da Kinect [14].

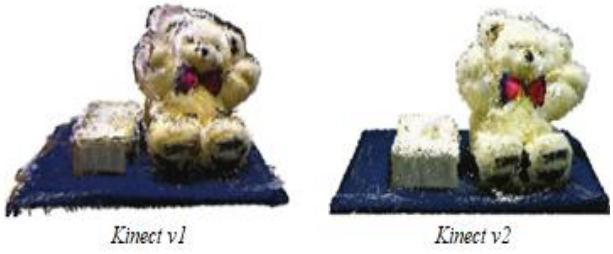


Figura 6 – Comparação da reconstrução 3D de urso de peluche das duas gerações da Kinect [14].

Figura 7 - Resultados qualitativos do seguimento de pessoas das duas gerações da Kinect [14].

Importa reforçar a ideia de que ambas as modalidades de sensores, *ToF* e luz estruturada, são sistemas de visão ativa e como tal a qualidade dos seus resultados é limitada pela distância a que a luz consegue ser projetada e refletida de volta [13].

Relativamente aos sistemas de captura de movimentos (*MOCAP*), embora apresentem vantagens no que toca à precisão da informação extraídas [15], a ponderação da sua utilização no presente projeto de dissertação foi descartada pelas limitações já referidas, das quais se destaca o

seu custo de aquisição, complexidade de configuração e o facto de ser intrusiva em abordagens com marcadores.

2.1.3 Adoção do sensor *Microsoft Kinect v2*

A vertente experimental do projeto de dissertação será feita com recurso ao sensor *Microsoft Kinect v2* pelas seguintes razões:

- Maior conveniência associada à aquisição de dados 3D quando comparada com as modalidades de sensores alternativas já discutidas devido, sobretudo, às ferramentas disponíveis que permitem a extração do esqueleto humano, i.e., o conjunto de articulações e respetivas ligações, que serão detalhadas mais à frente.
- Existe um elevado número de algoritmos propostos para o reconhecimento de atividades humanas via *Kinect* nas mais variadas áreas aplicacionais, constituindo-se um bom ponto de partida para o desenvolvimento do modelo proposto.
- Desafios de baixo nível como sombras e variações da luz são aliviados recorrendo ao canal de profundidade [4].
- A *Microsoft Kinect v2* demonstra ser um dispositivo de elevada qualidade e precisão quando comparado com tipos de sensores concorrentes.

No entanto, problemáticas como o limite do seu alcance ou a robustez perante oclusão de partes do corpo constituem-se como os maiores desafios na atual investigação com base neste tipo de sensores.

Tal como referido, as atuais técnicas de captura de profundidade apenas oferecem uma estimação 2.5D do cenário [4], emergindo a problemática da oclusão. Como tal, a construção do modelo proposto na presente dissertação irá procurar selecionar a abordagem mais robusta a esta problemática potenciando o seu funcionamento em cenários reais.

Verifica-se ainda que a generalidade das abordagens atuais lida com apenas um sujeito, em parte devido à limitação do alcance do sensor. Apesar disso, prevê-se que os sensores de alcance sejam alvo de desenvolvimentos futuros que venham aumentar o seu intervalo de alcance e resolução assim como reduzir o ruído do seu *output* [4].

Em geral, o rápido crescimento e melhoria contínua das técnicas para captar informação 3D, assim como a crescente disponibilidade e acessibilidade dos sensores de profundidade, cria um futuro promissor na investigação de visão por computadores [4].

Por fim, devido ao facto deste sensor estar disponível nos laboratórios da universidade, a sua adoção na presente dissertação torna-se uma opção viável.

2.2 Modelos para reconhecimento de ações humanas

Nesta secção, a revisão literária está estruturada em subsecções de acordo com o processo associado à construção dos modelos para reconhecimento de ações humanas.

Em cada subsecção, destacam-se as abordagens ou tecnologias consideradas mais relevantes, nomeadamente por serem amplamente adotadas no contexto do reconhecimento de ações humanas, e cuja aplicação na vertente experimental do projeto de dissertação será ponderada.

Importa referir que o termo inglês *feature(s)*, mencionado múltiplas vezes ao longo da dissertação, é a expressão adotada na área da aprendizagem automática para referir uma propriedade ou característica individual mensurável de um fenómeno observado. No contexto da presente dissertação irá remeter para os atributos das poses ou dos movimentos (como a localização, orientação ou velocidade das articulações), passíveis de captação através de sensores, que serão utilizados para modelar as ações humanas. Foi dada preferência ao uso do termo em inglês por ser mais representativo do conceito, quando comparado com uma eventual tradução, assim como o facto de ser geralmente aceite e compreendido.

Para reconhecer atividades humanas usando computadores, [16] afirmam, com base em resultados experimentais, que é suficiente conhecer a posição, orientação e movimento das articulações.

Assim, na primeira subsecção abordam-se as duas principais bibliotecas utilizadas em sensores de profundidade, *Microsoft Kinect SDK* e *OpenNI*, para a extração de informação de cor (*RGB*) e profundidade, assim como o conjunto estruturado de articulações do esqueleto humano e respetivas ligações.

Para reconhecer um gesto ou atividade humana a partir dos dados extraídos, o primeiro passo consiste em determinar o conjunto de *features*, nomeadamente das articulações, mais relevantes que permitam compreender a semântica do movimento [17]. Como tal, a segunda subsecção foca-

se nos modelos geralmente adotados para selecionar o conjunto de características mais descritivas dos movimentos, modelos de *filtro* e *wrapper*.

Em [17], salienta-se que os métodos de aprendizagem automática são a forma mais popular para desempenhar o reconhecimento de movimentos humanos, pelo que a subsecção 2.2.3 descreve as características gerais dos principais: Máquina de Suporte Vetorial, Árvores/florestas de Decisão, Modelos de *Markov* não Observáveis, Redes Neuronais Artificiais e *Adaboost*.

De forma a retratar a generalidade dos modelos para o reconhecimento de ações humanas com base em imagens de profundidade existentes na literatura e de contexto de aplicação indiferenciado, na subsecção 2.2.4 agregam-se os mais relevantes consoante o tipo de características movimento extraídas.

Por fim, na subsecção 2.2.5 reúnem-se as abordagens existentes na literatura para o reconhecimento de ações humanas cuja aplicação é exclusiva ao contexto de loja, retratando o seu estado-de-arte atual e destacando algumas das particularidades ou adaptações necessárias associadas a este cenário.

2.2.1 Ferramentas para extração de *features*

Em [18] refere-se que uma ação pode ser vista como uma sequência de poses distribuídas no tempo. Acrescenta também que a pose respeita certas orientações e posições relativas das articulações do esqueleto extraído. Seguidamente, com base nestas mesmas posições e orientações das articulações extraem-se diversas *features* que serão utilizadas para modelar movimentos desempenhados por pessoas.

Ao longo da revisão bibliográfica elaborada verifica-se uma elevada variedade de características passíveis de extrair e utilizar nos mais variados problemas de classificação de movimentos. A título elucidativo, podem-se extrair características como a localização, ângulo, velocidade, orientação ou distância da articulação. Atualmente, existem várias ferramentas complementares ao sensor *Microsoft Kinect* que permitem extrair o conjunto estruturado de articulações e respetivas ligações a partir dos mapas de profundidade produzidos pelo sensor [16].

Destaca-se a ferramenta oficial da *Microsoft* denominada *Microsoft Kinect SDK* e a *OpenNI* (*Open Natural Interaction*) que é de código aberto e mantida pela comunidade de utilizadores do sensor *Microsoft Kinect* [19].

As características principais destas duas bibliotecas são resumidas na Tabela 3.

	OpenNI	**Microsoft SDK**
Calibração de Câmera	√	√
Calibração automática do corpo	×	√
Esqueleto em pé	√	√
Esqueleto sentado	×	√
Reconhecimento de gestos do corpo	√	√
Análise de gestos da mão	√	√
Deteção Facial	√	√
Analisador de Cenário	√	√
Scanning 3D	√	√
Controlo Motor	√	√

Tabela 3- Comparação das bibliotecas OpenNI e Microsoft SDK [19].

As principais diferenças entre as duas ferramentas são mencionadas em [20], destacando-se, o facto da *Microsoft Kinect SDK* extrair 25 articulações de um esqueleto que esteja em pé e 19 de um esqueleto em que a parte inferior do corpo não seja visível, enquanto que o *OpenNI* extrai 15 articulações de um esqueleto em pé. Adicionalmente, no *OpenNI*, o detetor de esqueleto requer que o utilizador mantenha uma posição de calibração até que o detetor identifique articulações suficientes, sendo este tempo variável e altamente dependente das características do cenário e do poder de processamento. Contrariamente, o *Microsoft SDK* não requer uma pose de inicialização específica sendo, no entanto, mais propensa a falsos positivos, especialmente quando a pose humana inicial é demasiado complexa [20].

A *OpenNI* foca-se também na deteção de mãos e respetivo esqueleto, enquanto que a *Microsoft Kinect SDK* apenas realiza o reconhecimento de gestos simples como "agarrar" ou "empurrar" [20].

Uma abordagem alternativa presente noutros trabalhos de investigação prende-se com a extração de *features* diretamente das imagens de profundidade e/ou a cores (*RGB*). Tal abordagem pode ser útil quando as imagens *Kinect* do esqueleto não estão disponíveis ou não apresentam precisão suficiente [17]. Como a *Kinect* apenas consegue extrair o esqueleto humano quando este está dentro de uma gama de distâncias específica, esta abordagem poderá ser uma alternativa relevante a ter em conta nessas condições.

Os métodos utilizados para extrair *features* de tramas de cor/profundidade tipicamente correspondem a adaptações daqueles que são utilizados tradicionalmente em tramas de vídeo a cores. Verifica-se também que a informação adicional de profundidade permite facilitar a extração de caraterísticas mais distintivas para reconhecer atividades, quando comparada com os métodos meramente baseados em análise de tramas de vídeo a cores [17].

Para extrair o conjunto de *features* mais distintivas e posteriormente conceber um modelo capaz de detetar e classificar diferentes ações é fundamental ter em conta determinadas barreiras que derivam da natureza dos ambientes não estruturados.

Cada pessoa tem os seus próprios hábitos e maneirismos ao desempenhar ações, pelo que, consequentemente, estas irão ter variações na velocidade e estilo e assim criar dificuldades adicionais para os modelos de reconhecimento de ações [21].

A similaridade de movimentos em ações distintas é também problemática visto que estas apenas são distinguidas através de detalhes espácio-temporais consideravelmente subtis [4].

Outro desafio prende-se com variação nos ângulos de visão obtidos em relação ao sujeito em questão. Com isto, a mesma ação pode gerar diferentes aparências consoante a perspetiva captada por um dado ângulo de visão. No entanto, conseguindo inferir com precisão o esqueleto de articulações, é possível construir algoritmos invariantes à perspetiva de visão [4] através de operações de rotação.

Adicionalmente, sujeitos posicionados a distâncias distintas da câmera ou com diferentes tamanhos de corpo geram variações na escala, constituindo-se outro desafio. Recorrendo à informação da profundidade esta problemática poderá ser solucionada visto que as dimensões reais 3D do individuo são conhecidas [4].

2.2.2 Seleção de *features*

Partindo dos vetores de *features* extraídos, é fundamental determinar quais as articulações acrescentam maior valor ao desempenho do classificador e quais devem ser descartadas porque não são relevantes para a aplicação específica, dado introduzirem confusão, ruído e reduzirem a taxa de reconhecimento [16]. Um exemplo de articulações não relevantes são aquelas que não têm um movimento independente, como por exemplo as articulações da anca em relação às dos ombros.

Para podermos avaliar uma seleção especifica de f*eatures* é necessário aplicá-la num algoritmo de reconhecimento de ações humanas e comparar as métricas de desempenho obtidas.

Assim, o processo de seleção de *features* visa determinar o subconjunto de variáveis que maximizam o desempenho da classificação. Adicionalmente, reduzir a dimensão do conjunto de *features* também permite reduzir os custos computacionais [16]. A abordagem mais básica a este processo é considerar um vetor binário onde cada bit representa a consideração ou não de determinada *feature*. Para implementar isto existem dois modelos principais: o modelo de filtro e o modelo *wrapper*.

O modelo de filtro realiza uma análise a priori dos dados de forma a determinar a relevância das *features* com base nas suas propriedades intrínsecas. Estes métodos demonstram ser computacionalmente leves e evitam o sobreajuste (do inglês: *overfitting*). No entanto, ignoram o algoritmo de aprendizagem que está por baixo e o subconjunto de *features* resultante poderá conter redundância, sendo que não será o ótimo [16].

Para determinar a relevância das *features,* [22] sugere determinar a correlação existente entre as *features* (variáveis independentes) e a classificação (variável dependente) recorrendo ao Coeficiente de Correlação de *Pearson*, que determina a sua dependência linear. Adicionalmente, calcular a informação mútua entre *features* e a classificação é uma abordagem recorrente visto averiguar se uma variável independente pode dar informação sobre a variável resultado, indiciando a existência de dependência entre elas. Contudo, os métodos referidos não discriminam as *features* relativamente à correlação entre si o que pode resultar num subconjunto com redundância. Acresce ainda a possibilidade de serem descartadas *features* que por si só são pouco informativas para efeitos preditivos, mas quando combinadas com outras podem ser relevantes.

Por oposição, no modelo *wrapper* a seleção de *features* tem em conta o algoritmo de aprendizagem para desempenhar avaliações de vários subconjuntos de *features* na busca daquele que conduz ao melhor desempenho. Como tal, para cada subconjunto de *features* é criado um novo modelo com os dados de treino e aferido o seu desempenho (p. ex., precisão) com os dados de teste. Consequentemente, a principal desvantagem desta abordagem face à anterior prende-se com o considerável tempo consumido no processo de avaliação de subconjuntos de *features.*

Em [22] classifica-se os métodos de *wrapping* em duas classes gerais: Algoritmos de Seleção Sequencial e de Procura Heurística. Os primeiros caracterizam-se por começar com um conjunto

de *features* vazio (cheio) e adicionam (removem) *features* até que o máximo da função objetivo seja atingido (i.e., maximizar desempenho na classificação). Por outro lado, os Algoritmos de Procura Heurística avaliam diferentes subconjuntos para otimizar a função objetivo e diferencia-se da anterior por procurar uma solução suficientemente boa num período de tempo menor, mas que poderá não ser aquela que conduz ao melhor desempenho global. Como tal, a sua aplicabilidade predomina nos problemas de classificação em que a busca da melhor solução é impossível ou impraticável.

Apesar de tudo, abordagens baseadas no modelo *wrapper* são tipicamente preferíveis por demonstrarem melhores resultados [16].

2.2.3 Algoritmos para reconhecimento de ações humanas

Aqui são maioritariamente utilizados modelos estatísticos sofisticados como a máquina de suporte vetorial, árvores/florestas de decisão, Modelos de *Markov* Não Observáveis, redes neuronais artificiais e *AdaBoost*, respetivamente, do inglês: *SVM - Support Vector Machine; Decision Trees/Forests, Hidden Markov Model, ANNs – Artificial Neural Networks* - e *AdaBoost – Adaptive Boosting*.

Em seguida, ir-se-á fazer uma introdução sucinta aos vários modelos considerados mais populares segundo [17]:

O algoritmo *SVM* é um método popular de aprendizagem supervisionada para classificação com base em *codebooks*. Na sua essência, utiliza os dados de treino para determinar o plano ou hiperplano que maximiza a distância entre os elementos mais próximos e a fronteira de decisão. O referido plano ou hiperplano será então utilizado para classificar dados desconhecidos, como movimentos, em classes diferentes.

Árvores ou florestas de decisão adequam-se ao reconhecimento de gestos/movimentos se estes puderem ser modelados numa sequência de poses chave. O caminho entre o nó folha até ao nó raiz será assim usado para classificar gestos/movimentos desconhecidos com árvores/florestas treinadas onde cada nó interno representa uma pose chave.

Em alternativa, os modelos de *Markov* Não Observáveis são vulgarmente adotados. Estes capturam características temporais dos movimentos relacionando sequências de poses através de

probabilidades de transição de estado (pose). Assemelha-se às árvores/florestas de decisão no sentindo em que procura modelar gestos/movimentos numa sequência de poses chave.

Para o reconhecimento de atividades mais sofisticadas, o Modelos de *Markov* Não Observáveis Hierárquicos são adotados para modelar atividades em subatividades. Por exemplo, a atividade lavar os dentes compreende as subatividades: apertar a embalagem de pasta de dentes; aproximar a escova de dentes à cara; escovar os dentes e assim por diante. A natureza hierárquica das atividades poderá ser capturada usando um modelo gráfico hierárquico probabilístico. No entanto, este modelo apresenta complexidades acrescidas quando os movimentos são executados a velocidades distintas ou com variações no estilo dos movimentos. Para solucionar esta problemática, [17] afirma que inferindo uma estrutura gráfica das duas camadas, i.e. atividades e subactividades, e usando programação dinâmica é possível tratar automaticamente da adaptação a estas variações.

Relativamente às redes neuronais artificiais, estas consistem na criação de modelos compostos por muitos elementos computacionais não-lineares que operam em paralelo e são organizados em padrões reminiscentes às redes neuronais biológicas [23].

Rede Neuronal Convolucional (do inglês: *Convolutional Neural Networks*) é uma categoria de redes neuronais que prova ser eficaz na área do reconhecimento de imagens para classificação. Este tipo de modelos pode usar diretamente dados de entrada no seu estado bruto, no entanto a generalidade destes suporta apenas dados 2D [24]. Em [24] propõe-se um modelo 3D para reconhecimento de ações. Para tal, extrai *features* das dimensões espaciais e temporais desempenhando convoluções 3D e assim captura informação de movimento codificada em múltiplas tramas adjacentes.

AdaBoost, abreviatura para *Adaptive Boosting*, combina a soma ponderada de um conjunto de classificadores fracos [25]. Ao contrário de redes neuronais ou *SVMs*, o processo de treino do *AdaBoost* seleciona apenas as *features* que melhoram o poder preditivo do modelo, reduzindo a sua dimensão e potencialmente melhorando o tempo para execução visto que *features* irrelevantes não são computadas.

2.2.4 Reconhecimento de ações por categoria de *features* extraídas

No últimos anos muitos métodos foram desenvolvidos em torno do reconhecimento de atividades humanas a partir de imagens de profundidade, pelo que [4] propõe dividi-los em cinco categorias de acordo com as *features* que estes usam:

I - Silhuetas 3D

Inicialmente esta técnica consistia na extração de silhuetas das pessoas com base em imagens de intensidade ou a cores *RGB* e como tal fazia-se acompanhar de dificuldades no reconhecimento em ambientes com variações na luz ou planos de fundo com texturas confusas. Em imagens de profundidade, a silhueta da pessoa pode ser extraída com maior facilidade, precisão e riqueza informativa. Entre os vários algoritmos propostos baseados em silhuetas 3D, destacam-se dois dos mais sonantes no meio científico:

- Em [26] aplica-se um grafo de ação para modelar explicitamente as dinâmicas das ações e um "saco de pontos 3D" para caracterizar um conjunto de posturas salientes que correspondem aos nós no grafo de ações. Demonstra-se também o potencial do modelo de posturas com "saco de pontos 3D" para lidar com oclusões através de simulações.
- Na abordagem apresentada em [27], projetam-se os mapas de profundidade em três planos ortogonais que acumulam a energia do movimento para criar os Mapas de Profundidade de Movimento (conhecidos por *Depth Motion Maps (DMM)*). As formas criadas pelo acumular de energia do movimento criam aparências e formas especificas que podem ser usadas para caracterizar as categorias de ações. Os referidos *DMMs* são descritos através de Histogramas de Gradientes Orientados. Resultados obtidos com um conjunto de dados para avaliação de acesso público (*MSRAction3D*, disponível em [28]) indicam que esta abordagem superou o estado-de-arte dos métodos desenvolvidos até essa altura.

Em [4] destaca-se que os algoritmos baseados em silhuetas 3D propostos até então são mais apropriados para reconhecimento de ações de uma única pessoa e apresentam um melhor desempenho para ações simples e atómicas. Acrescenta que existem dificuldades no reconhecimento de atividades complexas devido a perdas de informação associadas ao cálculo de projeções 2D a partir de dados 3D ou ao acumular informação ao longo do domínio temporal. Adicionalmente, salienta que a qualidade das silhuetas é drasticamente afetada por oclusões e ruído e que a precisão da extração é dificultada quando o sujeito interage com objetos.

II - Articulações do esqueleto e deteção de partes do corpo

Esta vertente recorre frequentemente ao algoritmo de deteção de articulações do esqueleto que está incorporado no sensor *Kinect*. Consequente do fácil acesso à localização das articulações através desta tecnologia, muitos algoritmos foram propostos para reconhecer atividades através desta informação.

Uma das *features* mais intuitivas e diretas consiste no emparelhamento das diferenças de localização das articulações, que é uma representação compacta da estrutura da postura do esqueleto numa única trama. Calculando a diferença das posições das articulações da trama atual com a anterior é possível obter o movimento entre duas tramas.

Em [29] desenvolve-se um método denominado *EigenJoints* que, através da concatenação das *features* referidas anteriormente e recorrendo ao algoritmo de classificação *Naive* de Bayes, alega superar significativamente o estado de arte para o conjunto de dados *MSRAction3D*.

Além da concatenação direta apresentada anteriormente, em [30] as posições das articulações são transpostas para cones 3D e construídos histogramas destes pontos (conhecidos pela sigla *HOJ3D*) enquanto representativos das posturas. Estas posturas são projetadas usando *LDA* (*Latent Dirichlet Allocation*) e agrupadas segundo *k* posturas de palavras visuais representativas das poses típicas das ações. Por fim, modela a evolução temporal destas palavras visuais através dos modelos de *Markov* não Observáveis. Os autores acrescentam que, fruto do *design* do seu sistema esférico de coordenadas e da robustez da estimação do esqueleto via *Kinect*, o método demonstra uma significativa invariância ao ângulo de visão e um desempenho de classificação superior ao de [26] para o mesmo conjunto de dados (*MSRAction3D*).

A partir da localização das articulações do esqueleto pode-se calcular a sua orientação, o que constitui uma boa *feature* visto ser invariante à dimensão do corpo [4]. No mesmo estudo, refere-se ainda que outros investigadores agrupam articulações e constroem planos a partir destas e medem a distância e movimento entre o plano e a articulação enquanto *features*.

Comparativamente às *features* de silhuetas 3D, as *features* de articulações do esqueleto apresentam a vantagem de serem invariantes à localização do sensor (i.e., consegue extrair as articulações do esqueleto em vistas frontais, laterais e traseiras) e ao tamanho do sujeito captado. Posto isto, [4] acrescenta a possibilidade de modelar interações entre pessoas.

A limitação intrínseca às *features* baseadas no esqueleto humano corresponde ao facto de estas não nos darem informação relativa aos objetos envolventes. Para modelar interações pessoa-objeto terá que se combinar estes algoritmos com outros para deteção de objetos pelo que será esse o foco da revisão literária no subcapítulo 2.4.

III - *Features* de localização espácio-temporal

Em linhas gerais, estes algoritmos procuram detetar pontos de interesse espácio-temporais (do inglês: *spatio-temporal interest points (STIPs))* e em seguida construir descritores em torno destes pontos, os quais serão utilizados para classificação.

Encorajados pelo sucesso em vídeos a cor, são também explorados utilizando informação de profundidade. A generalidade dos algoritmos propostos trata os canais de cores *RGB* e profundidade de forma independentemente aquando da extração dos referidos pontos de interesse. Em [4] sugere-se que esta abordagem não será ótima visto que a profundidade e as cores residem juntas no mundo 3D.

Em contraste, [31] propõe um modelo 4D de *features* locais espácio-temporais que combina profundidade e cores. As *features* são detetadas aplicando diversos filtros ao longo das dimensões espaciais (3D) e temporal (1D). Seguidamente, o descritor de *features* calcula e concatena os gradientes da profundidade e cores dentro de um cubo 4D centrado na *feature* detetada. As *features* resultantes são usadas para reconhecimento de atividades recorrendo ao classificador *LDA* (*Latent Dirichlet Allocation*). O autor conclui que os resultados experimentais obtidos com as *features* mencionadas superam aquelas que derivam apenas de cores ou profundidade.

Por fim, [4] destaca que as principais limitações desta abordagem algorítmica são a dependência dos ângulos de visão, o facto de ser necessário o vídeo completo como entrada e a complexidade do algoritmo de computação das *features* que condiciona o seu uso em aplicações em tempo real.

IV - *Features* de ocupação local 3D

As *features* obtidas via padrões de ocupação local podem ser definidas no espaço *(x, y, z)*, descrevendo a aparência da profundidade local num determinado instante temporal, ou no espaço *(x, y, z, t)* descrevendo eventos atómicos dentro de determinado intervalo de tempo.

Nesta categoria, destaca-se [32] onde é proposta a *feature 3D Local Occupancy Patterns* (*LOP*) para descrever a aparência da profundidade local e assim caracterizar interações pessoa-objeto

para cada articulação. Para cada trama, esta *feature* computa a informação de ocupação local com base numa nuvem de pontos 3D em torno de uma articulação em particular. Com isto, captam-se as dinâmicas temporais, via Pirâmide Temporal de *Fourier*, de todos os padrões de ocupação, possibilitando a discriminação de diferentes tipos de interações.

O uso desta *feature* alia-se à representação do movimento humano enquanto um emparelhamento da posição relativa das articulações resultando em *features* mais discriminativas. Em suma, [32] conclui que o modelo proposto é discriminativo o suficiente para classificar ações humanas com diferenças subtis e interações pessoa-objeto com robustez a ruído e desalinhamentos temporais.

Salienta-se também que *features* espácio-temporais contêm informação do plano de fundo ao passo que as de ocupação local contemplam informação em torno de pontos específicos no espaço. Esta característica pode ser positiva ou negativa visto que o plano de fundo é útil em determinados cenários ao passo que noutros é perturbador [4].

V - *Features* extraídas do fluxo ótico 3D

O fluxo ótico corresponde à distribuição de velocidades aparentes do movimento de padrões de brilho numa imagem que surgem do movimento relativo do observador e dos objetos [4]. Amplamente usado em imagens a cores para estimação de movimentos ou segmentação de objetos, constitui-se também como uma fonte de *features* popular no reconhecimento de atividades em vídeos. As câmeras de profundidade oferecem informação geométrica útil que pode ajudar a suavizar as imagens bem como transformar de fluxos óticos 2D em 3D, e assim possibilitar um melhor reconhecimento de movimentos.

Atualmente, o fluxo 3D é tipicamente calculado para todos os pontos 3D do sujeito ou cena, resultando num custo computacional elevado e dificultando o seu uso em aplicações em tempo real [4].

No entanto, em [33] sugere-se o cálculo do fluxo ótico 3D apenas para algumas porções relevantes do cenário 3D. Cada sujeito detetado é representado por um cluster definido por uma nuvem de pontos 4D sendo estas dimensões representativas das coordenadas geométricas 3D e componente de cor RGB para cada ponto. Com esta informação é estimado o fluxo ótico 3D associado a cada *cluster* por cada imagem. O fluxo ótico é então codificado num vetor único

recorrendo ao descritor 3D proposto baseado em grelha e assim obtêm uma representação apropriada para ações humanas.

Em geral, [4] afirma que a exploração do fluxo ótico 3D usando profundidade e cores *RGB* é ainda limitada, encontrando-se numa fase preliminar. Sugere também que, emergindo novas técnicas mais eficazes para computar fluxos 3D, este tipo de *features* têm potencial para se tornar populares no reconhecimento de ações humanas.

2.2.5 Reconhecimento de ações em contexto de loja

Esta secção reúne as abordagens existentes na literatura em que a aplicabilidade do reconhecimento de ações é exclusiva ao contexto de loja.

Com vista ao desenvolvimento do modelo proposto, esta secção foca-se nas tecnologias utilizadas e modelos desenvolvidos, assim como na compreensão das especificidades deste ambiente e consequentes adaptações necessárias à correta modelação das atividades dos clientes. Adicionalmente, procura descrever a forma como o reconhecimento de ações humanas é um complemento fundamental para sistemas de lojas inteligentes.

Em [34] introduz-se um fluxo de trabalho para a constituição de um sistema para ambientes de loja inteligentes que procura melhorar a qualidade e eficiência dos estabelecimentos de retalho e aumentar a sua atratividade. Entre as várias camadas do sistema, salienta-se um bloco para análise de vídeo, constituída pelos módulos *tracking*, interação cliente-produto, deteção de faces e demografia, onde os primeiros dois coincidem com o âmbito do presente projeto de dissertação.

O *tracking* remete para a deteção humana e [34] considera-o como um passo preliminar visto que os módulos posteriores dependem da posição do sujeito a ser captado. Destaca que para os algoritmos testados as principais dificuldades estão associadas a alterações na iluminação, oclusões, movimentos da câmera, ruído, baixo contraste e reflexões especulares. Relativamente aos algoritmos para deteção e seguimento visual (*tracking*), num teste onde se experimentou um vasto conjunto de algoritmos (19 algoritmos distintos) conclui-se que o *Structured output Tracking with kernels (STR)* tem o melhor desempenho. Na abordagem proposta para este algoritmo, a sua principal característica corresponde à eliminação do passo intermédio de classificação que converte em tempo real a posição do objeto em exemplos de treino etiquetados. Alternativamente, propõe-se uma deteção visual de objetos adaptativa baseada na previsão estruturada do resultado,

permitindo ao espaço de resultados expressar explicitamente as necessidades do detetor e assim evitar o passo de classificação intermédio.

O algoritmo denominado *Tracking, Learning and Detection (TLD)* é considerado o segundo mais notável devido ao vasto número de vezes em que supera os demais ao desempenhar a deteção da posição humana em cenários caracterizados por momentos de oclusão e movimento da câmera.

Relativamente ao segundo módulo da camada de análise de vídeo, captação de interações cliente-produto, [34] salienta que esta informação é essencial para se determinar quão apelativos são para os clientes o visual dos produtos e a distribuição do espaço em loja. O autor destaca a abordagem proposta em [35] onde se recorre a uma rede distribuída de sensores *RGB-D* para criar zonas de monitorização, onde existe interação positiva quando o cliente agarra num produto, negativa quando o agarra e volta a colocá-lo na prateleira e neutra quando estende a mão sem pegar num produto.

O sistema proposto em [35] foca-se maioritariamente nestas interações e é capaz de detetar a presença de clientes identificando-os inequivocamente, no entanto, o algoritmo de processamento de imagem intrínseco a esta identificação requer otimizações para que o valor identificador atribuído se mantenha ao longo dos vários sensores (i.e. diferentes áreas da loja) e não apenas num determinado espaço de monitorização. Nesta abordagem utilizam-se os sensores de profundidade *Asus Xtion Pro* que são preferidos face ao sensor *Microsoft Kinect* por serem mais pequenos e alimentados apenas via porta USB. A este propósito, considera-se que as referidas vantagens face à *Microsoft Kinect* não desvalorizam a decisão da sua utilização na parte experimental desta dissertação visto que, para efeitos de investigação, as características básicas criticas para a captação do corpo humano são, em geral, as mesmas [36].

Importa destacar que em [35] os sensores *RGB-D* são montados verticalmente para fazer o seguimento das atividades dos clientes dentro da loja e captar as suas interações com prateleiras e não num plano lateral ou frontal como verificado nas abordagens previamente analisadas em 2.2.4.

O algoritmo para deteção de interações cliente-produto implementa um método para subtração do plano de fundo. Para evitar a falsa deteção de objetos (falsos positivos) fruto de planos de fundo ruidosos, a imagem de fundo é atualizada dinamicamente e um valor de *threshold* é definido para discriminar sinais positivos, indicativos de objetos em movimento. Cada *blob* referente a um produto é reconhecido e seguido dentro do fluxo de vídeo.

A implementação deste método permite o seguimento de movimentos de forma eficaz, nomeadamente por se basear em imagens de profundidade e não estar sujeita a alterações rápidas nas formas captadas.

Por fim, refira-se uma função interessante presente no sistema proposto em [35] que permite a construção de um mapa de calor das interações, em tempo real, que dá informações acerca das áreas das prateleiras onde houveram mais interações, discriminando-as de acordo com os três tipos de interações previamente mencionados (positivas, negativas e neutras).

Deduz-se através da revisão literária que vários artigos foram publicados referentes ao mesmo sistema proposto em [35] dados os autores em comum e sucessivas referências ao trabalho anterior, sendo que a análise de [35] será complementada por essa via:

- Em [37] refere-se que as "sugestões do movimento", habitualmente usadas para eliminar ambiguidades na deteção de ações, não são compatíveis com as imagens recolhidas visto se tratar de uma vista de topo (vertical). Refere ainda que para detetar e reconhecer interações cliente-produto é utilizado o método *Template Matching* que procura similaridades nas medições entre as *features* do *template* e as capturadas. Para construir o *template* captaram-se protótipos da forma das mãos (com e sem objetos).
- No artigo [38] aprofunda-se a aplicabilidade do sistema num contexto de loja inteligente, nomeadamente apelando à necessidade de novos serviços em loja que tenham por base estes sistemas. O seguimento de clientes permite analisar padrões de tráfego com vista à otimização da organização do espaço. A deteção de interações clientes-produtos dá-nos informação da atratividade da sua exposição, potenciando a sua otimização, e permite detetar falta de *stock* nas prateleiras em tempo real. Introduz ainda o conceito de expositor inteligente *Pick&Play* onde um ecrã apresenta informações adicionais ao cliente consoante as suas interações com o produto.
- Por fim, a publicação mais recente, [39], entra em detalhe relativamente aos processos de deteção de pessoas e respetiva interação com objetos, nomeadamente disponibilizando o pseudocódigo referente. Destaca-se o processo de monitorização de pessoas que, após subtração do plano de fundo, aplica uma segmentação multinível que, em momentos de oclusão entre pessoas, deteta a cabeça de cada pessoa servindo de elemento discriminativo. Por oposição, refere que usando segmentação de nível único, se duas

pessoas colidirem na mesma área de monitorização, são consideradas uma única pessoa. Os resultados experimentais em [39] demonstram que o algoritmo de deteção de pessoas apresenta um *recall* de 99% enquanto que o de deteção de interações com prateleira apresenta um *recall* de 80.5%.

Em [35] refere-se o trabalho de [40] no qual se propõe uma abordagem para a deteção e seguimento das poses do corpo humano, sendo que, com vista à sua aplicabilidade num contexto de loja para a análise de comportamentos de clientes, a câmera é posicionada no topo das prateleiras.

A revisão literária elaborada referente à presente secção sugere que o posicionamento do sensor de profundidade seja feito num ângulo de visão de topo. De facto, neste contexto, o posicionamento de câmeras nos tetos das lojas é a abordagem mais comum [40]. Embora não tenham sido encontradas as razões inerentes a esta decisão, à exceção da sua apropriação ao contexto (i.e. não intrusivo e posicionamento conveniente), será intuitivo concluir que existirão menos oclusões entre pessoas quando comparado com vistas laterais. Importa recordar que esta é uma das maiores problemáticas no reconhecimento de ações humanas.

No entanto, a maioria das abordagens presentes na literatura desenvolvem modelos adaptados à vista frontal da pessoa porque a sua forma é muito mais discriminativa nesta orientação [40]. Consequentemente, [40] propõe uma abordagem adaptada à vista de topo que seja eficiente na análise de comportamentos de clientes. Para tal, recorrem às "sugestões de profundidade" disponibilizadas pelo sensor *Microsoft Kinect* que oferecem features discriminativas da pose. Esta funcionalidade é incorporada num filtro de partículas para detetar as partes do corpo da pessoa, nomeadamente a deteção 2D da cabeça e ombros e deteção 3D dos braços. Em [40] o autor explica que o modelo é decomposto em modelos 2D e 3D de forma a reduzir a complexidade do filtro e alcançar processamento em tempo real. Acrescenta também que as "sugestões de profundidade" têm um forte impacto na redução da complexidade do modelo 2D e que o modelo 3D referente aos braços está constrangido pela posição dos ombros.

Relativamente à implementação do filtro de partículas em [40], apenas se modela a parte superior do corpo e como tal limitam a captação de imagens para apenas ter em conta os pixéis reconhecidos como elementos do torso, braços e cabeça. Refere também que o filtro de partículas demonstra ser bem-sucedido enquanto técnica de aproximação numérica para estimação de

sequências Bayesianas para modelos não lineares e não gaussianos. Na Figura 8 é possível visualizar os modelos 3D resultantes do modelo proposto por [40].

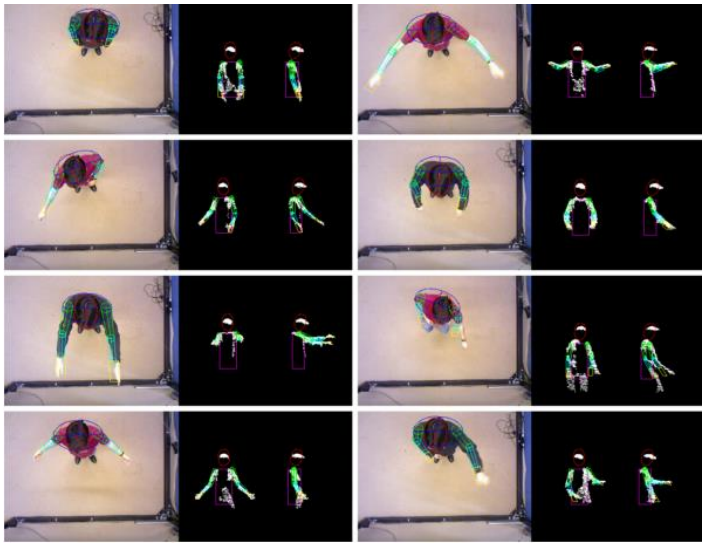


Figura 8- Visualização do estado da pose estimada pelas tramas adquiridas (à esquerda) e no espaço 3D (à direita), com a projeção dos pixéis da imagem de profundidade a branco [40].

2.3 Avaliação de modelos para reconhecimento de ações

2.3.1 Métricas de desempenho

A qualidade de um classificador pode ser avaliada através da matriz confusão [41], constituída pelos elementos :

- Verdadeiros Positivos (VP) – Número de exemplos de uma determinada classe corretamente reconhecidos.
- Verdadeiros Negativos (VN) – Exemplos que foram corretamente reconhecidos como não pertencentes a uma classe.
- Falsos Positivos (FP) – Reconhecimentos incorretos associados a uma classe.
- Falsos Negativos (FN) – Instâncias relevantes da classe que não foram reconhecidas.

A métrica de desempenho dos modelos propostos na literatura com maior frequência de reporte é a percentagem de instâncias corretamente identificadas (exemplos positivos e negativos), denominada *accuracy*[1]. Refira-se que para efeitos de comparação de desempenho entre modelos, estas deverão ser feitas para o mesmo conjunto de dados e será mais credível se mais métricas de

[1] Rácio de ações corretamente reconhecidas face ao total ($\frac{TP+TN}{TP+FP+FN+TN}$).

desempenho estiverem disponíveis visto que métricas distintas representam o desempenho do modelo relativamente a critérios diferentes [42]. Salienta-se, a título de exemplo, o fenómeno *Accuracy Paradox*[2] onde modelos com uma maior taxa de *accuracy* podem ter menor poder preditivo que outros com menor taxa de *accuracy*, apelando assim a uma análise complementar do desempenho de modelos com outras métricas de desempenho, nomeadamente *Precision, Recall* e *F1-Score*. No entanto, a generalidade dos estudos analisados reporta apenas a métrica *accuracy*.

Precision refere-se à percentagem de predições positivas que estão corretas ($\frac{TP}{TP+FP}$) e métrica *recall*, mede quão bem encontramos todos os casos positivos, dado pela fórmula: ($\frac{TP}{TP+FN}$).

A métrica *f1-score* consiste na média harmónica entre *precision* e *recall* ($2 * \frac{Precision * Recall}{Precision + Recal}$), pelo que é a métrica indicada quando se procura um equilíbrio entre ambas as métricas e/ou a proporção de exemplos por classe a classificar não está balanceada. Quando as classes não estão balanceadas deve-se considerar a problemática *Accuracy Paradox*, onde o modelo pode tender para a classificação da classe mais frequente e ainda assim atingir bons resultados em *accuracy*. Como tal, sugere-se o complemento com outras métricas, das quais prevalece o *f1-score*, sobre a *precision* e *recall*, dado considerar-se que o custo de um falso positivo é semelhante ao de um falso negativo, no contexto de aplicação considerado.

A avaliação do desempenho de modelos para o reconhecimento de atividades humanas consiste na comparação do resultado do modelo com o valor verdadeiro de referência (*ground truth*). Esta comparação é complicada visto que o reconhecimento de atividades é temporalmente contínuo (i.e., ambas as etiquetas e fronteiras temporais de cada atividade têm que ser determinadas).

Em [42] destacam-se as principais fontes de erro que dificultam a avaliação de resultados no reconhecimento de atividades humanas e discutem-se os diferentes níveis de análise de resultados de classificação com impacto na facilidade de deteção de determinados tipos de erro. Adicionalmente, destaca que no domínio do reconhecimento contínuo de atividades, os investigadores consideram diferentes conjuntos de resultados da classificação (Tabela 4 e Figura 9) que têm em conta o facto de não existir uma correspondência um-para-um entre os eventos previstos pelo modelo e os valores de referência verdadeiros.

[2] https://en.wikipedia.org/wiki/Accuracy_paradox

Resultados da classificação	Definição
Correct (C)	Representa classificações corretas
Substitutions (S)	Representa corretas deteções temporais, mas uma identificação da atividade incorreta
Insertions (I)	Deteção de uma atividade quando nada realmente aconteceu
Deletions (D)	Falha na deteção de uma atividade
Total Number of True Events (N)	Útil para cálculos estatísticos embora não seja estritamente um resultado da classificação: $N = (C + D + S)$
Underfill (U)	Atividade corretamente identificada, mas o momento temporal de início e fim não é detetado
Overfill (O)	Atividade corretamente detetada, mas temporalmente considera momentos em que observação não faz parte da atividade
Fragmentation (F)	Atividade longa detetada parcialmente múltiplas vezes separada por eventos nulos
Substitution-Fragmentation (SF)	Atividade longa detetada parcialmente múltiplas vezes e separada por atividades conhecidas, mas incorretamente identificadas
Merge (M)	Agrega múltiplos eventos distintos separados por eventos nulos num único evento longo
Substitution-Merge (SM)	Agrega múltiplos eventos distintos separados por eventos conhecidos num único evento longo

Tabela 4 - Resultados de classificação para reconhecimento continuo de ações [42].

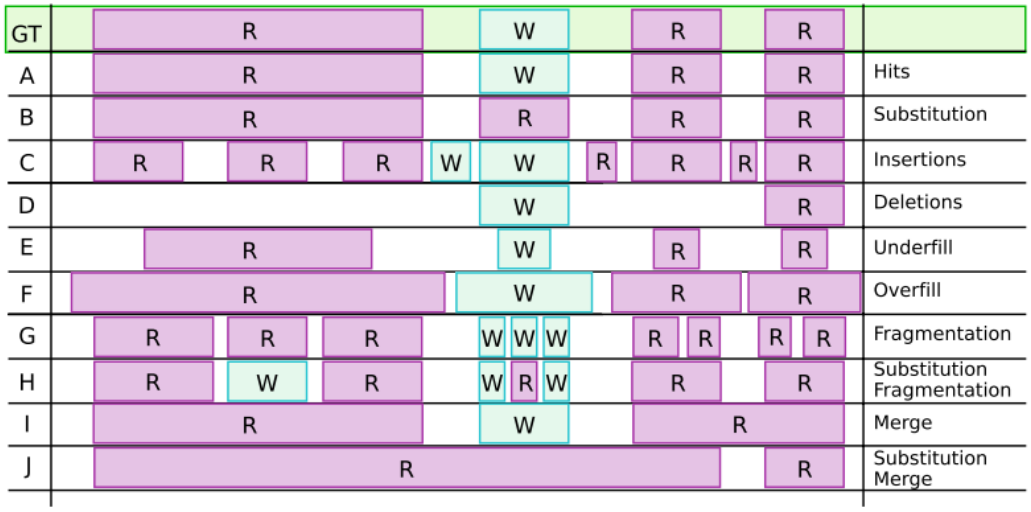


Figura 9 - Ilustração dos diferentes tipos de erro em para as atividades R (correr, do inglês run) e W (andar, do inglês walk). As etiquetas no topo são os valores referência e as letras A-J ilustram os diferentes tipos de erros da classificação [42].

2.3.2 Conjuntos de dados para avaliação de modelos

Como foi referido, o recurso a conjuntos de dados públicos de dados visuais de atividades humanas possibilita a comparação direta das métricas de desempenho do modelo a desenvolver com os existentes na literatura cujo desempenho foi aferida para os mesmos dados. Assim, poderá ser utilizado como forma de validação do modelo, no sentido de permitir concluir se o estado de arte associado ao reconhecimento de atividades humanas foi atingido ou superado.

Desta forma, a presente secção procura recolher os conjuntos de dados potencialmente mais apropriados às eventuais especificidades do modelo a desenvolver.

Destaca-se ao longo da revisão literária que a generalidade dos modelos está adaptada à vista frontal do corpo humano, no entanto, no contexto do retalho e comércio, a vista de topo aparenta ser a escolha mais apropriada embora, como foi referido, seja menos discriminativa para efeitos de reconhecimento da forma humana e respetivas ações.

Consequentemente, existe carência de conjuntos de dados públicos recolhidos na vista de topo o que condiciona a capacidade de validação de modelos adaptados a esta vista. Ao mesmo tempo, sugere existir necessidade de investigação adicional assim como a recolha e disponibilização pública de conjuntos de dados de acordo com esta particularidade.

O modelo proposto em [35] vai ao encontro da especificidades idealizadas do modelo a desenvolver, com destaque para o posicionamento da câmera (Figura 10). Porém, o conjunto de dados utilizado não está disponível para acesso público.

Figura 10- Zona e perspetiva de captação de imagens RGB-D em [35].

Em [4] apresentam-se alguns dos principais conjuntos de dados de acesso público para o reconhecimento de atividades humanas cuja captação de imagens é tipicamente feita na perspetiva

frontal. No entanto, nenhum dos conjuntos de dados acima referidos contém ações semelhantes àquelas que são típicas dos espaços em vendas, nomeadamente interações com prateleiras.

Na Tabela 5 reúnem-se as métricas de performance reportadas para os modelos analisados nos vários artigos científicos considerados na revisão literária. Considera-se que aquele que terá maior significância para efeitos de comparação será o modelo proposto em [39] visto que corresponde ao mesmo contexto aplicacional (i.e., reconhecimento de interações com prateleiras). Como o conjunto de dados não está disponível não será possível fazer uma comparação direta de resultados.

Modelo	**Categoria**	***Accuracy***	***Recall***	***Conjunto de dados***
[26]	Silhuetas 3D	94,2%[6]	-	-
[27]	Silhuetas 3D	97,4%	-	MSRAction3D
[29]	Articulações do esqueleto e partes do corpo	97,8%	-	MSRAction3D
[30]	Articulações do esqueleto e partes do corpo	97,2%	-	MSRAction3D
[31]	Pontos de interesse espácio-temporais	91,5%	-	-
[32]	Ocupações locais 3D	86,8%	-	MSRDailyActivity3D
[33]	Fluxo Ótico 3D	80,0%	-	-
[39]	*Features* espacio-temporais	-	80.52%	-

Tabela 5 - Métricas de desempenho disponíveis por modelo discutido em 2.2.4 e 2.2.5.

2.4 Modelos para reconhecimento de objetos

À semelhança dos algoritmos para deteção de gestos, procura-se uma solução para deteção de objetos de complexidade de implementação adequada ao tempo disponível, com capacidade para deteção em tempo real e compatível com os recursos disponíveis.

Adicionalmente, dado o contexto de aplicação, importa que o sistema seja preciso e capaz de reconhecer um espectro alargado de objetos.

De forma a permitir uma comparação dos vários modelos que constituem o estado-de-arte da deteção e classificação de objetos importa compreender a forma como estes são avaliados.

[6] *Accuracy* média quando 2/3 da amostra é usada para treino (restante 1/3 para teste) sobre o conjunto de ações AS1, AS2 e AS3 especificado em [22 - 25].

2.4.1 Avaliação de modelos para reconhecimento de objetos

A métrica *mAP* (do inglês, *Mean Average Precision*) tornou-se uma forma aceitável para avaliar e comparar sistemas para deteção de objetos em competições como *PASCAL VOC*[7], *ImageNet*[8] e *COCO*[9]. Para compreender o propósito e lógica intrínseca a esta métrica, detalha-se a sua construção em seguida.

De forma a avaliar corretamente um detetor de objetos importa aferir o resultado das seguintes tarefas:

1. Determinar se um dado objeto existe numa imagem (classificação);
2. Determinar a localização do objeto (deteção).

Para isso, recorre-se às métricas: *Accuracy*, *precision*, *recall*, *f1-score*, *intersection over union* (*IoU*) e *mean average precision* (*mAP*).

Intersection over Union (*IoU*) permite determinar quão bem o objeto foi localizado. Para isso, divide a área de sobreposição entre as fronteiras de referência (*ground truth*) e as detetadas pela sua área de união (Figura 11), através da fórmula: $\frac{Área\ de\ sobreposição}{Área\ da\ União}$.

Assim, deteções de objetos podem ser consideradas pelo modelo como verdadeiras ou falsas consoante o limiar (*threshold*) definido para o *IoU*. Adicionalmente, para cada valor de *IoU* computa-se a curva *precision-recall*, formada pela *precision* máxima para *N* níveis de *recall* (Figura 12). Para obter a *precision* por nível de *recall* ordenam-se os resultados por grau de confiança do detetor e consideram-se limiares do grau de confiança distintos. Como tal, à medida que incluímos mais predições do modelo, com menor grau de confiança, o *recall* aumenta e a *precision* baixa.

Average Precision resulta da média dos valores máximos em precision para os N níveis diferentes de *recall*. Seguidamente, a métrica final, *mean Average Precision* (*mAP*), é calculada fazendo a média da *average precicion* (*AP*) de todas as classes de objetos e/ou de todos os limiares de *IoU* definidos (Figura 13).

[7] http://host.robots.ox.ac.uk/pascal/VOC/
[8] https://www.kaggle.com/c/imagenet-object-localization-challenge
[9] http://cocodataset.org/#detection-eval

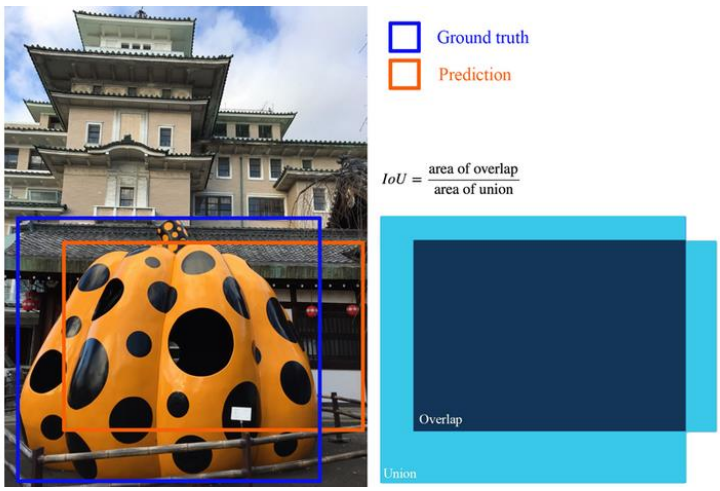


Figura 11 - Ilustração do cálculo da métrica Intersection over Union (IoU) [43].

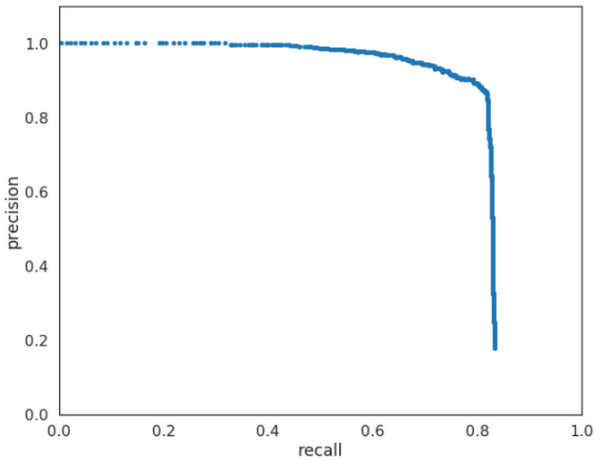


Figura 12 - Exemplo ilustrativo de curva da precision-recall [44].

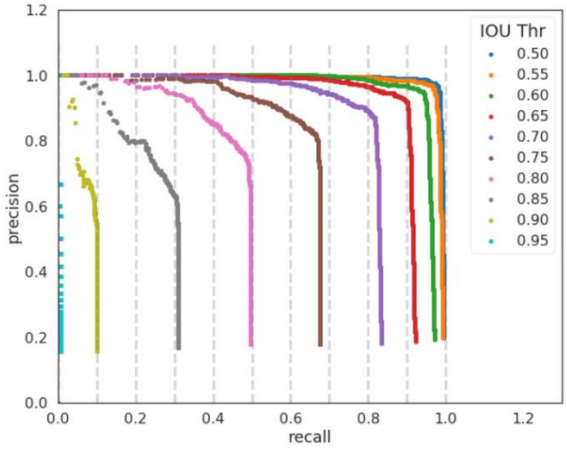


Figura 13 – Exemplo ilustrativo de curvas precision-recall calculadas para vários limiares de IoU [44].

2.4.2 Modelos para reconhecimento de objetos

Desde a introdução das redes neuronais e progresso nos algoritmos de *deep learning* na área da visão por computadores, as estruturas para deteção tornaram-se mais rápidas e precisas [45].

Na revisão da literatura elaborada em torno deste tema, presta-se particular atenção ao sistema para deteção de objetos denominado *You Only Look Once* (*YOLO*), que disponibiliza a terceira versão do sistema (*Yolo v3*) no final do primeiro trimestre de 2018.

Comparando com outras redes para classificação por região, como o *Fast R-CNN* que desempenha a deteção em várias propostas de região e como tal desempenha múltiplas previsões para várias regiões numa imagem, a arquitetura *YOLO* reformula a deteção de objetos num único problema de regressão.

Sumariamente, consiste numa única rede convolucional que simultaneamente prevê múltiplas caixas delimitadoras e probabilidades de classe para essas caixas (Figura 14). Ao dividir a imagem numa matriz *SxS*, a célula que contiver o centro do objeto é responsável por detetar esse objeto.

Adicionalmente, cada célula prevê B caixas delimitadoras e o grau de confiança para essas caixas. O grau de confiança reflete o quão confiante o modelo é relativamente ao facto de existir um determinado objeto nessa caixa e quão bem a caixa se enquadra no objeto. Formalmente, a confiança é definida pela fórmula: *Pr (Objeto) * IoU*.

Com isto, se não existir objeto nessa célula o valor da confiança será zero, caso contrário, pretende-se que seja igual ao produto entre a interseção sobre a união (*IoU*) entre a caixa prevista e os valores de referência (*ground truth*) e a probabilidade desse objeto constar nas fronteiras detetadas.

O treino é feito em imagens completas e o desempenho de deteção é diretamente otimizado.

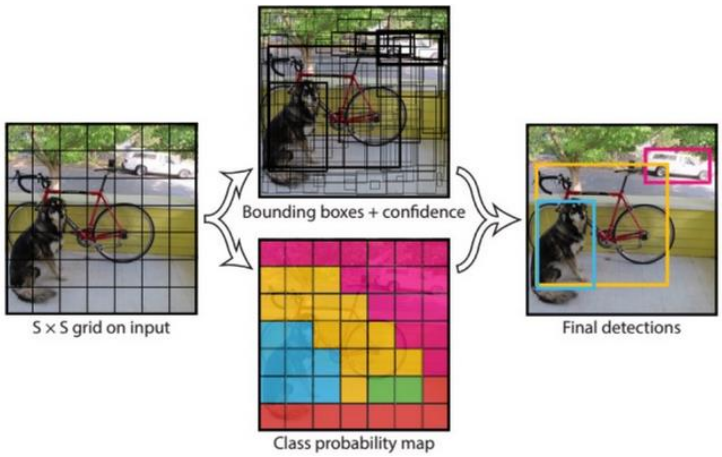


Figura 14 - Modelo YOLO. Divide a imagem numa matriz SxS e para cada célula prevê B caixas delimitadoras, grau de confiança para essas caixas e C probabilidades de classe [46].

Backbone	**Top-1**	**Top-5**	**Bn Ops**	**BFLOP/s**	**FPS**
Darknet-19	74.1	91.8	7.29	1246	**171**
RestNet-101	77.1	93.7	19.7	1039	53
RestNet-152	**77.6**	**93.8**	29.4	1090	37
Darknet-53	77.2	**93.8**	18.7	**1457**	78

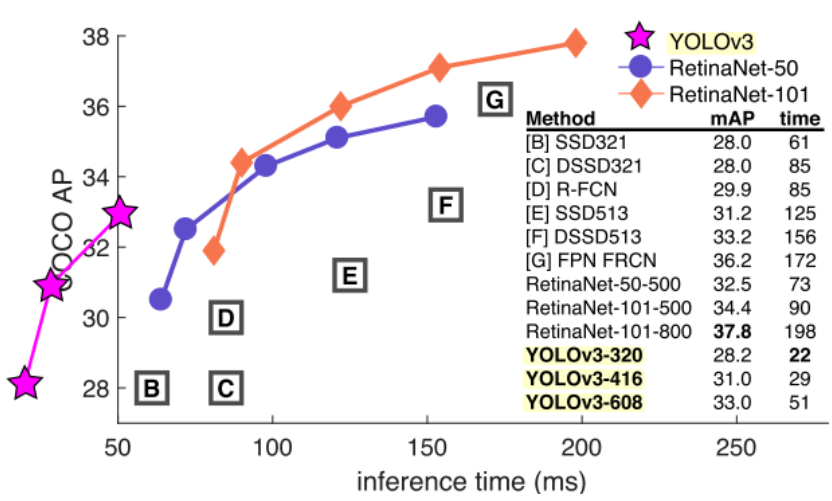
Tabela 6 - Comparação de Backbones com base na precisão e biliões de Operações (Bn Ops e Biliões de operações de ponto flutuante por segundo (BFLOP/s) [47].

Este modelo unificado traz vários benefícios quando comparado com os métodos mais tradicionais para deteção de objetos:

1. **Rapidez na classificação** – A versão 3 incorpora uma rede convolucional de 53 camadas com tamanhos 3x3 e 1x1 (Figura 15), maior e mais potente que a versão anterior *Darknet-19* (inerente ao *YoloV2*), mais preciso que a concorrente *ResNet-101* e duas vezes mais rápida que a *ResNet-102* (Tabela 6). Assim, o *backbone Darknet-53* tem um desempenho equiparável aos classificadores do estado-de-arte, mas com maior rapidez de processamento (Figura 16), menos operações em vírgula flutuante e a maior taxa de operações em vírgula flutuante por segundo registada para o efeito.
2. **O contexto global da imagem é considerado** – Vê a imagem completa durante as fases de treino e teste codificando informação contextual acerca das classes assim como a sua aparência. Assim, para prever cada caixa delimitadora a rede usa *features* da imagem toda. Adicionalmente, para uma dada imagem, prevê todas as caixas delimitadoras ao longo de todas as classes em simultâneo.
3. **Alto índice de generalização do treino** – Quando treinado em imagens naturais e testado em obras de arte, supera significativamente métodos de deteção de topo como *DPM* e *R-CNN*. Isto significa que é altamente generalizável e que é menos propenso a falhar quando aplicado a novos domínios ou dados de entrada inesperados.
4. **Melhor deteção de objetos pequenos** – O Yolo v3 faz predições a três escalas diferentes (Figura 17) através do *downsampling* (i.e., reduzir o mapa de *features* dimensionalmente) das imagens de entrada por 32, 16 e 8, respetivamente. As camadas de *upsampling* são concatenadas com as camadas anteriores para ajudar a preservar *features* que auxiliam na deteção de objetos pequenos. A camada 13 x 13 é responsável por detetar objetos grandes enquanto que a camada 52 x 52 deteta objetos pequenos, com a camada 26 x 26 a detetar objetos de dimensão média.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figura 15 - Arquitetura Darknet-53 [43].



Method	mAP	time
[B] SSD321	28.0	61
[C] DSSD321	28.0	85
[D] R-FCN	29.9	85
[E] SSD513	31.2	125
[F] DSSD513	33.2	156
[G] FPN FRCN	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	**37.8**	198
YOLOv3-320	28.2	**22**
YOLOv3-416	31.0	29
YOLOv3-608	33.0	51

Figura 16 - Tradeof entre average precision (AP) e tempo de inferência de diferentes métodos para deteção de objetos [48].

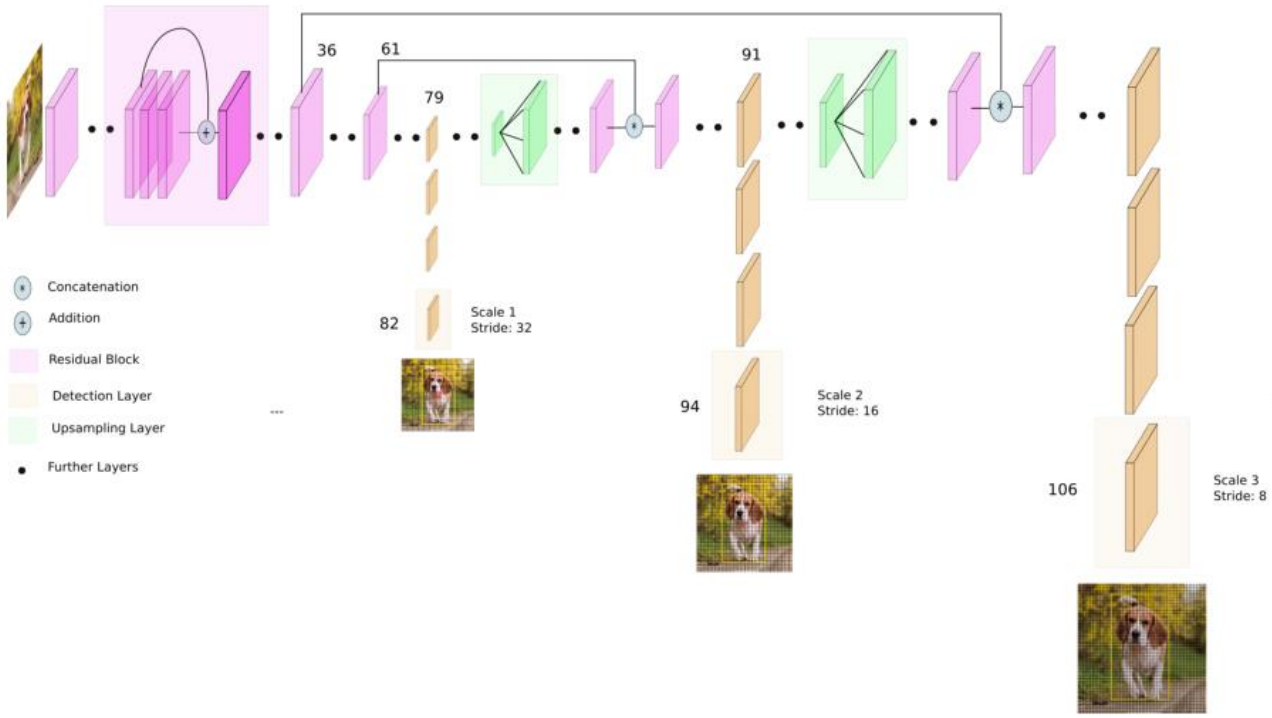


Figura 17 - Arquitetura da rede Yolo v3 [49].

Capítulo 3 – Desenvolvimento da Solução

3.1 Modelo conceptual

Em resultado da revisão literária elaborada e de acordo com a metodologia de investigação selecionada, sucede-se a fase de proposta de novos esforços de investigação no âmbito dos problemas identificados.

Tal como foi delineado previamente, pretende-se criar um modelo para o reconhecimento de interações de clientes com produtos posicionados em prateleiras de lojas.

Verificou-se através da revisão da literatura, nomeadamente no capítulo 2.2.5 referente às abordagens já existentes com o mesmo propósito aplicacional, que existe um conjunto reduzido de estudos em torno do tema. Adicionalmente, as particularidades do contexto aplicacional - interação de clientes com produtos em prateleiras, seguimento e distinção de clientes, posicionamento e ângulo de captação do sensor com mais restrições - refletem-se em adaptações que deverão ser tidas em conta aquando da construção do modelo.

Relativamente às interações de clientes com produtos, o modelo deve ser capaz de detetá-las e classificá-las, devendo por isso ter em conta a presença de produtos nas mãos dos clientes.

Para fazer o seguimento e distinção de clientes é importante manter um identificador único do esqueleto do cliente.

O posicionamento do sensor de forma a que ângulo de captação ocorra na vista de topo irá condicionar a capacidade de implementar as funcionalidades descritas anteriormente.

Dado que a revisão da literatura revelou que existe uma forte carência de modelos desenvolvidos com base nestas particularidades, deduz-se ser importante para a comunidade científica a construção e disponibilização de um conjunto de dados para desenvolver e avaliar modelos destinados ao reconhecimento de ações de clientes em espaços de vendas.

Em resposta aos problemas identificados, nesta secção propõe-se um modelo conceptual que define, em linhas gerais, os vários módulos e respetivas relações que compõem o modelo final.

O mesmo foi desenhado de acordo com os objetivos de investigação e do conhecimento adquirido na revisão da literatura. Dessa forma, corresponde à segunda fase da metodologia *Design Science Research,* denominada *Suggestion,* e dá origem a um modelo inovador ao articular de

forma diferente o conhecimento já existente para resolver o problema enfatizado na secção anterior.

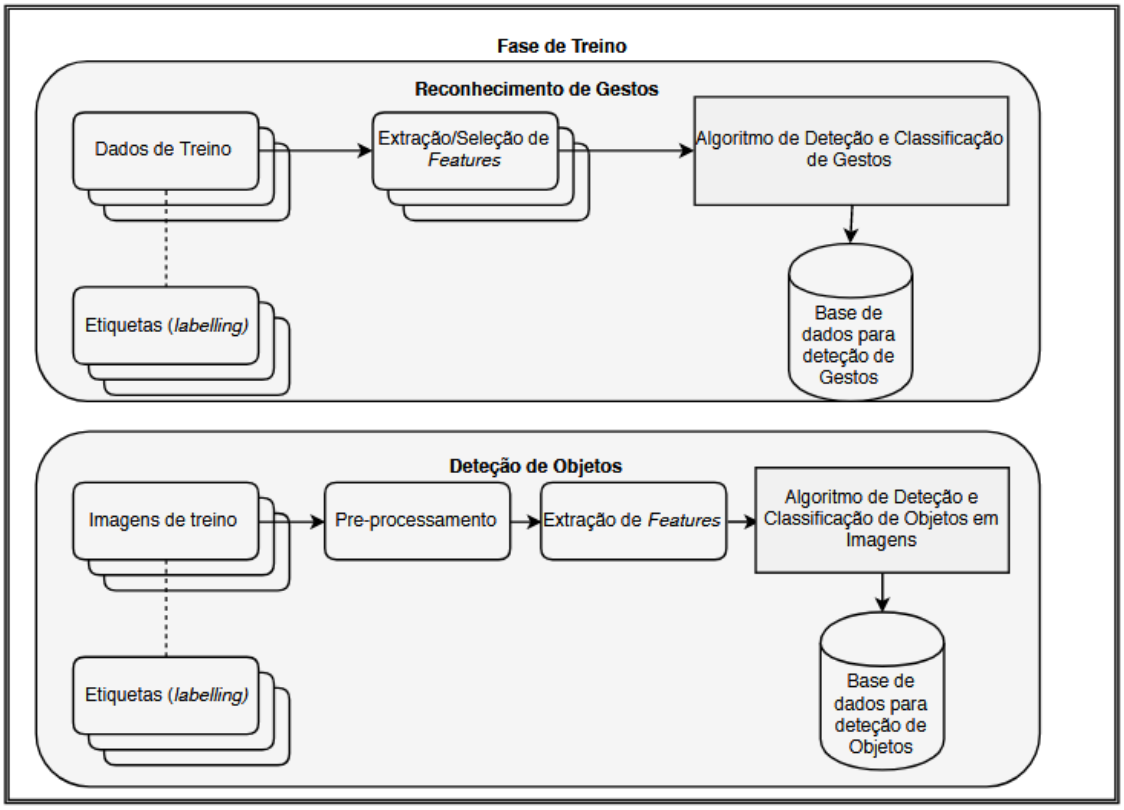


Figura 18 - Modelo conceptual (Fase de Treino)

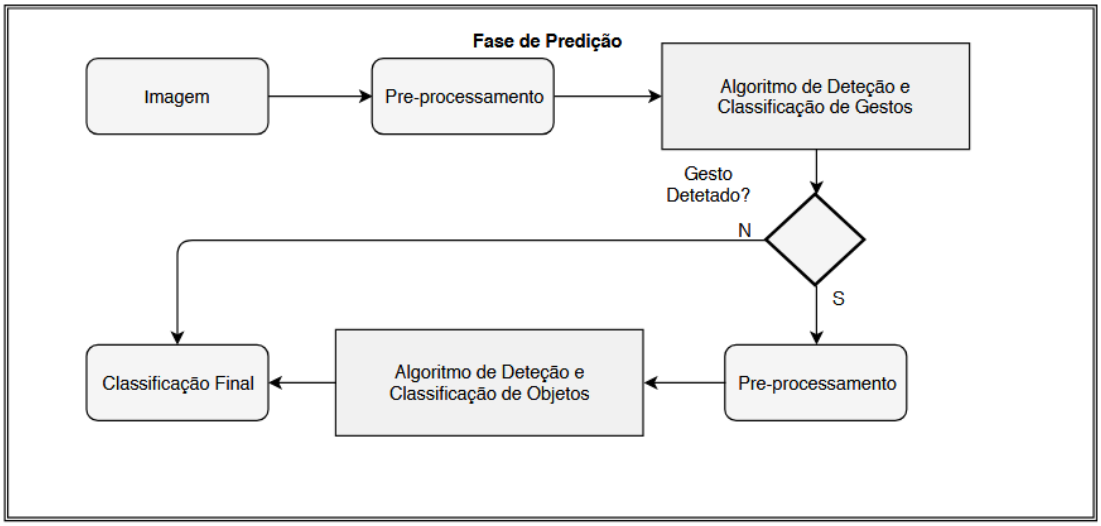


Figura 19 - Modelo conceptual (Fase de Teste)

Em termos gerais, o modelo final (Figura 19) deverá receber um fluxo contínuo de imagens RGB-D (cores e profundidade) e reconhecer as interações dos clientes com os produtos das prateleiras em loja por meio da deteção de gestos e objetos.

Para conceber ambos os algoritmos para reconhecimento de gestos e objetos é necessária uma fase de treino para cada um destes que, conceptualmente, se sumariza na Figura 18. Tal como sugere a Figura 18, para treinar qualquer um dos algoritmos de classificação (gestos e objetos) será necessário fornecer dados etiquetados, sucedendo-se uma fase de extração de features, obtendo-se, por fim, o modelo para detetar e classificar gestos/objetos.

Para testar o modelo proposto (Figura 19) será necessário integrar os modelos para deteção e classificação de gestos e objetos num único sistema. Em linhas gerais, no modelo para teste deverão existir duas etapas de pré-processamento anteriores a cada fase de classificação de forma a que os dados sejam tratados de acordo com os dados de entrada esperados pelos classificadores de gestos e objetos.

Em suma, o pré-processamento necessário ao classificador de gestos será responsável pelo reconhecimento do esqueleto, extraindo a informação das articulações que o compõem. Caso seja reconhecido um gesto por parte do classificador importa averiguar a existência de objetos nas mãos da pessoa, sendo esta a região de interesse ao detetor de objetos. Como tal, o pré-processamento anterior ao detetor de objetos passa por corretamente extrair a referida área de interesse em torno das mãos.

Partindo do modelo conceptual, importa agora selecionar técnicas e métodos para classificar gestos e detetar objetos tendo em consideração o seu enquadramento contextual, complexidade de implementação, recursos disponíveis e o seu desempenho.

Remetendo para as cinco categorias de abordagens para o reconhecimento de ações previamente revistas na secção 2.2.4, os métodos baseados em *features* extraídas das articulações do esqueleto e deteção de partes do corpo sugerem ser uma abordagem mais apropriada face às restantes. Esta abordagem diferencia-se das abordagens baseadas em *features* obtidas via silhuetas 3D, nomeadamente por serem invariantes à localização e respetivo ângulo de captação do sensor e invariantes à escala do corpo captado.

Modelos baseados em *features* de localização espácio-temporal apresentam, igualmente, dependência em relação ao ângulo de captação do sensor. Adicionalmente, refere-se na literatura que o recurso a esta abordagem condiciona o desempenho da aplicação em tempo real [4], sendo esta uma das características desejadas no modelo.

As *features* de ocupação locais 3D aparentam ser as mais indicadas para a deteção de interações pessoa-objeto, potencialmente cliente-produto, e como tal um eventual complemento interessante para as *features* extraídas das articulações do esqueleto visto que estas não nos dão informação relativa aos objetos envolventes. Importa relembrar que [32] desenvolve um modelo que combina estes dois tipos de *features*. No entanto, para efeitos de reconhecimentos de gestos, a complexidade de implementação é superior pelo que é dada preferência ao recurso a *features* baseadas nas articulações do esqueleto humano.

Relativamente às *features* via fluxo ótico 3D constata-se que o seu custo computacional é elevado e, consequentemente, dificulta o desempenho em tempo real. Acresce ainda o facto destas técnicas estarem numa fase preliminar de desenvolvimento e que se traduz em complexidades de implementação adicionais.

Importa ainda destacar que, ao converter a imagem humana em informação do seu esqueleto, salvaguardam-se problemáticas relacionadas com a privacidade, visto que esta é a única informação guardada e necessária ao algoritmo de reconhecimento de gestos. Tendo em conta o contexto de aplicação do modelo proposto, esta característica é uma mais valia do sistema.

Em resultado da análise anterior, optar-se-á por abordagens baseadas em *features* extraídas do conjunto de articulações do esqueleto.

No que toca a modelos para deteção e classificação de objetos optou-se pelo uso da rede convolucional *Darknet Yolo v3*, em resultado da revisão literária, nomeadamente os benefícios apresentados no final da secção 2.4.2.

3.2 Tecnologias utilizadas

De forma a desenvolver o modelo proposto na secção anterior deverão ser feitas decisões relativas às tecnologias, ferramentas ou técnicas a utilizar para implementar as seguintes componentes do modelo:

- Deteção e classificação de gestos (sec. 3.2.1);
- Deteção e classificação de objetos (sec. 3.2.2);
- Sistema integrado para deteção e classificação de interações (sec. 3.2.3).

Como tal, ao longo das secções seguintes aprofunda-se e fundamenta-se as escolhas tecnológicas realizadas para cada módulo que constitui o modelo conceptualizado anteriormente.

Importa referir que, dada a decisão tomada na secção 2.1.3 – Adoção do sensor *Microsoft Kinect v2.0* – o leque de opções tecnológicas disponíveis para cada componente do modelo será restringido àquelas que forem compatíveis com o sensor selecionado.

3.2.1 Deteção e classificação de gestos

Tal como referido, a classificação de gestos deverá ser feita recorrendo a *features* baseadas em articulações e partes do esqueleto.

Recordando a revisão literária elaborada, nomeadamente a secção 2.2.1, para extrair o conjunto estruturado de articulações e respetivas ligações a partir dos mapas de profundidade produzidos pelo sensor existem as ferramentas *Microsoft Kinect SDK* e *OpenNi*.

Das várias funcionalidades disponibilizadas pela *Microsoft Kinect SDK 2.0* destaca-se o *Kinect Studio* e o *Visual Gesture Builder*.

O *Kinect Studio* é uma ferramenta que permite gravar e reproduzir fluxos de tramas de profundidade e cor produzidos pelo sensor *Kinect*.

Por sua vez, o *Visual Gesture Builder* (*VGB*) é um construtor de detetores de gestos que usa aprendizagem automática e informação do esqueleto extraído via *Kinect Studio* para definir um gesto. Para tal, múltiplos clipes com informação do esqueleto são etiquetados nos momentos em que o gesto ocorre de forma a serem utilizados para extrair *features* que definem o gesto. Após a fase de treino, resulta um classificador dos gestos definidos que pode ser usado numa aplicação a desenvolver para detetar diferentes gestos de múltiplos sujeitos.

Para definir um gesto nesta ferramenta deverá ser selecionado o seu tipo, discreto ou contínuo, o que resultará no recurso a um algoritmo de classificação distinto. No caso dos gestos discretos recorre-se ao *AdaBoostTrigger* e para gestos contínuos o *Random Forest.* O primeiro consiste num detetor binário que determina se o sujeito está a desempenhar determinado gesto e o grau de confiança do classificador, pelo que se adequa a gestos simples. O segundo mostra o progresso, em termos percentuais, enquanto o sujeito desempenha o gesto e é particularmente útil no mapeamento dos movimentos do sujeito em animações ou na combinação de vários gestos discretos num contínuo. Na Figura 20 sumariza-se o processo de criação de um detetor de gestos recorrendo ao *VGB*.

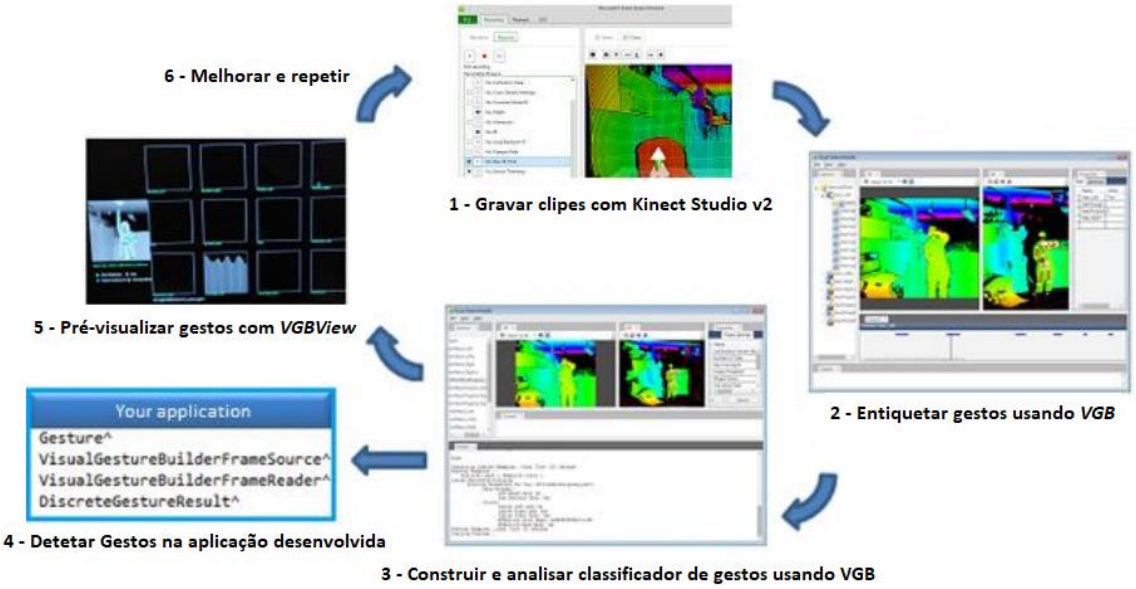


Figura 20 - Processo orientado por dados para criação do detetor de gestos usando VGB [50].

Para a correta utilização desta ferramenta recorre-se ao *whitepaper* [50] associado à mesma, disponibilizado pela Microsoft, assim como aos vídeos instrucionais disponíveis em [51] e [52] onde são descritos os passos envolvidos na criação de um detetor de gestos customizado.

Em alternativa à ferramenta *VGB*, considera-se o projeto *Gesture Recognition Toolkit* (*GRT*), de código aberto e com uma biblioteca de aprendizagem automática C++ para reconhecimento de gestos em tempo real. Esta ferramenta é compatível com o sensor *Kinect*, recorrendo aos programas complementares *OpenFrameworks* e *Synapse*, e é integrável com projetos C++. À semelhança do *VGB*, permite treinar um modelo com quaisquer gestos pretendidos. No entanto, difere do mesmo ao permitir o uso de algoritmos para processamento e extração de *features* customizados, bem como a alteração do algoritmo de aprendizagem automática do classificador (*AdaBoost*, Árvores de Decisão, *Dynamic Time Warping*, *Support Vector Machines*, entre outros especificados em [53]). Adicionalmente, o *GRT* inclui um conjunto de algoritmos para tratar o pré-processamento, pós-processamento e extração de *features*.

Verifica-se assim que o *Gesture Recognition Toolkit* apresenta um grau de customização consideravelmente superior ao do *Visual Gesture Builder*, acompanhado, no entanto, de uma maior complexidade de implementação.

Ainda assim, o recurso à ferramenta *GRT* foi a primeira opção no que toca aos desenvolvimentos iniciais do modelo. No entanto, a sua instalação de acordo com o guião adaptado

à *Kinect*, disponibilizado em [54], originou diversos problemas associados à instalação dos *drivers* necessários visto que recorre a versões do *OpenNi* desatualizadas. O problema foi recriado numa máquina virtual *Windows 10 Enterprise* [55] através do *VMware WorkStation 14 Player* [56] e o problema permanece.

Adicionalmente, os dados de treino eventualmente recolhidos seriam extraídos com recurso a software desatualizado para extração de articulações das imagens de profundidade. Acresce o facto da fase de treino recorrendo ao *GRT*, segundo o guião em [54], não ser a desejada visto que o processo de treino suporta apenas clipes de vídeo com uma duração máxima de 5 segundos. Contrariamente, o *VGB* não impõe limites para a duração dos clipes usados no treino.

Para classificar gestos em tempo real, a ferramenta *Visual Gesture Builder* oferece a funcionalidade *VGBView* onde é possível ver a confiança/progresso de cada gesto discreto/contínuo a ser detetado em tempo real.

No entanto, pretende-se criar uma aplicação que integre tanto o classificador de gestos como o de objetos. O *VGBView* é útil para rapidamente prototipar diferentes detetores de gestos testando o seu comportamento em tempo real mas, visto ser de código fechado, não é possível interligá-lo com um eventual classificador de objetos. Para tal, recorre-se à biblioteca de classes *NtKinect* de código aberto e sobre a licença *MIT*, que permite usar as funções da *Kinect v2* na linguagem de programação C++ em ambientes de desenvolvimento integrados, como o *Visual Studio Community 2015/2017*.

Esta biblioteca está disponível *online* em [57] e os guias de utilização relativos às suas várias funcionalidades foram usados como referência.

Com o recurso a esta biblioteca consegue-se uma abordagem tecnológica de complexidade de implementação adequada ao tempo disponível para desenvolvimento da dissertação e que vai ao encontro dos recursos disponíveis.

3.2.2 Deteção e classificação de objetos

Para criar um modelo customizado baseado na rede *Darknet Yolo v3* para detetar os objetos usados na experiência, especificados em 3.3.2, foi usado como referência o projeto de código aberto disponível em [58].

O projeto *Darknet* foi compilado no ambiente de desenvolvimento *Visual Studio 2017* com recurso às tecnologias:

- *CUDA 9.1*: Biblioteca para uso generalizado do GPU, utilizada para processamento das imagens, tanto para treino como para deteção e classificação.
- *cuDNN 7.0*: Biblioteca de *deep learning* otimizada para GPU.
- *OpenCV 3.0*: Biblioteca de visão por computador, utilizada para o tratamento de fluxos de vídeo em tempo real.

A rede *Darknet Yolo v3* é posteriormente adicionada via *DLL* ao projeto já integrado com a biblioteca *NtKinect*. Após configurar a rede de acordo com as classes/objetos que se pretendem detetar e classificar, para fazer deteções numa dada imagem, resta eleger um dos estados da rede guardados durante o seu treino.

Após aprofundar as ferramentas para extração/seleção de *features* e os algoritmos selecionados para deteção de gestos e objetos, resta detalhar a forma de como estes são integrados numa aplicação desenvolvida na linguagem de programação *C++*.

3.2.3 Sistema integrado para deteção e classificação de interações

Na Figura 21 apresenta-se uma visão macro de todo o sistema. Como se pode observar, o sistema está dividido a tracejado segundo três partes principais: as duas partes superiores referem-se ao processo de treino dos modelos, discutido em 4.2.1 e 4.3.1, e a inferior ao teste da integração de ambos os modelos cujo desenvolvimento será objeto de análise nesta secção.

No que toca ao teste do modelo proposto, para que seja simulada a transmissão de tramas em tempo real, recorre-se ao *Kinect Studio v2.0* e ao *Serviço Kinect*. Desta forma, ao reproduzir um clipe no *Kinect Studio* ir-se-á simular a aquisição de tramas pelo sensor *Kinect* visto que, após ligá-lo ao *Serviço Kinect*, a deteção de esqueletos passa a ser feita em tempo real por meio do *Serviço Kinect*.

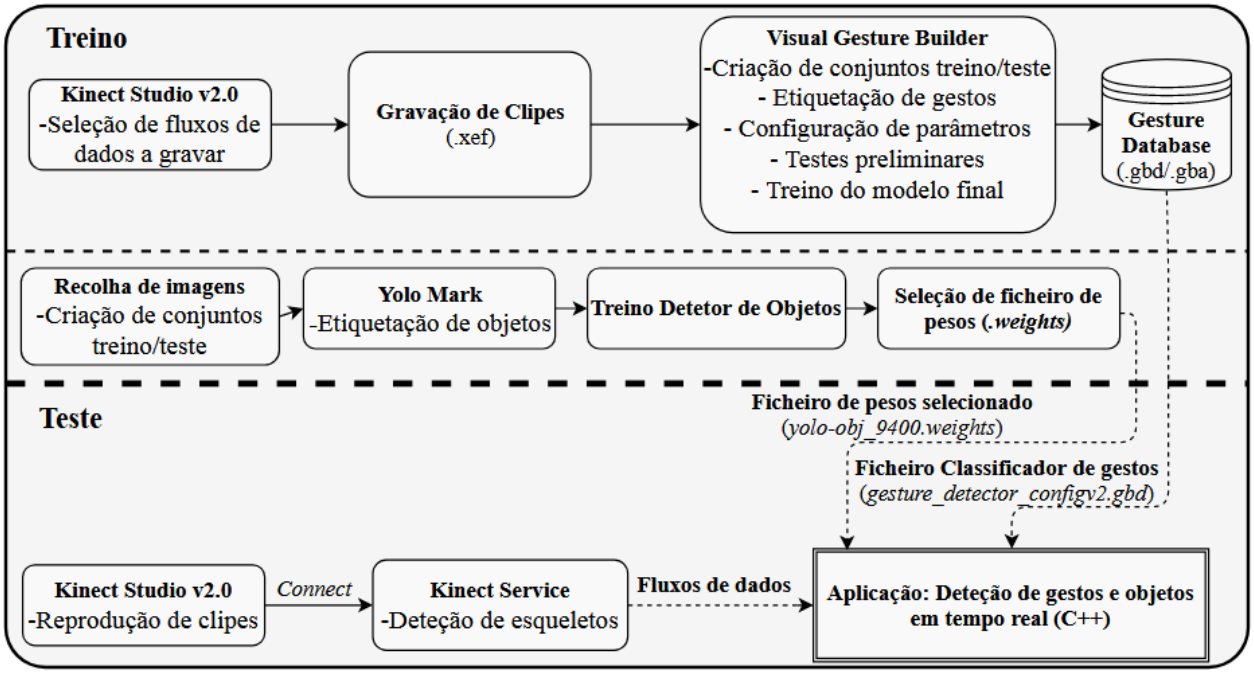


Figura 21 - Sistema para deteção de gestos e objetos em espaços de vendas (visão macro).

Na Figura 22 apresenta-se o fluxograma da aplicação C++ desenvolvida no ambiente Visual Studio 2017 que integra a deteção de objetos recorrendo ao *DLL* da *Darknet Yolo* e deteção de gestos através do pacote de desenvolvimento *Kinect Windows SDK 2.0* e das bibliotecas *NtKinect* e *OpenCv 3.0* (igualmente necessário para deteção de objetos).

Refira-se que a Figura 22 consiste num fluxograma simplificado e meramente ilustrativo do algoritmo desenvolvido, acompanhado dos principais métodos usados. Como tal, nesta secção importa aprofundar detalhes relativos à aplicação desenvolvida, nomeadamente a estrutura do sistema, o algoritmo proposto, os principais problemas/desafios encontrados e respetivas soluções propostas.

Para compreender o algoritmo proposto considera-se mais intuitivo detalhar, em primeiro lugar, os seguintes problemas encontrados:

Inconsistências no número de tramas processadas

É importante que o sistema seja capaz de processar todas as tramas transmitidas pelo sensor, pois o treino dos gestos é feito com um fluxo contínuo de tramas e como tal o detetor espera um fluxo de dados igualmente contínuo. Adicionalmente, os factos (*ground truth*) relativos aos gestos estão associados a intervalos de tramas específicos, pelo que é necessário que a contagem de tramas seja consistente para que se possam avaliar os resultados das deteções corretamente.

Foram detetadas duas causas principais que provocam perdas ocasionais de tramas:

1. Durante a reprodução de clipes via *Kinect Studio* e Serviço *Kinect* podem ocorrer atrasos no processamento do esqueleto e consequentemente perdas de tramas. Como este fenómeno ocorre com raridade, neste tipo de situação optou-se por repetir o teste de determinado clipe e respetiva extração de resultados.
2. Na aplicação desenvolvida com base na *API* do *SDK* da *Kinect* apenas existe um método para adquirir a trama mais recente. Se o processamento referente a uma trama demorar mais do que o intervalo de tempo entre as duas novas tramas seguintes, a primeira trama que se segue será perdida.

Para colmatar a última situação, optou-se pela introdução de múltiplas linhas de execução (*multithreading*). A linha de execução 1 é responsável por adquirir novas tramas e guardá-las num buffer. Este buffer é partilhado com a linha de execução 2 que irá tratar do processamento da trama mais antiga no buffer.

Falhas no acompanhamento de identidade do esqueleto

Cada esqueleto, num dado instante de um clipe, tem um número inteiro aleatório identificador.

No entanto, se o esqueleto deixar de ser detetado e mais tarde voltar a ser reconhecido é-lhe atribuído um novo identificador.

A necessidade reidentificação do esqueleto é fundamental não só para manter o registo das ações de um dado sujeito como para permitir a correta extração de resultados.

A abordagem adotada para solucionar esta problemática está adaptada a clipes de um ou dois sujeitos e baseia-se na atribuição da etiqueta ‘A’ ao esqueleto mais próximo do sensor e ‘B’ ao esqueleto mais longe.

Desta forma, um dado facto contém a etiqueta do esqueleto que desempenhou o gesto (A/B) e sempre que ocorre uma deteção de gesto o seu registo irá ser acompanhado da etiqueta do esqueleto. Para determinar se uma dada deteção pertence ao esqueleto ‘A’ ou ‘B’ os valores da profundidade da articulação da anca são comparados e a etiqueta é atribuída.

Embora se considere uma boa solução para o problema identificado, esta abordagem irá falhar sempre que, num dado instante, o esqueleto A deixar de ser reconhecido e o esqueleto ‘B’ desempenhar um gesto, visto que será atribuído de forma errada a etiqueta ‘A’.

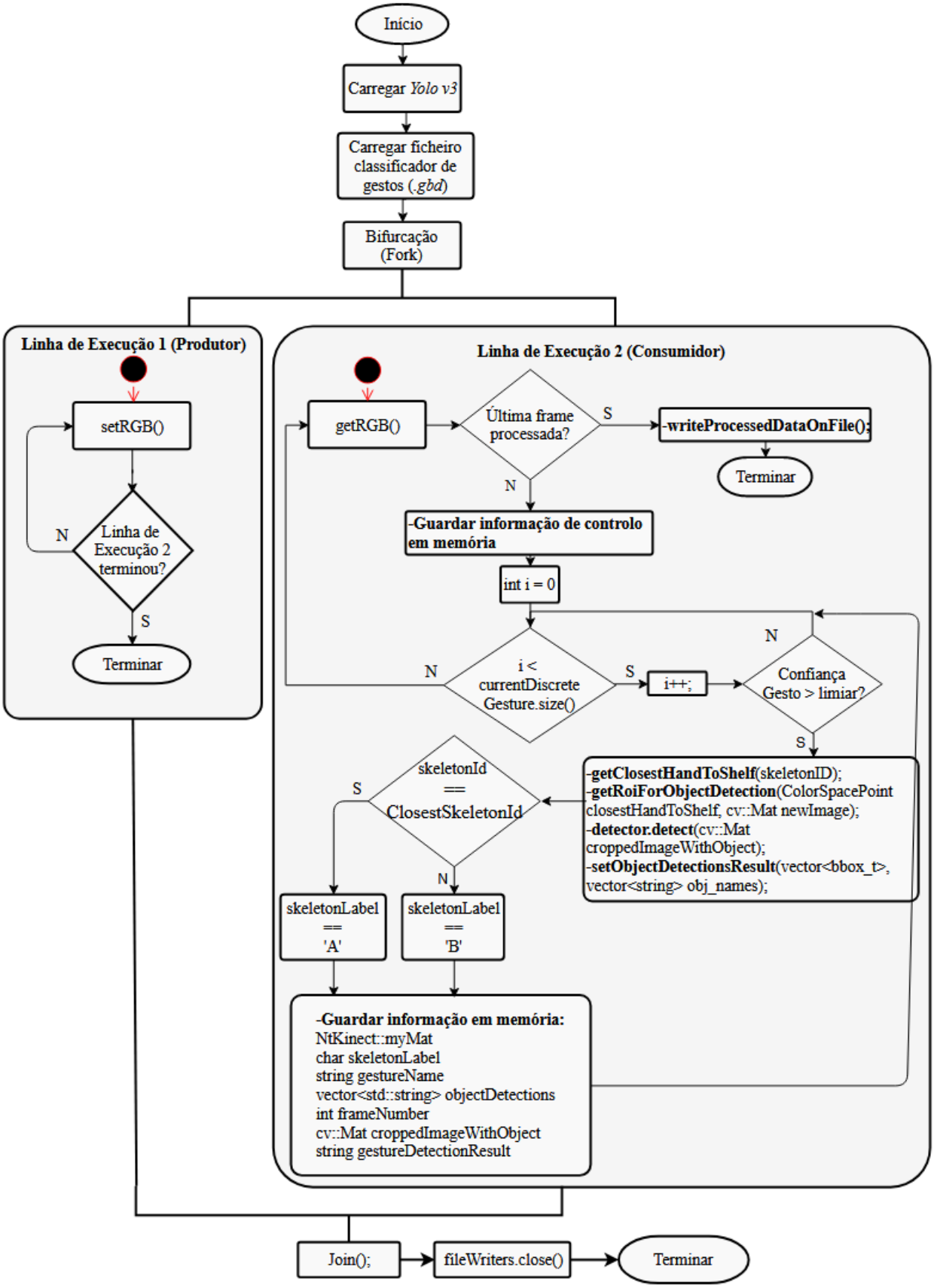


Figura 22 - Fluxograma aplicação C++ que integra deteção de objetos e gestos.

Fluxograma do modelo para deteção de interações com prateleiras

Face aos problemas identificados, desenvolveu-se um método para registar informação de controlo por trama processada, permitindo quantificar o impacto destes problemas no desempenho.

Importa referir que a criação do *buffer* é necessária para minimizar problemas relacionados com inconsistências no número de tramas, no entanto, irá pôr em causa o funcionamento da aplicação em tempo real. Para avaliarmos o atraso introduzido na aplicação, a informação de controlo regista por cada trama processada o número de tramas em *buffer*.

Por outro lado, regista também o número de esqueletos detetados por trama. Isto permite associar deteções incorretas a falhas na correta captação dos esqueletos.

Tendo em conta os problemas/soluções discutidos, o algoritmo desenvolvido para tratar do registo de deteções ilustrado na Figura 22 funciona da seguinte forma:

1. **Inicialização** – A rede neuronal convolucional Darknet Yolo v3 é carregada em memória no arranque do programa para permitir a rápida deteção de objetos.
2. **Aquisição de tramas** – A linha de execução 1 invoca num ciclo infinito o método *setRGB()*. Este método está codificado no ficheiro *NtKinect.h* e foi alterado para permitir a criação do *buffer* de estruturas *myMat* (*std::queue<myMat>)* onde cada objeto contem, entre outras informações, dados dos esqueletos, gestos detetados e o identificador do esqueleto mais próximo.
3. **Processamento de tramas** – Em paralelo com o passo 2, a linha de execução 2 executa o seguinte algoritmo:
    a. Se a última *trama* de um clipe já foi processada a execução termina e as ocorrências de gestos guardadas em memória são registadas num ficheiro de texto, assim como os objetos detetados.
    b. Caso contrário:
        i. Guarda informação de controlo em memória: Número da trama em processamento, o número de esqueletos detetados, os identificadores dos esqueletos e a profundidade para cada esqueleto.
        ii. Processa o objeto *myMat* mais antigo na fila:
            1. Para cada gesto detetado, caso o grau de confiança do detetor seja superior a um *threshold* (definido empiricamente a 0.2, visto ser o grau de confiança

médio que capta a generalidade das deteções) executam-se os próximos passos, caso contrário continua a iteração sobre o vetor de deteções.

2. Extrai-se a localização da mão mais próxima da prateleira.
3. Determina-se a *ROI* (região de interesse) da imagem completa para fazer a deteção de objetos. Corresponderá a um quadrado de 130 x 130 pixéis em torno da mão mais próxima da prateleira. A dimensão do quadrado (130 x 130) é o valor resultante de vários ajustes feitos até que se atingisse o valor mais baixo possível que capte os objetos na sua integridade.
4. Executa-se a deteção de objetos sobre a imagem recortada.
5. Guarda-se em memória os resultados de deteção.

4. **Terminar** – Ambas as linhas de execução são terminadas, o fluxo associado à escrita em ficheiros é fechado e o programa termina.

3.3 Construção do conjunto de dados

Projetos igualmente inseridos na área da visão por computadores são vulgares, no entanto, as particularidades do contexto de aplicação deste modelo (espaço de vendas), criam a necessidade de recolher dados específicos desse contexto, não permitindo a reutilização daqueles que já estão disponíveis publicamente para fins de investigação.

3.3.1 Aquisição de material

Para simular os cenários pretendidos contou-se com o apoio das entidades: (1) *ISTAR-IUL*, na cedência do sensor *Microsoft Kinect v2*, (2) *Vitruvius Fablab*, pelo empréstimo de uma estrutura em madeira para colocar o sensor e captar uma visão de topo e (3) *IT-IUL*, pela sala disponibilizada para recolha de dados.

A gravação de clipes foi feita via *Kinect Studio 2.0*, captando os fluxos de dados apresentados na Tabela 7.

Para recolher os dados pretendidos angariaram-se 20 sujeitos voluntários que, num ambiente de loja simulado com duas prateleiras e quatro produtos distintos em quantidades variáveis, desempenharem três guiões de ações distintos, enquadrados nos objetivos da tese.

Fluxo de Dados	Descrição
NUI Body Frame	Fluxo de tramas que contem toda a informação de deteção computada em tempo real acerca dos sujeitos que estão no campo de visão do sensor.
NUI Depth	Fluxo de tramas onde cada pixel representa a profundidade associada ao objeto mais próximo visto nesse pixel.
NUI IR	Fluxo de tramas adquiridas por infravermelhos que permitem visualizar a cena com brilho consistente independentemente do local ou brilho deste.
NUI Sensor Telemetry	Fluxo de dados internos necessário à reprodução de gravações no *Kinect Studio*

Tabela 7 - Fluxos de dados captados na ferramenta Kinect Studio.

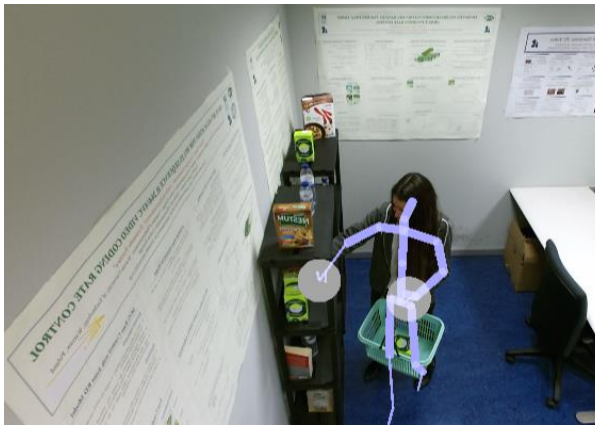


Figura 23 - Trama ilustrativa do cenário em clipe de um sujeito.

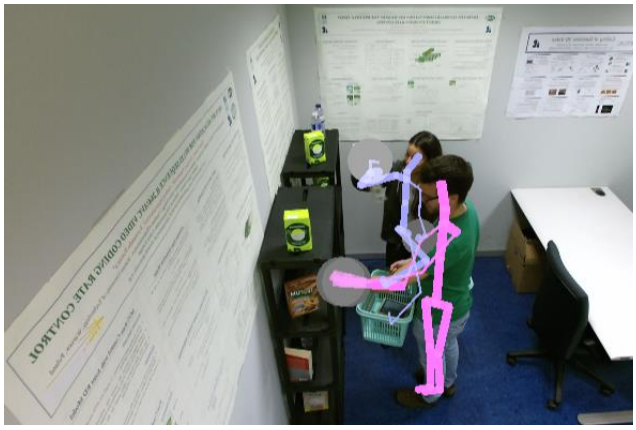


Figura 24 - Trama ilustrativa do cenário em clipe de dois sujeitos.

O conjunto de dados é composto por clipes de um e dois sujeitos captados a interagir com o cenário ilustrado nas Figuras 23 e 24, durante aproximadamente um minuto.

Nos três guiões desempenhados pelos sujeitos, presentes no Apêndice A, solicita-se que estes desempenhem cinco tipos de interações diferentes sobre os diversos andares das prateleiras. Desta forma, é possível obter uma amostra das várias posturas e movimentos associados às interações com uma prateleira em loja. O propósito dos vários guiões está relacionado com o teste de cenários com diferentes complexidades:

- Guião A – De complexidade mais reduzida, capta apenas um sujeito a interagir com uma prateleira. No entanto, existem oclusões provocadas pelo próprio corpo que irão dificultar a correta captação do esqueleto ou do objeto alvo de interação.
- Guião B – Com complexidade intermédia visto que dois sujeitos interagem, cada um, com apenas uma prateleira.

- Guião C – Corresponde ao de maior complexidade porque são captados dois sujeitos que trocam de posições e podem interagir com a prateleira do sujeito ao seu lado. Com isto, simulam-se as movimentações típicas dos clientes nos espaços de vendas, assim como a concorrência no acesso às prateleiras. Espera-se por isso que seja mais difícil manter uma correta captação dos dois esqueletos por existirem mais oclusões entre sujeitos.

Relativamente aos produtos posicionados nas prateleiras que serão objetos para deteção e classificação, foram selecionados estrategicamente 4 tipos distintos de produtos. O desafio é o sistema ser capaz de distinguir produtos que são visualmente muito diferentes e outros que apresentem semelhanças visuais entre eles.

Posto isto, os produtos selecionados são:

1. Caixas de cereais *Kellogg's* e *Nestum* dado terem dimensões e cores diferentes.
2. Embalagens de leite de uma só marca.
3. Garrafas de água 0.5l com diferentes rótulos/marcas.
4. Livros de diferentes dimensões, formas e cores.

O desafio na distinção de produtos pode ser feito a vários níveis. Optou-se por classificá-los ao nível do tipo de produto em si, não discriminando a marca, sendo estas as classes para deteção Ainda assim, as caixas de cereais são potencialmente confundíveis com livros dada a similaridade das suas formas, ao contrário das embalagens de leite e águas que têm cores e formas distintas.

3.3.2 Classes de interações, gestos e objetos

Tendo em conta o contexto aplicacional do modelo, definiram-se as seguintes interações com as prateleiras em loja:

- **Interação positiva** – O cliente retira um produto da prateleira.
- **Interação neutra** – O cliente estende o braço, interagindo com a prateleira, mas não retira qualquer produto.
- **Interação negativa com novo produto** – O cliente volta a colocar um produto na prateleira e retira outro.
- **Interação negativa sem novo produto** – O cliente volta a colocar um produto na prateleira.

- **Interação nula -** Quando não existe qualquer interação para um dado intervalo de tramas.

Para conseguir distinguir as referidas interações propõe-se segmentar os movimentos associados às interações com a prateleira em dois gestos principais a detetar e classificar: a extensão do braço para alcançar a prateleira e a recolha/flexão do braço.

Se forem detetadas uma extensão de braço sem produto e uma flexão com produto esta informação pode generalizar-se numa interação do tipo positivo. Por outro lado, se a flexão não contiver um produto infere-se uma interação neutra.

Quando é detetado um produto na mão durante a extensão resta saber se existe um produto nas mãos durante a flexão seguinte para se poder inferir o tipo da interação como negativa com/sem novo produto.

Relativamente aos objetos, as classes a classificar correspondem aos vários tipos de produto – Caixas de Cereais, Embalagens de leite, Garrafas de Água e Livros – assim como a classe Mão Vazia.

3.3.3 Conjuntos de treino e teste

Uma extensão ou recolha do braço pode ser feita através de diferentes movimentos e posturas (agachando, inclinando o tronco, mantendo uma postura vertical, entre outras). Visto que se segmenta uma interação em apenas dois gestos, está-se implicitamente a determinar a capacidade da generalização dos diferentes movimentos/posturas associadas a este num só gesto, em alternativa à sua segmentação nessas diferentes posturas (p.ex., extensão do braço agachando).

Consequentemente, considera-se apropriado que a forma/postura do gesto seja utilizada como um critério de seleção de clipes para treino e teste.

Assim, propõe-se testar um conjunto de treino constituído pelos clipes que perfazem 50% das ocorrências de cada forma de desempenhar o gesto e o conjunto de teste constituído pelos restantes clipes que juntos vão ao encontro da distribuição real das diferentes formas de desempenhar o gesto.

Considera-se que um gesto de uma dada interação com a prateleira pode ser desempenhado de cinco formas distintas, conforme o descrito na Tabela 8.

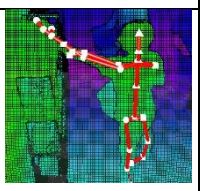
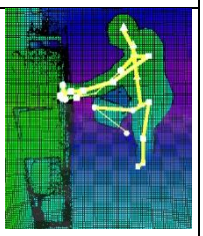
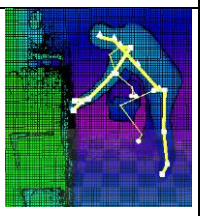
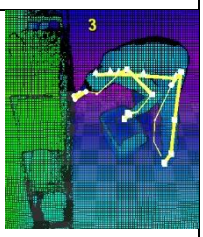
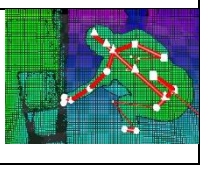
Postura do gesto	Descrição	Exemplo
P1 - Alto	Torso direito e o braço acima do ombro	
P2 - Médio	Torso direito e braço até nível dos ombros	
P3 - Baixo parcialmente inclinado	Torso parcialmente inclinado e braço abaixo dos ombros	
P4 - Baixo e significativamente inclinado	Torso inclinado formando ângulo perpendicular com as pernas e braço abaixo dos ombros	
P5 - Agachamento	Qualquer forma de agachamento	

Tabela 8 – Formas/posturas de gestos extensão/flexão do braço consideradas.

Em seguida, determinou-se manualmente o número de ocorrências de cada postura associada ao desempenho dos gestos para todos os clipes (Tabela 9) e constitui-se o conjunto de treino e teste conforme referido anteriormente.

P1	***P2***	***P3***	***P4***	***P5***	**Total**
575	999	565	153	245	2537

Tabela 9 - Distribuição de posturas de gestos do conjunto de dados recolhido.

No entanto, recorrendo ao *Visual Gesture Builder*, para fazer o treino do classificador de gestos apenas é possível considerar os gestos do sujeito mais próximo (A). Posto isto, consideraram-se duas possíveis configurações:

1. "Ignorar" o facto do conjunto de treino conter menos exemplos por não considerar os gestos do sujeito B, resultando conjuntos de treino e teste com dimensão aproximada em número de clipes.
2. Considerar que o conjunto de treino deve conter 50% das ocorrências totais de cada forma de desempenhar o gesto, pelo que o conjunto de teste será substancialmente reduzido.

De facto, ambas as configurações são plausíveis no que toca ao teste do seu desempenho. No entanto, após extração de resultados de ambas as configurações, verifica-se que os resultados obtidos são bastante próximos. Consequentemente, considerou-se mais apropriado apresentar apenas resultados do primeiro conjunto de treino/teste, tornando a análise de resultados menos exaustiva e repetitiva relativamente às conclusões tiradas e tabelas/gráficos de resultados apresentados.

O termo "conjunto de treino útil" corresponde à contabilização de gestos do sujeito A para treino, denominando-se "útil" visto conter os gestos que de facto são considerados durante o treino.

	P1	***P2***	***P3***	***P4***	***P5***	**Total**
Total	575	999	565	153	245	2537
Conjunto de Treino	290	498	264	70	120	1242
Conjunto de Treino Útil	174	338	150	48	84	794
Conjunto de Teste	285	501	301	83	125	1295

Tabela 10 - Distribuição de formas de desempenhar o gesto por conjuntos de treino/teste.

	P1	***P2***	***P3***	***P4***	***P5***	**Total**
Total	22.66%	39.38%	22.27%	6.03%	9.66%	100.00%
Conjunto de Treino	11.43%	19.63%	10.41%	2.76%	4.73%	48.96%
Conjunto de Treino Útil	6.86%	13.32%	5.91%	1.89%	3.31%	31.30%
Conjunto de Teste	11.23%	19.75%	11.86%	3.27%	4.93%	51.04%

Tabela 11 - Distribuição percentual de formas de desempenhar o gesto com base na Tabela 10.

Para a configuração de conjuntos de teste/treino selecionada obtém-se uma distribuição de 31.30% do total de gestos para o conjunto de treino (útil) e 51.04% para teste, conforme descrito na Tabela 11.

	P1	*P2*	*P3*	*P4*	*P5*
Distribuição real	22.66%	39.38%	22.27%	6.03%	9.66%
Conjunto Teste	22.01%	38.69%	23.24%	6.41%	9.65%

Tabela 12 - Distribuição de formas de desempenhar o gesto face à distribuição real.

Como se pode confirmar na Tabela 12, a distribuição das várias formas de desempenhar o gesto está próxima da distribuição real do conjunto de dados total.

3.4 Testes preliminares ao treino do modelo para deteção/classificação de gestos

Antes de proceder ao treino do modelo foi dada atenção aos parâmetros configuráveis do projeto relativos ao processo de treino do modelo.

No Anexo A estão contempladas duas tabelas, a primeira (Tabela 32) com as descrições de cada parâmetro, disponíveis em [59], e a segunda (Tabela 33) com o racional associado ao valor definido para cada um destes parâmetros. Relativamente aos valores definidos, existem casos em que se optou pelas recomendações em [59] e outros em que se considerou-se mais apropriado efetuar testes preliminares ao treino final do modelo de forma a aferir, com um conjunto de teste mais reduzido, que valores maximizam as métricas de desempenho.

Na Tabela 13 encontram-se as várias configurações de parâmetros de treino consideradas nos referidos testes preliminares.

Para efeitos de comparação, serão isolados todos os parâmetros à exceção daquele que se pretende aferir o impacto no desempenho. Refira-se que o conjunto de testes preliminares considerados não são todos os possíveis, mas sim os considerados mais relevantes, dado ser impraticável considerar todas as combinações possíveis. Adicionalmente, o conjunto de clipes usados neste teste não corresponde ao conjunto completo que será usado na avaliação final porque consumiria bastante mais tempo e o conjunto reduzido aparenta ser suficiente para perceber o impacto das diferentes configurações no desempenho do modelo.

Na Tabela 14 apresenta-se a distribuição das formas/posturas de desempenhar os gestos relativa ao conjunto de clipes para testes preliminares. Adicionalmente, para cada clipe indica-se tipo de guião associado. Para efeitos de teste preliminar, considerou-se suficiente um conjunto composto por dois clipes de um sujeito (Guião A) e três clipes de dois sujeitos onde dois destes pertencem ao Guião C e o restante ao Guião B.

Configuração de Parâmetros	***Número de classificadores fracos em tempo de execução***	***Peso dos Falsos Positivos durante filtragem***	***Ignorar parte inferior do corpo***
1	0 (ilimitado)	50%	Verdadeiro
2	0 (ilimitado)	50%	Falso
3	1000	50%	Falso
4	1000	50%	Verdadeiro
5	0 (ilimitado)	75%	Falso
6	0 (ilimitado)	25%	Falso
Modelo Final	**0 (ilimitado)**	**50%**	**Falso**

Tabela 13 - Configurações testadas relativas aos parâmetros de treino.

Id clipe	**Distribuição de gestos por classe**					**Tipo de Guião**	**Total**
	P1	***P2***	***P3***	***P4***	***P5***		
134711	4	6	0	8	4	A	22
133806	10	24	6	2	4	B	46
114634	2	8	6	8	0	A	24
163634	8	12	12	4	2	C	38
104657	10	14	10	2	2	C	38
Total #	34	64	34	24	12	-	168
Distribuição %	20.73%	32.93%	20.73%	14.63%	7.32%	-	100%
Desvio face à Distribuição Real	1.93%	6.45%	1.54%	(8.60%)	2.35%	-	-

Tabela 14 - Distribuição das formas de desempenhar os gestos por clipe e tipo de guião de clipes para testes preliminares.

Os testes preliminares são feitos analisando o desempenho ao nível da deteção e classificação de gestos, através da média das métricas de desempenho relativas às extensões e flexões do braço.

Com as configurações 1 a 4 pretende-se aferir os valores para os campos *Ignorar parte inferior do corpo* e *Número de classificadores fracos em tempo de execução* que maximizam o desempenho, visto corresponder a todas as combinações possíveis para estes dois parâmetros.

Para comparar o desempenho entre as várias configurações é necessário um critério objetivo, isto é, a métrica de desempenho que melhor representa a realidade do desempenho do modelo.

Dado que o modelo apresenta as classes balanceadas (uma flexão para cada extensão) e se considera que os falsos positivos e negativos têm um custo similar, considera-se que a *accuracy* será a métrica com maior peso para comparar o desempenho entre as várias configurações.

No entanto, todas as métricas referidas serão calculadas e serão objeto de análise, dado fornecerem perspetivas de análise diferentes e relevantes.

Para que a largura das tabelas não exceda os limites laterais da página, as métricas de desempenho e componentes da matriz confusão serão abreviadas – *Accuracy* (*Acc*.), *Precision* (*Pre*.), *Recall* (*Rec*.), *F1-Score* (*F1*) e número de verdadeiros/falsos positivos/negativos (#VP/VN/#FP/#FN).

Na Tabela 15, respeitante ao desempenho dos testes preliminares, apresenta-se também a matriz confusão que resulta da agregação dos resultados da deteção ao nível do gesto. A matriz confusão não é apresentada no seu formato matricial típico, mas sim em linha, de forma a que o espaço seja melhor aproveitado.

Posto isto, para os parâmetros ***Ignorar parte inferior do corpo*** e ***Número de classificadores fracos em tempo de execução***, irão ser usados os valores correspondentes à configuração que teve um melhor desempenho em *accuracy*.

Config. Parâmetros	***Acc.(%)***	***Pre.(%)***	***Rec.(%)***	***F1(%)***	***#FP***	***#FN***	***#VP***	***#VN***
1	58.93	55.36	92.26	69.20	125	13	**155**	43
2	**61.60**	**57.20**	92.26	**70.62**	**116**	13	**155**	**52**
3	50.59	50.30	**99.40**	66.80	165	**1**	167	3
4	52.38	51.23	98.81	67.48	158	2	166	10

Tabela 15 – Resultados de testes preliminares aos parâmetros Ignorar parte Inferior do corpo e Nº de classificadores fracos.

Desta forma, é possível concluir que se o modelo incluir as articulações inferiores do corpo e for usado um número ilimitado de classificadores fracos deverá obter-se um melhor desempenho.

Consequentemente, a configuração 5 irá contemplar os valores da configuração 2 para os dois parâmetros testados e será testado o parâmetro *Peso dos Falsos Positivos durante filtragem*. Dado existir uma proporção significativamente superior de falsos positivos face aos falsos negativos,

será testado um valor superior (75%) para o parâmetro Peso dos Falsos Positivos durante filtragem (acréscimo de 25% face ao valor por defeito) na configuração 5.

Config. Parâmetros	***Acc.(%)***	***Pre.(%)***	***Rec.(%)***	***F1(%)***	***#FP***	***#FN***	***#VP***	***#VN***
2	**61.60**	**57.20**	92.26	**70.61**	**116**	13	155	52
5	56.8	53.72	**98.81**	69.60	143	**2**	**166**	**25**
6	59.82	55.96	92.26	69.66	122	13	155	46

Tabela 16 - Métricas de desempenho para as configurações 2, 5 e 6.

Embora em [59] sugiram aumentar o valor deste parâmetro para reduzir os falsos positivos, verifica-se a tendência contrária, de acordo com os testes preliminares elaborados, pelo acréscimo em falsos positivos e decréscimo em falsos negativos, refletindo-se, em perdas na métrica *precision* e ganhos na métrica *recall*.

Dado o comportamento inverso ao expectável, testou-se na configuração 6 o valor 25% em alternativa aos 75%. No entanto, verifica-se um acréscimo de seis falsos positivos e o mesmo número de falsos negativos face à configuração 2. Com o ligeiro decréscimo em verdadeiros negativos, os resultados segundo a configuração 6, são piores em qualquer uma das métricas de desempenho consideradas. Visto que o critério de seleção é a métrica *accuracy*, resulta que a configuração final a usar no modelo será a segunda.

Capítulo 4 – Extração e Análise de Resultados

Com o conhecimento adquirido ao longo do desenvolvimento do modelo a questão de investigação – "*Com base em algoritmos conhecidos para classificação, é possível criar um modelo adaptado ao reconhecimento de ações humanas no contexto de uma loja?"* – deverá ser complementada com as seguintes:

- Qual o desempenho individual do detetor de gestos e de objetos?
- Quais as diferenças no desempenho dos detetores relativamente às diferentes complexidades dos cenários?
- Qual o impacto das falhas do detetor de esqueletos nos resultados do modelo?
- Quais as principais causas associadas a falhas da classificação de gestos?

Para extrair resultados do modelo são necessários dois elementos principais, o registo das deteções do modelo desenvolvido e informação de referência ou factos (*ground truth*) acerca das interações com as prateleiras dos vários clipes para teste.

Os factos foram extraídos manualmente analisando cada clipe e cada facto contém o nome do gesto, o intervalo de tramas, a etiqueta do esqueleto (A/B) e informação do objeto interagido.

Para podermos responder às questões de investigação e aprofundarmos a compreensão do comportamento do modelo, os resultados do modelo são apresentados sobre diferentes níveis de agregação.

Uma deteção correta de uma interação com a prateleira depende da correta deteção da extensão e flexão do braço assim como, para cada um destes gestos, a correta deteção do objeto.

Posto isto, considera-se apropriado analisar o desempenho ao nível da deteção das(o):

- Interações – Representa a capacidade do modelo generalizar informação sobre deteções de gestos e objetos em interações.
- Gestos – Testa o desempenho do detetor de gestos isoladamente
- Objetos no contexto de aplicação – Determina a capacidade do algoritmo desenvolvido selecionar o objeto detetado correto com base em uma ou mais deteções sucessivas.
- Objetos fora do contexto da aplicação – Testa o desempenho do detetor de objetos isoladamente.

A razão intrínseca às quatro análises referidas deve-se aos sucessivos acréscimos em detalhe face à(s) análise(s) anterior(es).

Caso ainda não seja evidente as diferenças entre as quatro perspetivas de análise referidas, nas próximas secções, dedicadas a cada uma destas análises distintas, será aprofundado o significado dos resultados obtidos.

4.1 Resultados ao nível da interação

Os resultados do modelo ao nível da interação correspondem aos resultados de alto nível visto que generalizam a informação de deteção de gestos e objetos trama-a-trama em interações associadas a intervalos de tramas.

4.1.1 Extração de resultados ao nível da deteção e classificação da interação

Uma interação será detetada sempre que, para um dado esqueleto, se detetar uma extensão seguida de uma flexão do braço. Para um dado sujeito, quando existem múltiplas deteções de extensão de braço seguidas de múltiplas flexões de braço considera-se a primeira extensão de braço e a última flexão de braço detetados para definir o intervalo de tramas da interação. Os objetos detetados na extensão e flexão de braço consideradas serão os objetos associados à interação.

A razão pela qual se seleciona apenas a primeira extensão de braço e a última flexão de braço está relacionada com uma deteção de objetos menos propensa a falha por oclusão ou confusão entre objetos. Tanto nos momentos finais de uma extensão como nos iniciais de uma flexão a mão poderá estar ocludida pela prateleira, caso se interaja, por exemplo, com os andares intermédios da prateleira, ou existem múltiplos objetos em torno da mão, cenário este típico quando se interage com o andar superior da prateleira.

Note-se que, basta que um gesto seja incorretamente detetado, ou o objeto detetado associado a este, para que o tipo de interação inferido esteja errado.

Para avaliar o desempenho relativo à classificação do tipo de interação recorre-se a uma matriz confusão onde cada classe é um dos tipos de interações já especificados (3.3.2).

Assim, resulta uma matriz confusão, como a Tabela 17, onde é contabilizado:

- **Verdadeiro Positivo**: Sempre que a interação detetada corresponde à interação factual, nomeadamente por corresponder no intervalo de tramas e no(s) produto(s) envolvido(s).
- **Verdadeiro Negativo**: Quando não existe nem é detetada interação com a prateleira.
- **Falso Positivo**: Quando a interação de determinado tipo foi detetada mas não existe.
- **Falso Negativo**: Quando a interação de determinado tipo não foi detetada mas existe.

Desta forma, será possível determinar a capacidade do modelo generalizar os vários gestos e objetos detetados em interações. As métricas extraídas serão a *precision*, *recall* e *f1-score* por tipo de interação com a prateleira. Assim, avalia-se a percentagem de deteções positivas de determinada interação que estão corretas (*precision*) e quão bem detetamos todos os casos positivos desse tipo de interação (*recall*).

4.1.2 Apresentação e análise de resultados ao nível da interação

Os resultados ao nível da interação de um modelo podem ser apresentados por clipe, sujeito (*A/B*) e tipo de cenário (*A/B/C*). No entanto, visto ser impraticável apresentar todas estas, procura-se selecionar aquela que tem maior significado à luz das questões de investigação.

Assim, apresenta-se a matriz que agrega os resultados de todos os cenários, esqueletos e gestos (Tabela 17) e as restantes serão colocadas no Apêndice B para possibilitar uma análise mais aprofundada ao leitor.

		Atual					
		Nula	**Positiva**	**Negativa c/ Novo Produto**	**Negativa s/ Novo Produto**	**Neutra**	**Total instâncias Detetadas**
Detetado	**Nula**	**551**	308	77	59	56	1051
	Positiva	83	**77**	9	7	8	184
	Negativa c/ Novo Produto	55	53	**5**	5	9	127
	Negativa s/ Novo Produto	48	9	0	**18**	5	80
	Neutra	461	210	57	65	**43**	836
	Total Instâncias Reais	643	385	82	77	99	-

Tabela 17 - Matriz confusão agregada dos cenários A, B e C e resultados ao nível da interação.

Os valores a negrito correspondem aos verdadeiros positivos por tipo de interação. O total de instâncias reais por tipo de interação equivale ao número de factos de determinada interação e, como tal, corresponde ao denominador do cálculo da métrica *recall*. Relativamente ao total de

instâncias detetadas por tipo de interação, este equivale ao somatório dos verdadeiros positivos e falsos positivos, correspondendo ao denominador do cálculo da métrica *precision*.

Desta forma, com os resultados apresentados na Tabela 17 calculam-se as métricas de desempenho apresentadas na Tabela 18.

Tipo de interação	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Nula	**52.43**	**85.69**	**65.05**
Positiva	**41.85**	20.00	**27.07**
Negativa c/ Novo Produto	3.94	6.10	4.78
Negativa s/ Novo Produto	22.50	23.38	22.93
Neutra	5.14	**43.43**	9.20

Tabela 18 - Métricas de desempenho precision, recall e f1-score por tipo de interação para os todos os cenários e esqueletos.

Através das métricas de desempenho apresentadas na Tabela 18 determina-se que a *precision* média ponderada entre os vários tipos de interação é 30.47%, *recall* 53.97% e *f1-score* 38.95%

Como o custo de um falso positivo é equivalente ao de um falso negativo e as classes não estão balanceadas, a métrica *f1-score* é considerada a que melhor representa o desempenho do modelo ao nível da deteção de interações.

Considera-se que as métricas de desempenho médias calculadas apresentam valores baixos, sugerindo uma capacidade fraca no que toca à generalização de informação de gestos individuais e produtos detetados respetivos em interações com prateleiras. Para explicar estes resultados será relevante analisar a capacidade para detetar gestos individuais e a capacidade para distinguir tipos de produtos.

Excluindo as interações nulas (quando corretamente não é identificada uma interação), as deteções de interações do tipo positivo apresentam o melhor desempenho. Contrariamente, interações negativas com novo produto e neutras apresentam o pior desempenho e, consequentemente, maior impacto na *precision* média.

O modelo está correto 41.85% das vezes que classifica determinada interação como positiva, no entanto, apenas encontra 20% das ocorrências reais deste tipo de interação.

Por outro lado, interações negativas com novo produto deverão ser as mais complexas de detetar dado necessitar que ambos os objetos associados à flexão e extensão sejam corretamente detetados, ao contrário dos restantes onde só existe objeto num dos gestos.

As interações do tipo neutro apresentam o melhor desempenho relativamente à métrica *recall* (43.43%), sendo o tipo de interação onde mais observações relevantes são detetadas. Por outro lado, apenas 5.14% das interações neutras detetadas estão realmente corretas. Consequentemente, corresponde ao tipo de interação com pior desempenho. Como tal, a análise da quantidade de falsos positivos associados à deteção de gestos e a capacidade para deteção de mãos vazias durante a extensão e flexão de braço poderá explicar esses resultados.

Para compreender as principais tendências do modelo apresenta-se o Gráfico 1 onde se cruza para cada tipo de interação detetado as interações reais correspondentes para esse intervalo de tramas e sujeito em questão (*ground truth*).

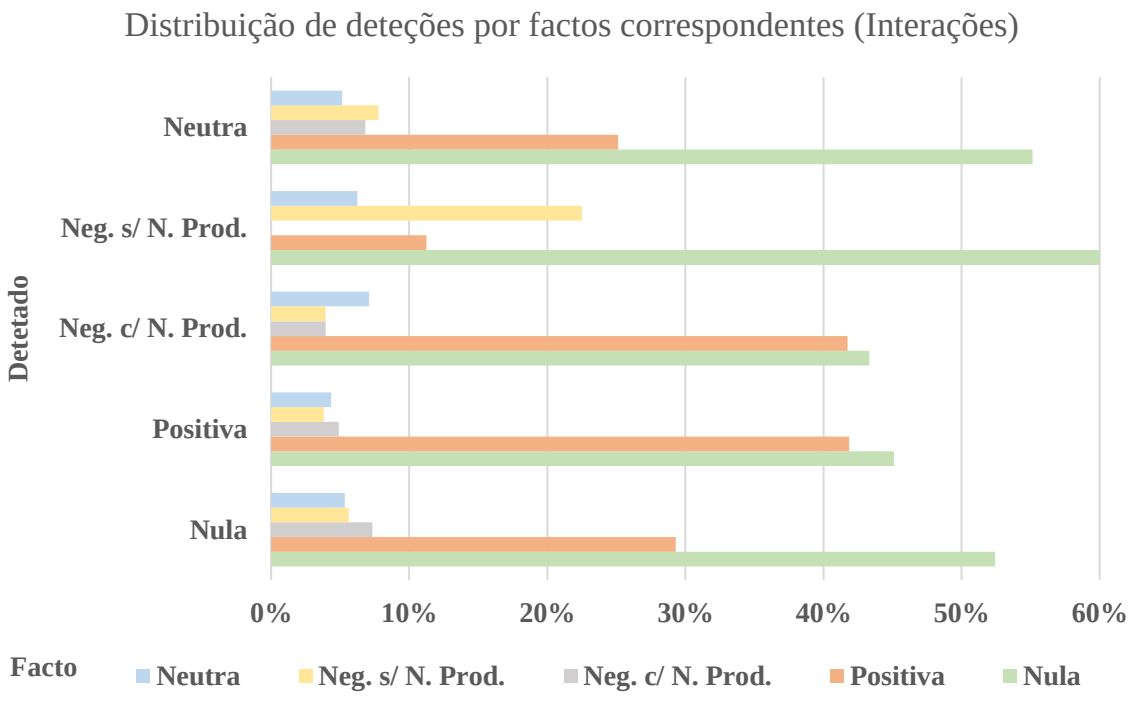


Gráfico 1 - Distribuição de deteções por factos de interações.

Verifica-se por este meio que entre 40% a 60% das vezes que é detetada qualquer interação esta corresponde a um falso positivo detetado durante um intervalo de tempo em que nenhuma interação ocorreu (i.e., associado ao facto Interação Nula). Como tal, para estes casos não se pondera uma hipotética confusão por parte do modelo em distinguir o tipo de interação com base

nos produtos detetados. Esta proporção significativa de falsos positivos está correlacionada unicamente com falsos positivos na deteção e classificação de gestos.

Adicionalmente, a confusão existente entre interações mais significativa seguinte ocorre entre Interações Positivas factuais e os restantes tipos detetados, sugestiva de uma fraca capacidade para detetar objetos embora os gestos inerentes tenham sido corretamente reconhecidos.

As restantes confusões entre tipos de interação apresentam uma proporção individual inferior a 10%, próximas, em média, dos 5%, no entanto, somadas correspondem a cerca de 10 a 20% das deteções erradas destas.

Nos dois casos anteriores, um ou ambos os objetos detetados para os gestos estão errados e levaram à determinação do tipo de interação incorreto. A título de exemplo, uma interação neutra é confundida com interação positiva quando ocorre uma deteção falsa de objeto durante a flexão do braço.

Para compreendermos o impacto da complexidade associada aos diferentes guiões no desempenho do modelo, nos Gráficos 2 a 4 compara-se as várias métricas de desempenho por tipo de interação entre guiões.

As matrizes confusão e métricas de desempenho por guião que deram origem aos Gráficos 2 a 4 estão presentes no Apêndice B (Tabelas 36 a 41).

Verifica-se que o Guião A prevalece em desempenho para interações dos tipos Nula, Negativa Sem novo Produto e Neutra, não existindo diferenças superiores a 10% em *f1-score* comparativamente aos outros guiões/cenários testados. Por outro lado, para as interações Negativas com Novo Produto, todos os guiões apresentam um desempenho baixo e próximo entre si.

Relativamente ao reconhecimento de interações Positivas, o Guião C lidera em desempenho, posicionando-se cerca de 10% acima do Guião B e 3% acima do A.

Fazendo a média ponderada do *f1-score* dos tipos de interação por guião (Tabela 19), o Guião A apresenta um *f1-score* médio de 43.02%, seguindo-se o Guião B e C com desempenhos próximos (respetivamente, 37.99% e 37.26%).

No que toca à *precision* verificam-se diferenças similares (Tabela 19). No entanto, a métrica *recall* já apresenta diferenças superiores entre guiões sendo que o Guião A capta mais de metade das interações relevantes (61.04%), seguindo-se o B com 52.06% e o C com 51.42%.

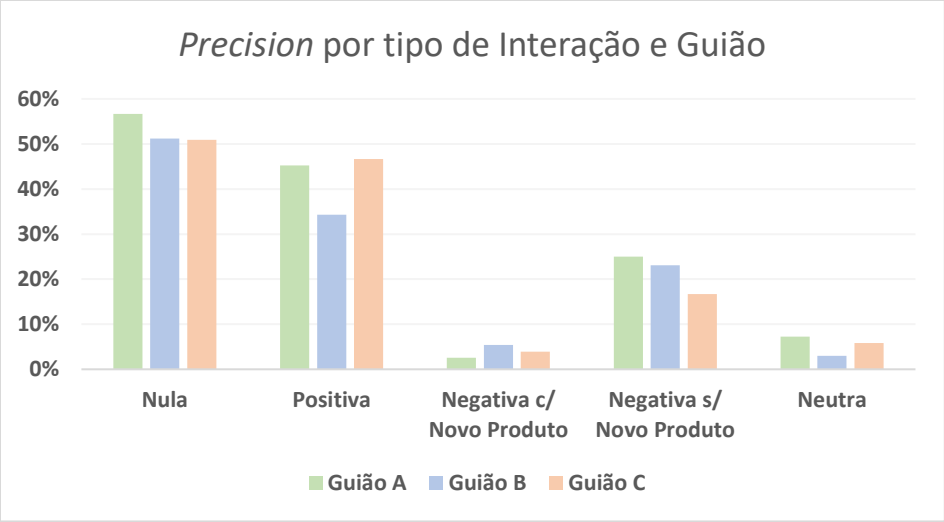


Gráfico 2 - Comparação da métrica precision entre os vários guiões por tipo de interação.

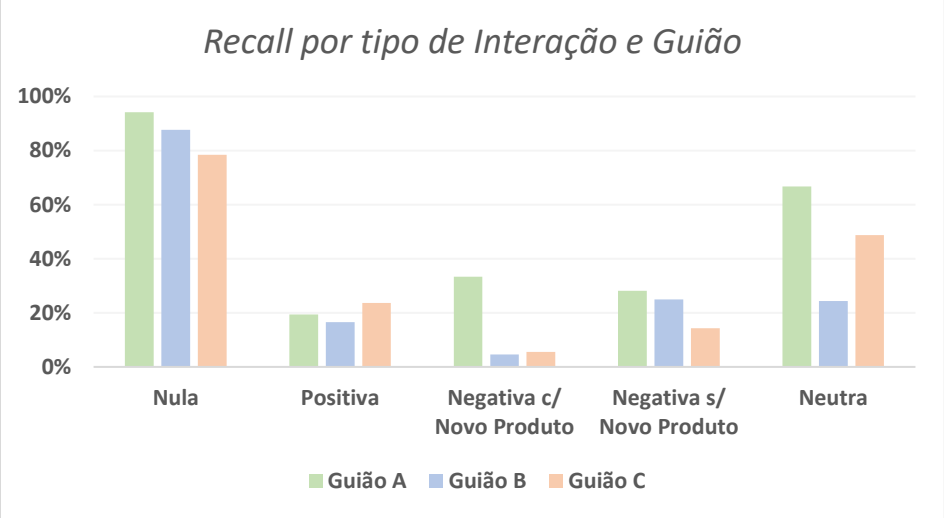


Gráfico 3 - Comparação da métrica recall entre os vários guiões por tipo de interação.

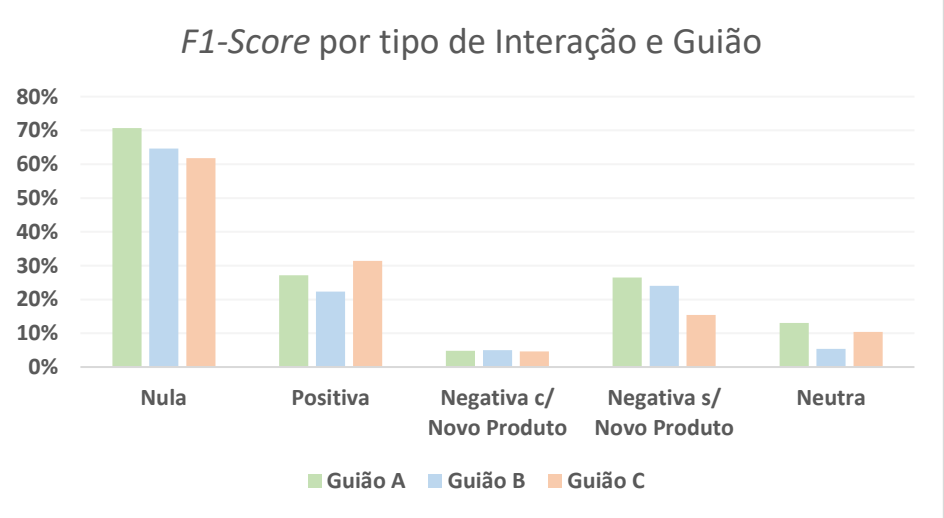


Gráfico 4 - Comparação da métrica f1-score entre os vários guiões por tipo de interação.

Tipo de Guião	*AVG(Precision (%))*	*AVG(Recall (%))*	*AVG(F1-Score (%))*
A	**33.22**	**61.04**	**43.02**
B	29.91	52.06	37.99
C	29.21	51.42	37.26

Tabela 19 - Média de métricas de desempenho por interação para cada tipo de guião.

Desta forma, é possível afirmar que cenário A está mais vezes associado a melhores resultados, embora não existam diferenças significativas.

Visto que o Guião A apenas contem um sujeito e tem resultados ligeiramente melhores, importa verificar se existem diferenças em desempenho no reconhecimento de interações entre esqueletos diferentes.

Posto isto, nas Tabelas 20 e 21 apresentam-se as métricas de desempenho agregadas dos vários cenários relativas, respetivamente, aos esqueletos A e B.

As referidas tabelas foram calculadas de acordo com as matrizes confusão presentes no Apêndice B (Tabelas 42 e 44).

Tipo de interação	Precision (%)	Recall (%)	F1-Score (%)
Nula	**53.06**	**86.45**	**65.76**
Positiva	**45.71**	20.17	**27.99**
Negativa c/ Novo Produto	1.56	5.88	2.47
Negativa s/ Novo Produto	22.50	23.38	22.93
Neutra	4.95	**42.37**	8.87

Tabela 20 - Métricas de desempenho por tipo de interação para os resultados agregados do esqueleto A.

Tipo de interação	Precision (%)	Recall (%)	F1-Score (%)
Nula	**51.45**	**84.52**	**63.96**
Positiva	**36.71**	19.73	**25.66**
Negativa c/ Novo Produto	6.35	6.15	6.25
Negativa s/ Novo Produto	-	-	-
Neutra	5.44	**45.00**	9.70

Tabela 21 - Métricas de desempenho por tipo de interação para os resultados agregados do esqueleto B.

Verifica-se que ambos os esqueletos apresentam desempenhos próximos para qualquer tipo de interação.

A maior diferença verifica-se para interações do tipo Negativo Com Novo Produto, onde o esqueleto B tem um desempenho superior. Tal pode acontecer por existir um menor número de observações relevantes para detetar associadas ao esqueleto A (17) comparativamente ao esqueleto B (65).

Visto que não existem amostras de interações do tipo Negativo Sem Novo Produto para o esqueleto B, não é possível comparar o desempenho entre esqueletos para este tipo de interação.

Em geral, verifica-se uma fraca capacidade na deteção de todas as interações relevantes, sobretudo para as interações do tipo Negativo com Novo Produto com *recall* de 6.10%, e uma baixa proporção de deteções corretas dentro do conjunto de deteções, com destaque para o tipo Neutro e, novamente, interações Negativas com Novo Produto.

Como foi analisado, os maus resultados podem estar associados falsas interações detetadas, fruto de falsos positivos na deteção de gestos, e/ou, quando os gestos que compõe a interação são corretamente detetados, o tipo de interação inferido está errado fruto de falhas na deteção e classificação do objeto intrínseco.

4.2 Resultados ao nível da deteção e classificação de gestos

No que toca aos resultados ao nível da deteção de gestos, detalha-se em primeiro lugar o processo de treino do detetor/classificador desenvolvido (4.2.1), seguido da lógica implementada para extrair resultados com base nas deteções e factos (4.2.2) e, por fim, o desempenho sobre diversas perspetivas de análise (4.2.3).

4.2.1 Fase de treino

Para treinar o modelo de deteção de gestos recorre-se à função *Build* presente no *VGB*, obtendo-se o detetor de gestos acompanhado de resultados do seu desempenho determinados com base nos clipes para treino.

Foram usados 31 clipes para treino, que totalizam 394 exemplos positivos de extensões de braço e 384 exemplos de flexões de braço.

Dos 8664 classificadores fracos gerados, as nove *features* que mais contribuíram para a classificação da extensão do braço são: velocidades no eixo horizontal da espinha do ombro, cotovelo esquerdo e direito, mão esquerda e direita, ângulos entre o centro da espinha, cabeça e pulso esquerdo e torque muscular do pulso direito no eixo z (profundidade).

Para associar o nome da articulação referida à articulação correspondente apresenta-se no Apêndice C (Figura 25) o esqueleto captado e na Tabela 46 a correspondência entre o identificador da articulação e respetivo nome.

No que toca à classificação de flexões do braço, dos 88547 classificadores fracos gerados, as nove *features* com mais impacto são: velocidades da cabeça, espinha do ombro, mão esquerda, ombro esquerdo e pulso direito no eixo horizontal, velocidades da cabeça no eixo vertical, diferença em posição do pulso esquerdo e ponta da mão esquerda relativamente ao eixo vertical e, por fim, velocidades dos ângulos entre o centro da espinha, cabeça e cotovelo esquerdo.

Nível dos resultados	**Métrica de resultado *VGB***	**Gesto**	
		Extensão	**Flexão**
Trama	*Accuracy* VP (%)	97.42	97.77
	Erro FP (%)	2.51	2.42
Gesto	Accuracy VP (%)	100	100
	Erro FP (%)	111.27	126.52

Tabela 22 – Métricas Accuracy (verdadeiros positivos) e erro (falsos positivos) testados com conjunto de dados para treino.

Na Tabela 22 apresenta-se as métricas de desempenho apresentadas pelo *Visual Gesture Builder* calculadas com base no conjunto de dados para treino.

Como o desempenho será aferido com base no conjunto de clipes para teste, os resultados obtidos baseados no conjunto de treino foram usados como forma de perceber rapidamente se existem problemas relativamente aos dados de treino (como a coerência na etiquetação de exemplos), analisando o erro.

As métricas obtidas ao nível da trama são consideradas boas. No entanto, a reduzida proporção de tramas que são falsos positivos dá origem uma proporção grande de falsos positivos para ambos os gestos. Dos inúmeros classificadores criados, constata-se que a configuração final do modelo apresenta métricas de desempenho, tanto ao nível da trama como do gesto, melhores que as restantes configurações testadas. No entanto esse detalhe é considerado supérfluo e como tal é omitido.

Posto isto, importa agora avaliar o desempenho do detetor e classificador de gestos com base no conjunto de dados para teste e no artefacto desenvolvido para extração de resultados, explicado em seguida.

4.2.2 Extração de resultados ao nível do gesto

As métricas para avaliação de desempenho do modelo ao nível do gesto serão calculadas com base na matriz confusão resultante, pelo que importa detalhar como vai ser considerada a contabilização de verdadeiros positivos/negativos e falsos positivos/negativos.

Para o conjunto de factos de determinado gesto que obedeçam às condições abaixo contabiliza-se uma ocorrência de verdadeiro positivo ou falso negativo:

- **Verdadeiro Positivo**: Se existir pelo menos uma deteção durante o intervalo de tramas real (*ground truth*) e se houver correspondência na etiqueta do esqueleto e nome do gesto.
- **Falso Negativo**: Se não existirem deteções durante o intervalo de tramas real (*ground truth*) onde a etiqueta do esqueleto e nome do gesto correspondem.

Para contabilizar falsos positivos e verdadeiros negativos analisam-se todos os intervalos de tramas onde o gesto não ocorre (*ground truth*). Para cada intervalo, contabiliza-se um falso positivo ou verdadeiro negativo:

- **Falso Positivo**: Se existir pelo menos uma deteção do mesmo gesto para uma determinada etiqueta de esqueleto.
- **Verdadeiro Negativo**: Se não existirem deteções para esse esqueleto e gesto.

A informação obtida complementa os resultados em 4.1 nos seguintes aspetos:

- Análise separada dos gestos extensão e flexão, nomeadamente o seu impacto nos resultados generalizados;
- Determinar, sempre que possível, as causas de erros na classificação de cada gesto, como a falha na deteção de esqueleto e/ou incorreta atribuição de etiqueta de esqueleto durante a deteção.

Como foi referido, para que seja possível determinar as causas de erros na deteção de gestos regista-se também dados de controlo relativos ao estado do(s) esqueleto(s) para cada trama.

A partir dessa informação, a cada esqueleto de uma dada trama associa-se uma de três potenciais causas de erro na classificação:

- **Foco de esqueleto errado (*C1*):** Associado ao esqueleto B que é confundido com A visto que o esqueleto A não é detetado nessa trama e, consequentemente, as deteções de B são registadas como se fossem desempenhadas pelo sujeito A.
- **Esqueleto Não Detetado (*C2*):** Associado a qualquer um dos esqueletos que numa dada trama não seja detetado.

- **Esqueleto de fraca qualidade ou falha do classificador (*C3*):** Quando o esqueleto é detetado e não é confundido com outro a única potencial causa de erro será a fraca qualidade do esqueleto (esqueletos fundidos num ou localização de articulações errada) e/ou a incapacidade do classificador em detetar o gesto.

As referidas causas de falhas na classificação serão particularmente úteis para distinguir falsos negativos, no sentido de apurar se o gesto não foi reconhecido porque o classificador ou o esqueleto não são suficientemente bons ou se é impossível reconhecer o gesto porque o sujeito não foi detetado ou foi identificado incorretamente.

Para implementar a associação entre esqueletos e causas de erro possíveis de uma dada deteção considerou-se, no caso dos clipes de dois sujeitos:

1. Quando **ambos os esqueletos são detetados** o esqueleto com profundidade menor será o esqueleto A e o restante o B. Em seguida associa-se ambos à causa *C3* (*Esqueleto de fraca qualidade ou falha do classificador*).
2. Quando **apenas um esqueleto é detetado**: Se em tramas anteriores ambos os esqueletos foram captados, o atual único esqueleto captado terá a mesma etiqueta que o esqueleto que apresente menor diferença em profundidade face ao atual, relativamente ao momento mais recente em que ambos foram captados. Caso ambos ainda não tenham sido detetados optou-se por uma simplificação baseada na profundidade. Se o esqueleto está a mais de 2,3 metros então será o B, caso contrário será o A. Tal deve ao facto de, tipicamente, o sujeito A se encontrar entre 1,7 a 2,2 metros do sensor e o sujeito B entre 2,3 a 3,5 metros.

Seguidamente, para os casos em que se deteta apenas um sujeito, em função da etiqueta atribuída ao esqueleto definem-se as potenciais causas de falha na classificação:

a. Se for etiqueta A então associa-se o esqueleto à potencial causa de falha de classificação *C3* e ao esqueleto B a causa *C2* (esqueleto não detetado).
b. Se for etiqueta B então associa-se *C1* (Foco de esqueleto errado) ao B e *C2* ao A.

No caso dos clipes de um sujeito, quando este é detetado associa-se à potencial causa de falha na classificação *C3*, caso contrário associa-se à causa *C2*.

Em seguida, atribuem-se as causas de falha na deteção de gestos com base na informação de controlo associada a deteções e os factos existentes.

1. Se a deteção pertencer a um esqueleto não detetado ou incorretamente etiquetado (*C1* ou *C2*), esta foi mero acaso e deverá ser diferenciada.

2. Se a deteção pertencer a um esqueleto detetado (*C3*), o resultado da classificação está efetivamente correto, seja um verdadeiro positivo ou falso positivo/negativo.

Desta forma, separam-se Verdadeiros Positivos e Falsos Positivos registados de forma incorreta daqueles que efetivamente foram bem identificados. Quanto aos Falsos Negativos, separam-se os gestos não detetados devido a um mau desempenho do detetor ou fraca qualidade do esqueleto daqueles que eram impossíveis detetar visto que o esqueleto não foi reconhecido.

4.2.3 Resultados ao nível da deteção e classificação do gesto

Dado o nível de detalhe dos resultados obtidos resultam múltiplas matrizes confusão que foram colocadas no Apêndice D em páginas horizontais.

Neste apêndice apresentam-se resultados globais, por guião/cenário, por cada tipo de esqueleto e por gesto.

Para efeitos de análise, criou-se a Tabela 23 com base no Apêndice D apresentando-se as métricas de desempenho resultantes.

Resultados	***Accuracy (%)***	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Globais	61.31	57.25	89.34	69.78
Guião A	65.26	59.92	92.21	72.63
Guião B	62.58	58.19	89.37	70.48
Guião C	57.63	54.77	87.55	67.39
Esq. A	63.69	58.65	92.87	71.89
Esq. B	57.65	55.01	83.92	66.46
Extensão	61.14	57.20	88.48	69.48
Flexão	61.49	57.30	90.22	70.08

Tabela 23 – Métricas de desempenho para deteção e classificação ao nível do gesto por perspetiva de análise.

Tal como referido em 3.4, considera-se que a métrica *accuracy* é a que melhor representa o desempenho do modelo quando medido ao nível do gesto. Logicamente, é essencial considerar as restantes para uma análise completa, no entanto, para efeitos de comparação, será dada maior importância à *accuracy*.

É agora mais evidente o impacto da complexidade de cada guião nos resultados do detetor de gestos, dado verificar-se decréscimos em *accuracy* de 2.6 a 5% sempre que a complexidade do

guião aumenta. Adicionalmente, o esqueleto A apresenta melhor desempenho comparativamente ao B, com uma diferença em *accuracy* de 6%.

Por outro lado, as diferenças em desempenho entre o tipo de gesto são pouco evidentes, onde a extensão apresenta uma *accuracy* de 61.14% e a flexão 61.49%, sugerindo não existirem diferentes graus de dificuldade no reconhecimento dos dois tipos de gestos considerados.

Em geral, o modelo é sensível à deteção de ambos os gestos tal como sugere os valores altos em *recall* (89.34%). No entanto, dos gestos detetados, apenas 57.32% (*precision*) estão corretos.

O desequilíbrio entre *precision* e *recall* ao nível da deteção dos gestos pode explicar os resultados fracos das deteções ao nível das interações. Nomeadamente devido à quantidade de falsos positivos (explícitos nas matrizes apresentadas no Apêndice D), que condicionam a capacidade de generalizar um conjunto de extensões seguidas de flexões numa única interação.

Para explicar os resultados das deteções ao nível das interações menos bons existem ainda dois elementos fundamentais a avaliar:

- O impacto de falhas na deteção do esqueleto e respetivas causas (*C1*, *C2* e *C3*)
- A capacidade do modelo desenvolvido detetar um objeto corretamente.

Posto isto, na Tabela 24 apresenta-se a contabilização de potenciais causas de falha na deteção por componente da matriz confusão. Estes resultados são apresentados sobre diferentes níveis de agregação à semelhança dos resultados anteriormente apresentados, de forma poderem ser relacionados.

Usa-se o termo "potencial causa de falha" porque todos os gestos são associados a uma potencial causa de falha (*C1*, *C2* ou *C3*). Quando associado a *C3* (*Esqueleto de fraca qualidade ou falha do classificador***)** a deteção está associada a um esqueleto corretamente identificado. Logo, se se tratar de um falso positivo/negativo apenas se pode concluir o classificador errou ou o esqueleto não tinha qualidade suficiente, embora seja captado. Quando se trata de Verdadeiros Positivos/Negativos os casos *C3* podem ser ignorados porque a deteção está correta. No entanto, um Verdadeiro Positivo/Negativo associado a *C1* ou *C2* significa que o Verdadeiro Positivo ocorreu por mero acaso ou "não intencional".

Através da sua análise da Tabela 24, verifica-se que falhas na deteção de esqueleto explicam uma parte significativa dos falsos negativos.

Existem 21 gestos (9 extensões e 12 flexões de braço) que não foram detetados porque o esqueleto não estava a ser reconhecido e 3 gestos (2 extensões e 1 flexão) não detetados visto que o sistema de captação de esqueletos confundiu o sujeito B com o A.

Resulta uma proporção de falsos negativos causados por falhas na deteção do esqueleto de 17.39%.

Importa realçar também que não se verifica uma correlação evidente entre o cenário e a quantidade de casos *C2*, por existir um total de situações *C2* associadas a Falsos Negativos/Positivos em quantidades similares ao longo dos guiões. Por outro lado, apenas ocorrem casos *C1* no Guião C (Esqueleto B).

	VP			***VN***			***FP***			***FN***		
Resultados	***C1***	***C2***	***C3***	***C1***	***C2***	***C3***	***C1***	***C2***	***C3***	***C1***	***C2***	***C3***
Globais	0	**3**	1154	0	0	431	0	**3**	861	**3**	**21**	114
Guião A	0	0	284	0	0	118	0	0	190	0	**9**	15
Guião B	0	0	437	0	0	175	0	**1**	313	0	**9**	43
Guião C	0	**3**	433	0	0	138	0	**2**	358	**3**	**3**	56
Esq. A	0	**3**	726	0	0	271	0	**3**	510	0	**11**	45
Esq. B	0	0	428	0	0	160	0	0	350	**3**	**10**	69
Extensão	0	**2**	574	0	0	220	0	**2**	429	**2**	**9**	64
Flexão	0	**1**	580	0	0	211	0	**1**	432	**1**	**12**	50

Tabela 24 - Contabilização de causas de falha na deteção de gestos por componente da matriz confusão.

Na Tabela 25 calcula-se o ganho percentual em desempenho para cada métrica calculada anteriormente (Tabela 23) quando estas são calculadas considerando apenas os casos *C3*.

Assim, obtém-se o desempenho dos modelos sobre diversas perspetivas de análise (por cenário/esqueleto/gesto) isentos de casos em que o sistema de captação de esqueletos falhou na correta identificação do(s) esqueleto(s). Desta forma, desagrega-se a capacidade do modelo identificar um gesto da dependência que estabelece com o sistema de captação de esqueletos.

Para visualizar as métricas de desempenho resultantes considerando apenas casos C3, em alternativa ao ganho percentual, apresentam-se as mesmas no Apêndice E.

	Accuracy (%)	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Globais	0.60	0.02	1.67	0.52
Guião A	0.97	0.00	2.77	0.85
Guião B	0.64	0.08	1.67	0.58
Guião C	0.34	-0.03	1.00	0.27
Esq. A	0.55	0.09	1.29	0.46
Esq. B	0.74	0.00	2.20	0.68
Extensão	0.55	0.03	1.49	0.48
Flexão	0.65	0.01	1.77	0.54

Tabela 25 – Impacto nas métricas de desempenho ao nível do gesto quando se exclui deteções associadas a C1 e C2.

Verifica-se que os ganhos em *accuracy* não excedem os 0.97%, com um impacto médio de 0.60%.

Para a métrica *precision* os resultados permanecem praticamente os mesmos porque existem as mesmas quantidades de situações *C1* e *C2* associadas a Verdadeiros Positivos e Falsos Positivos.

Contrariamente, a métrica *recall* apresenta impactos superiores, sobretudo devido à maior proporção de *C1* e *C2* associada a falsos negativos quando comparada com Verdadeiros Positivos.

Consequentemente, os ganhos em *f1-score* são resultado das melhorias em *recall.*

Apesar de tudo, considera-se que as falhas do detetor de esqueletos não têm um impacto significativo nos resultados, sendo a principal causa associada aos maus resultados ao nível da interação consequente dos falsos positivos associados à deteção de gestos.

Resta compreender até que ponto a deteção ao nível dos objetos está também a condicionar esta incapacidade de generalização em interações, pelo que será alvo de análise em seguida.

4.3 Resultados ao nível da deteção e classificação de objetos

4.3.1 Fase de treino

O treino do modelo para detetar os objetos pretendidos é feito com um conjunto imagens etiquetadas, exemplificadas no Apêndice F. Estas imagens foram recolhidas após a integração do detetor de gestos na aplicação desenvolvida. Isto é, sempre que foi detetado um gesto do sujeito

guardou-se localmente uma imagem recortada em torno das mãos desse sujeito. Após a reprodução do conjunto de clipes para treino obtêm-se inúmeras imagens que serão manualmente etiquetadas e selecionadas para treino.

A etiquetação destas imagens consistiu na delimitação das fronteiras do objeto e foi feita recorrendo à ferramenta *Yolo Mark*, disponível em [60].

Nesta ferramenta delimitam-se as fronteiras do objeto criando uma caixa em torno do objeto. Em resultado disso, por cada imagem etiquetada é gerado um ficheiro de texto contendo a informação referente a cada objeto identificado nessa imagem (uma linha de texto por cada objeto). A informação associada a cada objeto irá conter um identificador, as coordenadas do seu centro, a largura e o comprimento da caixa que delimita esse objeto.

Na Tabela 26 apresenta-se o total de imagens consideradas por produto para treino.

Depois de etiquetados todos os objetos para treino/teste inicia-se o treino do modelo com um ficheiro de pesos pré-treinados e a cada 100 iterações é guardado um *backup* do estado da rede, isto é, os valores dos vários pesos que foram ajustados durante o treino.

O modelo foi treinado durante aproximadamente 44 horas e 30 minutos com uma unidade de processamento gráfico *Nvidia GeForce GTX 1050 Ti OC 4GB GDDR5* e processador *AMD Ryzen 5 1600 6-Core 3.2GHz.* No que toca ao *hardware*, o mais relevante para o treino e teste do modelo é o processamento gráfico visto que, em geral, é substancialmente mais rápido treinar/testar uma rede neuronal via processamento gráfico do que com um processador [61].

Importa agora detalhar os critérios utilizados para parar o treino do modelo que estão de acordo com as recomendações em [58]. Regra geral, é suficiente fazer 2000 iterações por classe (objeto) pelo que resultava num total de 8000 iterações (4 x 2000). No entanto, para obter melhores resultados e evitar o sobreajuste do modelo, devem ser testados os vários estados da rede guardados com um conjunto de imagens para teste diferente do treino.

As métricas consideradas para determinar o desempenho do modelo são a *IoU* e o *mAP*, explicadas em 2.4.1, sendo selecionado o ficheiro de pesos que apresente maior média entre estas.

No Gráfico 5 apresenta-se a evolução da média entre as métricas *mAP* e *IoU* ao longo das várias iterações feitas durante o treino (Apêndice G). Resulta que o ficheiro de pesos com melhor

desempenho corresponde à iteração número 9400 (78.27%), apresentando um *mAP* de 81.53% e *IoU* de 75%.

Tipo de Produto (Classes)	**Elementos da classe**	**Nº Exemplos**
Caixas de cereais	Nestum	108
	Kellogg's	105
	Subtotal	**213**
Livros	Capa vermelha, branca e cinza	134
	Capa azul e cinza	103
	Capa azul	57
	Capa vermelha, azul e cinza	131
	Capa preta	118
	Subtotal	**543**
Embalagens de Leite	-	**125**
Garrafas de água	-	**115**
Total imagens para treino		**996**

Tabela 26 - Distribuição de exemplos de objetos usados para treino da rede Darknet Yolo v3.

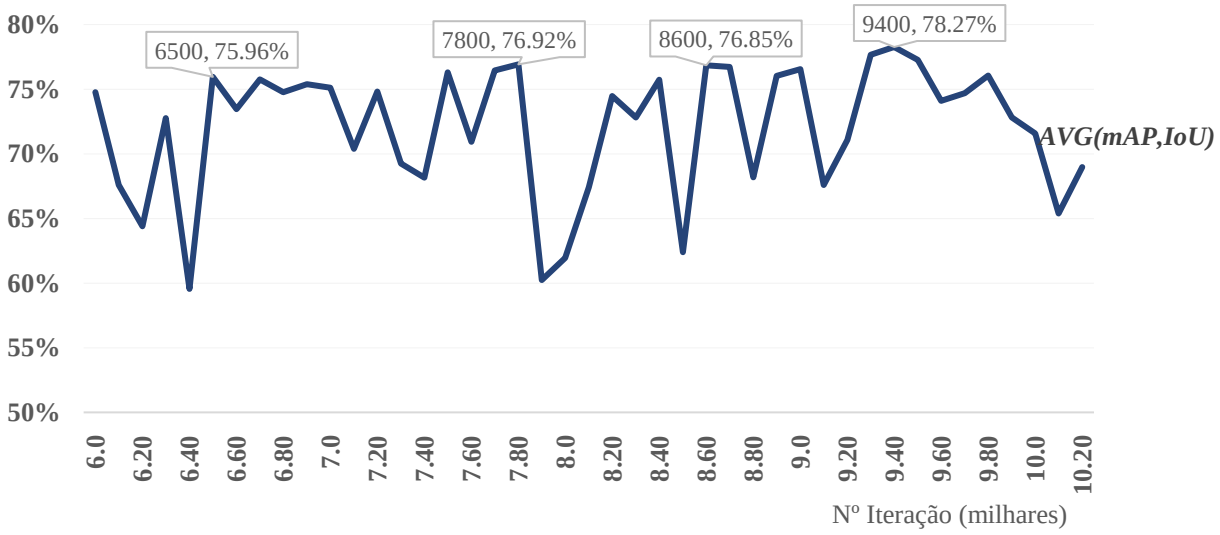


Gráfico 5 - Evolução do desempenho (média(mAP,IoU)) por número de iteração de treino.

4.3.2 Resultados da deteção/classificação de objetos no contexto de aplicação

Relativamente à avaliação do detetor de objetos, importa recordar o subcapítulo 2.4.1 onde é abordado o estado-de-arte na avaliação de modelos para reconhecimento de objetos.

As métricas tipicamente usadas são a *precision*, *recall*, *Intersection over Union* (*IoU*) e *mAP*.

Visto que as métricas *IoU* e *mAP* apenas podem ser calculadas a posteriori porque as fórmulas associadas se baseiam em factos apenas determináveis após a tarefa de deteção e classificação, estas métricas são avaliadas na secção 4.3.3.

Dado se tratar de um problema de classificação de múltiplas classes, a matriz confusão resultante assemelha-se às apresentadas na secção 4.1.2, relativas aos resultados do modelo ao nível da interação. Em ambas as situações contabiliza-se o total de instâncias reais e detetadas por classe, que correspondem, respetivamente, aos denominadores das métricas *precision* (TP + FP) e *recall* (TP + FN).

As matrizes de classificação global, por guião, esqueleto e gesto estão presentes no Apêndice H, acompanhadas de uma tabela com as métricas desempenho por tipo de produto.

Com base nos resultados obtidos, a Tabela 27 agrega as métricas de desempenho resultantes relativas à deteção de objetos no contexto de aplicação. Para que a tabela não exceda os limites abrevia-se *precision (%)* em *P*, *recall (%)* em *R* e *f1-score (%)* em *F*.

	Água			**Livro**			**Caixa de Cereais**			**Leite**			**Mão Vazia**		
	P	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
Globais	81	15	**26**	73	54	**62**	69	40	**51**	88	24	**37**	65	94	**77**
Guião A	57	14	23	90	78	**84**	75	41	**53**	85	37	**51**	67	93	**78**
Guião B	86	12	21	59	47	52	62	40	48	83	24	37	64	93	76
Guião C	92	19	**32**	83	48	60	73	40	52	100	15	26	63	97	76
Esq. A	73	11	19	78	56	**65**	75	45	**57**	88	20	32	65	95	**77**
Esq. B	87	21	**33**	67	50	57	54	29	37	87	32	**46**	63	93	75
Extensão	50	19	**28**	68	54	60	52	29	37	80	27	**40**	83	95	**88**
Flexão	100	15	25	77	53	**63**	76	45	**56**	91	23	36	43	93	59

Tabela 27 – Resultados percentuais da deteção de objetos sobre diversas perspetivas de análise por tipo de produto.

Para analisar o desempenho de cada modelo considera-se apropriado fazer comparações entre os cenários/guiões (A, B e C), esqueletos (A e B), gestos (flexão e extensão) e tipos de produtos.

Como a métrica *f1-score* consiste na média harmónica entre a *precision* e *recall* esta será a melhor representação do desempenho dos resultados de um dado tipo de produto, não desprezando a análise individual das restantes métricas. Por essa razão, apresenta-se a negrito o *f1-score* do tipo de produto que tem melhor desempenho dentro dos vários grupos de comparação previamente enunciados (i.e., entre modelos, cenários, esqueletos e gestos).

O Cenário/Guião A apresenta, em geral, melhores resultados com um *f1-score* médio ponderado de 70.07%, seguindo-se o Cenário C com 66.28% e o Cenário B com 64.30%.

No que toca aos esqueletos, o esqueleto A apresenta um *f1-score* médio de 67.49% e o esqueleto B 64.42%, pelo que se conclui que o detetor de objetos apresenta um desempenho semelhante entre esqueletos.

Relativamente aos gestos, a extensão tem um *f1-score* médio ponderado de 79.86% e a flexão 53.18%.Tal deve-se ao facto de se detetarem corretamente mais instâncias de 'Mão Vazia' em casos de extensão de braço, com uma *precision* aproximada de 83%, do que flexão de braço, com uma *precision* de 43%. Isto explica parte dos maus resultados obtidos na deteção de interações visto existir uma proporção elevada de falsos positivos associados à errada deteção de flexões de braço com 'Mão Vazia'.

Por sua vez, verifica-se a tendência contrária no caso dos restantes tipos de produto onde existe uma maior precisão na deteção desses produtos quando se trata do gesto flexão do braço. Tais dificuldades constituem-se a principal causa dos baixos resultados ao nível do reconhecimento de Interações Negativas, onde há extensões de braço com produtos.

Analisando as diferenças em desempenho entre tipos de produtos, a deteção da 'Mão Vazia' prevalece em desempenho qualquer outro tipo de produto, seguindo-se os livros, caixas de cereais, leite e água. Tal acontece devido aos altos valores em *recall* (94.2%) e *precision*, próxima de 65%.

De facto, não se verificam valores tão altos em *recall* para qualquer um dos outros tipos de produto devido, em grande parte, à quantidade de falsos negativos. Como tal, infere-se erradamente que a mão está vazia (Apêndice H, Tabela 57).

No Apêndice H apresentam-se todas as matrizes confusão que deram origem à Tabela 27. A partir dessas matrizes confusão criou-se também a Tabela 28 onde se apresenta a proporção de falsas deteção de cada tipo de produto quando o facto corresponde à 'Mão Vazia' face ao total de deteções. A título de exemplo, aproximadamente 20% das deteções do objeto água ocorrem quando o sujeito tem a mão vazia.

Embora se verifique uma *precision* alta para a generalidade dos tipos de produto, esta métrica é principalmente afetada pela proporção de falsas deteções de qualquer um dos tipos de produto quando se trata de 'Mão Vazia' (Tabela 28).

Proporção de falsas deteções por Tipo de Produto associadas ao facto 'Mão Vazia' (%)	
Água	15.38
Livro	14.63
Caixa de Cereais	**19.78**
Leite	9.38

Tabela 28 – Proporção de falsas deteção por tipo de produto associadas ao facto 'Mão Vazia'.

No entanto, a elevada proporção de falsos negativos é a maior condicionante ao reconhecimento de objetos. Na Tabela 29 apresenta-se a proporção de deteções de 'Mão Vazia' quando o facto correspondente é um dos quatro tipos de produto.

Proporção de Deteções de 'Mão Vazia' por factos dos tipos de produtos (%)	
Água	**77.21**
Livro	44.64
Caixa de Cereais	56.69
Leite	**72.03**

Tabela 29 - Proporção de deteções de 'Mão Vazia' quando o objeto corresponde a um dos quatro tipos de produto.

Verifica-se que 77.21% das instâncias reais de objetos água não são reconhecidos e é inferido erradamente que o sujeito tem a Mão Vazia.

Conclui-se também que, embora a proporção de deteções de 'Mão Vazia' corretas seja aproximadamente 65%, o impacto dos falsos positivos é significante no *recall* dos vários tipos de produto.

Verifica-se assim que a capacidade do modelo detetar um determinado produto está limitada, principalmente por existir uma certa proporção de falsos positivos por tipo de produto, associados ao facto 'Mão Vazia' (Tabela 28), afetando a *precision*, e bastantes falsos negativos (Tabela 29) de cada tipo de produto com impacto no *recall*. Para compreender se tal se deve à incapacidade do detetor de objetos reconhecer os quatro tipos de produtos ou se isso é provocado pelo algoritmo

de seleção de imagens mais relevantes devem ser analisados os resultados do detetor de objetos isolado do contexto de aplicação.

4.3.3 Resultados ao nível da deteção/classificação de objetos isolado do contexto de aplicação

Para compreender o significado desta avaliação de desempenho isolada do contexto de aplicação, considera-se mais apropriado recorrer a um exemplo que o ilustre. Nesse sentido, suponha-se: (1) uma extensão de braço entre as tramas 10 e 30 com as mãos vazias (factos); (2) deteções de extensão de braço entre as tramas 15 e 25 onde entre as tramas 15-18 se deteta, erradamente, uma embalagem de leite e entre as tramas 19-25 se deteta, corretamente, a mão vazia.

Tendo em conta o algoritmo desenvolvido, no caso das deteções de objetos relativas a extensões de braço, é dada prioridade à primeira deteção ocorrida.

Assim, a classificação final atribuída para o objeto envolvido neste gesto será a embalagem de leite, resultando um falso negativo para a etiqueta ‘Mãos Vazias’ e falso positivo para a etiqueta ‘Embalagem de Leite’.

Este resultado permite avaliar o desempenho do modelo no seu contexto de aplicação visto que está dependente do algoritmo desenvolvido, isto é, da deteção e respetiva trama selecionada como mais relevante.

Por outro lado, imagine-se que modelo detetou a embalagem de leite entre as tramas 15 e 18 porque a imagem recortada em torno da mão apanha parte do cesto de compras no qual é visível uma embalagem de leite. Neste cenário, o modelo detetou corretamente uma embalagem de leite e, como tal, não considera que a mão está vazia.

Posto isto, cria-se a necessidade de averiguar o desempenho do detetor de objetos isolado do algoritmo de seleção de deteções mais relevantes.

Para que isso seja possível, desenvolveram-se métodos para extrair cada imagem considerada em cada deteção de gesto. Seguidamente estas imagens foram etiquetadas através da ferramenta *Yolo Mark* e testa-se o desempenho do mesmo detetor de objetos sobre estas imagens.

Para podermos avaliar o seu desempenho ir-se-á calcular as métricas *precision*, *recall*, *f1-score*, *IoU* e *mAP* com base em factos que representem os objetos que constam nas imagens.

Na Tabela 30 apresenta-se a métrica *mean average precision* por tipo de produto.

O desempenho é superior na deteção de livros, seguindo-se as caixas de cereais, leite e águas.

	mAP (%)
Água	31.43
Livro	**56.40**
Caixa de Cereais	41.22
Leite	31.37

Tabela 30 – Mean Average Precision (mAP) por tipo de produto.

Não existem resultados para 'Mão Vazia' visto que esta classe é automaticamente atribuída pelo modelo quando nenhum objeto é detetado nas mãos.

Total Deteções	**Total Factos**	***mAP (%)***	***Avg(IoU) (%)***	***Prec. (%)***	***Rec. (%)***	***F1 (%)***
530	598	40.10	58.46	81	30	43

Tabela 31 - Métricas de desempenho para deteção de objetos isoladas do contexto de aplicação.

Relativamente ao desempenho global do detetor de gestos, pode-se concluir que existe elevada precisão dentro do conjunto de predições visto que 81% das predições feitas para o conjunto de teste estão corretas. Por outro lado, encontra apenas 30% dos casos positivos no conjunto de teste, confirmando o principal problema levantado anteriormente, isto é, a fraca capacidade para detetar todas as instâncias reais. A métrica *mean Average Precision* resultante é 40.1%, significando que o modelo classifica corretamente um objeto a cada duas ou três imagens testadas, em média.

No que toca à capacidade de localização do objeto detetado apresenta um *IoU* médio entre classes de 58.46%. Tipicamente considera-se um *IoU* superior a 50% um bom resultado. O facto de uma parte significativa das deteções ser feita em imagens onde o produto está desfocado consequente do movimento do braço, reforça-se a conclusão de que o *IoU* obtido é satisfatório.

Transpondo o significado dos resultados obtidos para a análise feita em 4.3.2, sempre que a deteção de um objeto falha porque não foi detetado (70%) irá ocorrer um falso positivo de Mão Vazia e um falso negativo para o objeto correto não detetado.Acresce ainda os casos em que o tipo de produto é confundido com outro (19%), ocorrendo um falso positivo para o tipo de produto confundido e um falso negativo para o correto. Como tal, a principal causa dos resultados menos bons na deteção de cada tipo de produto deve-se à capacidade limitada do detetor de objetos reconhecer corretamente todas as instâncias reais, sendo os casos restantes causados pela lógica implementada para seleção de tramas mais significativas, onde o objeto pode ou não estar presente na trama selecionada.

Capítulo 5 – Conclusões e Recomendações

A presente dissertação foca-se no desenvolvimento de um modelo para extração e classificação de ações de clientes em espaços de loja. Em particular, na deteção e classificação de interações de clientes com prateleiras em loja.

O trabalho desenvolvido pode ser sumarizado nos seguintes pontos:

- Proposta de um sistema que interliga dois modelos para deteção e classificação - gestos e objetos – com base em tramas *RGB-D*.
- Criação de um conjunto de dados em ambiente simulado para desenvolvimento e teste de modelos inseridos no mesmo contexto de aplicação, caracterizado por cenários com diferentes níveis de complexidade para deteção de gestos e objetos
- Análise das principais particularidades dos espaços de vendas que condicionam o desempenho do modelo e quantificação do seu impacto nos resultados.
- Avaliação do desempenho do sistema sobre diversas perspetivas de análise – desempenho do sistema global e por componente (gestos e objetos) - e diferentes níveis de agregação de resultados – por tipo de esqueleto, cenário e gesto.

A validação do modelo desenvolvido foi feita unicamente recorrendo conjunto de dados recolhido.

Tal acontece porque não foram encontrados conjuntos de dados de acesso público enquadrados no tema da presente dissertação. Por um lado, é problemático por não existir uma base de comparação de desempenho face ao estado-de-arte, por outro é encarado como uma oportunidade de contributo científico.

5.1 Principais conclusões

Com base nos esforços de investigação relatados no desenvolvimento da solução (Capítulo 3) e respetiva extração e avaliação de resultados (Capítulo 4), considera-se que, embora existam desvios no comportamento do artefacto desenvolvido, os resultados são suficientemente bons porque:

- Grande parte (89.34%) dos gestos relevantes são detetados e classificados, embora apenas aproximadamente 57.25% das deteções/classificações de gestos estejam corretas, resultando num *f1-score* médio entre modelos de 69.78%.

- Para um total de cinco classes de produto, a generalidade das deteções de produtos estão corretas e reconhecem-se 66.5% do total de instâncias reais.
- O sistema é capaz de distinguir deteções feitas sobre condições normais, isto é, esqueletos corretamente captados e identificados, daquelas que são feitas quando o sistema de reconhecimento de esqueletos falha.

Apesar do desempenho individual das componentes ser satisfatório, revelou-se insuficiente quando a informação obtida por estes é generalizada nos quatro tipos distintos de interações com prateleiras. A informação individual das componentes permite reconhecer, aproximadamente, 20% das interações positivas existentes sendo que, das detetadas, quase metade destas estão corretas (41.85%). No entanto, em média, os modelos detetam 54% das interações existentes e estão corretos 30.5% das vezes.

O grau de detalhe dos resultados das componentes individuais (gestos e objetos) explica parte desta incapacidade, destacando-se:

- A proporção significativa de falsos positivos na deteção dos gestos flexão e extensão do braço (Apêndice D) que condiciona a capacidade de inferir os intervalos de tramas corretos correspondentes às interações reais, assim como a seleção das tramas relevantes para deteção de produtos.
- A elevada proporção de falsos negativos dos quatro tipos de produto (garrafas de água, embalagens de leite, caixas de cereais e livros) inferindo-se erradamente que o cliente tem a mão vazia (Tabela 29), o que se traduz num *recall* por tipo de produto seja bastante mais baixo (Tabela 27) e condiciona a capacidade de inferir corretamente o tipo de interação com a prateleira (Tabela 18)

O referido grau de detalhe dos resultados permitiu ainda concluir, à luz das questões de investigação, que:

- Existem diferenças em desempenho não superiores a 5% em *accuracy* entre cenários de diferentes complexidades, ao nível da deteção e classificação de gestos (Tabela 23).
- O impacto das falhas do sistema de reconhecimento de esqueletos quantificáveis é reduzido, não excedendo perdas de 0.97% em *accuracy*. No entanto, justificam uma parte significativa dos Falsos Negativos na deteção e classificação de gestos (Apêndice D).

- A falha mais recorrente do sistema para reconhecimento de esqueletos é o não reconhecimento/deteção de um determinado esqueleto, ocorrendo em proporção semelhante para esqueletos mais próximos (A) e distantes (B).

5.2 Contributos para a comunidade científica e empresarial

Em resultado da revisão literária, tornou-se evidente que a comunidade científica carece em estudos específicos do contexto de aplicação da presente dissertação, nomeadamente pela ausência de conjuntos de dados RGB-D públicos e número limitado de artigos científicos relacionados.

Ao nível académico-científico, o presente estudo permite uma análise detalhada do desempenho do sistema criado, salientando-se as principais causas associadas aos desvios do comportamentais do modelo desenvolvido que, por sua vez, deverão ser problemas centrais em trabalhos de investigação futuros.

Adicionalmente, serve de base comparativa para avaliação de modelos semelhantes eventualmente desenvolvidos no futuro, fazendo-se acompanhar do conjunto de dados recolhido e das métricas tipicamente usadas na avaliação de modelos para deteção e classificação de ações humanas e objetos.

Por outro lado, o contexto de aplicação do modelo desenvolvido está direcionado à inovação do setor do retalho e comércio, nomeadamente pelo contributo no estudo comportamental dos clientes.

Como umas das particularidades do modelo é o recurso a tecnologias consideradas de baixo custo e complexidade de utilização, permite a empresas, com menor capacidade de investimento em sistemas mais robustos, compreender as atuais capacidades e limitações de sistemas baseados apenas em informação de um sensor RGB-D.

Consequentemente, pode despertar interesse ao nível empresarial na investigação e desenvolvimento de sistemas mais precisos.

5.3 Limitações do estudo

Embora as limitações do presente estudo tenham sido minimizadas tanto quanto foi considerado possível, existem certas problemáticas cujo controlo não é total, com particular destaque para o sistema de reconhecimento de esqueletos

A título de exemplo, fusões de esqueletos de sujeitos diferentes não são reconhecidos ou diferenciados pelo sistema.

Adicionalmente, inferir que certos gestos não foram detetados porque o esqueleto não foi reconhecido é apenas possível em ambientes de teste controlados como os da presente dissertação.

Acresce ainda que a qualidade dos esqueletos captados, isto é, a precisão com que é localizada cada articulação do esqueleto, é variável e não foi quantificada nem relacionada com melhores ou piores resultados da classificação de gestos.

Por outro lado, a necessidade de criação de um *buffer* para minimizar a perda de tramas transmitidas em tempo real pelo sensor impede-nos de afirmar que o sistema funciona em tempo real. No conjunto clipes para testes usados existe uma média de 7 tramas em *buffer* por trama processada.

Importa também referir que o ritmo de transmissão de tramas nem sempre foi 30 tramas por segundo, por se verificar que, em função da localização dos clipes em unidades de disco rígido (*HDD*) ou de estado sólido (*SSD*) as velocidades de leitura são diferentes, existindo mais tramas em *buffer* quando o ritmo de transmissão está próximo das 30 tramas por segundo.

Por fim, considera-se que o facto da presente dissertação ser apresentada na língua portuguesa limita o seu impacto no meio científico, acrescendo que, dadas as limitações temporais para desenvolvimento da dissertação, não foi possível desenvolver um artigo científico na língua inglesa para publicação.

5.4 Propostas de investigação futura

Relativamente à investigação futura em torno do reconhecimento de interações de clientes com prateleiras em loja, sugere-se:

- Avaliar a capacidade de distinção de produtos pertencentes à mesma categoria de produto através da deteção e classificação de objetos com imagens *RGB*.
- Desenvolver métodos para reidentificar sujeitos/esqueletos sempre que estes se desloquem por áreas de captação de diferentes sensores ou quando se deixa de detetar, por momentos, determinado esqueleto.
- Melhorar a precisão da deteção e classificação de gestos, nomeadamente através da redução de falsos positivos.
- Propor um algoritmo mais robusto para reconhecer objetos em mãos com base numa sequência de tramas que compõe um gesto, de forma a que a deteção de objetos seja mais precisa.

- Avaliar o desempenho de modelos com o mesmo propósito, mas que recorram a sensores diferentes, como o futuro sucessor da *Microsoft Kinect v2*, denominado *Project Kinect for Azure*, ou sensores já existentes concorrentes à *Kinect*, como o *Asus Xtion Pro*.

Ao nível da investigação comportamental de clientes, excluindo as ideias presentes no estado-de-arte mencionadas em 2.2.5, considera-se relevante aprofundar as potenciais aplicabilidades dos resultados obtidos com o modelo desenvolvido ou semelhantes, tratando os dados de forma a serem informativos no contexto da(o):

- Análise de preferências de clientes– Relacionar situações onde o cliente hesita entre os produtos A e B, hipoteticamente concorrentes, com históricos de compras passados.
- Validação de estratégias de marketing – Como por exemplo, quantificar a adesão a companhas temporárias servindo de métrica de validação do sucesso das mesmas (p. ex. a percentagem de clientes que interagiu com a campanha e/ou que compraram).
- Desenvolvimento de estratégias de marketing aplicáveis em tempo real:
    - Desenvolver mecanismo de promoções em tempo real, comunicadas via aplicações móveis, durante a visita de clientes, potenciando a compra de produtos relacionados ou inseridos no seu perfil de interesses.
- Gestão de produtos ou marcas – Segmentar clientes em perfis por grau de interesse/aceitação a determinados tipos de produto.
- Melhoraria da experiência em loja:
    - Através dos resultados do sistema fomentar a proatividade dos colaboradores ao oferecer assistência em loja no momento certo aos clientes certos.
    - Extrair métricas relativas à experiência em loja de clientes, como o tempo médio até encontrar produto ou secções em loja mais visitadas.

Como nota final, e considerando a metodologia de investigação selecionada, o tipo de contributo em conhecimento da dissertação desenvolvida pode ser classificado como Adaptação, visto que o modelo proposto inova na adaptação de conhecimento existente para problemas novos [3]. Desta forma, considerando o estado atual da investigação em torno do reconhecimento de interações de clientes com produtos em loja, o contributo é considerado interessante à comunidade e significante, nomeadamente pelo interesse existente em torno do tema e da baixa maturidade do problema estudado, tal como retrata o estado de arte elaborado.

Bibliografia

[1] L. Xia, C.-C. Chen, and J. K. Aggarwal, "Human detection using depth information by Kinect," *Cvpr 2011 Work.*, pp. 15–22, 2011.

[2] D. P. Lacerda, A. Dresch, A. Proença, and J. A. V. Antunes Júnior, "Design Science Research: método de pesquisa para a engenharia de produção," *Gestão & Produção*, vol. 20, no. 4, pp. 741–761, 2013.

[3] B. Kuechler and S. Petter, "Design Science Research in Information Systems," *Des. Sci. Res. Inf. Syst.*, pp. 1–66, 2017.

[4] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognit. Lett.*, vol. 48, pp. 70–80, 2014.

[5] "ea-capture-studio-05 @ media.contentapi.ea.com." [Online]. Available: https://media.contentapi.ea.com/content/dam/news/e3-news-2015/en-us/common/ea-capture-studio-05.jpg.

[6] "Motion capture @ pt.wikipedia.org." [Online]. Available: https://en.wikipedia.org/wiki/Motion_capture.

[7] "vicon-cameras-vantage @ www.biometrics.fr." [Online]. Available: http://www.biometrics.fr/V4/83-197-thickbox/vicon-cameras-vantage.jpg.

[8] C. Wu, C. Stoll, L. Valgaerts, and C. Theobalt, "On-set performance capture of multiple actors with a stereo camera," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 1–11, 2013.

[9] Z. Cai, J. Han, L. Liu, and L. Shao, "RGB-D datasets using microsoft kinect or similar sensors: a survey," *Multimed. Tools Appl.*, vol. 76, no. 3, pp. 4313–4355, 2017.

[10] "The difference between Kinect v2 and v1 @ skarredghost.com." [Online]. Available: https://skarredghost.com/2016/12/02/the-difference-between-kinect-v2-and-v1/.

[11] J. Smisek, M. Jancosek, and T. Pajdla, "3D with Kinect," *Image (Rochester, N.Y.)*, pp. 1154–1160, 2011.

[12] A. J. Stoyanov, Todor and Louloudi, Athanasia and Andreasson, Henrik and Lilienthal, "Comparative evaluation of range sensor accuracy in indoor environments," *Proc. 5th Eur. Conf. Mob. Robot. ECMR 2011*, pp. 19–24, 2011.

[13] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 1995–2006, 2013.

[14] S. Zennaro *et al.*, "Performance Evaluation of the 1st and 2nd Generation Kinect For Multimedia Applications," *Multimed. Expo (ICME), 2015 IEEE Int. Conf.*, pp. 1–6, 2013.

[15] D. Jardim, L. Nunes, and M. Dias, "Human activity recognition from automatically labeled data in RGB-D videos," *2016 8th Comput. Sci. Electron. Eng. Conf. CEEC 2016 - Conf. Proc.*, pp. 89–94, 2017.

[16] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta, "Evolutionary joint selection to improve human action recognition with RGB-D devices," *Expert Syst. Appl.*, vol. 41, no. 3, pp. 786–794, 2014.

[17] W. Zhao, "A concise tutorial on human motion tracking and recognition with Microsoft Kinect," *Sci. China Inf. Sci.*, vol. 59, no. 9, pp. 1–5, 2016.

[18] D. Jardim, L. Nunes, and M. Dias, "Human activity recognition from automatically labeled data in RGB-D videos," *2016 8th Comput. Sci. Electron. Eng. Conf. CEEC 2016 - Conf. Proc.*, pp. 89–94, 2017.

[19] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, 2013.

[20] Jungong Han, Ling Shao, Dong Xu, and J. Shotton, "Enhanced Computer Vision With Microsoft Kinect Sensor: A Review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013.

[21] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," *Robot. Autom. (ICRA), 2012 IEEE Int. Conf.*, pp. 842–849, 2012.

[22] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[23] R. Lippmann, "An introduction to computing with neural nets," *IEEE Assp Mag.*, vol. 4, no. 2, pp. 4–22, 1987.

[24] S. Ji, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition.," *Pami*, vol. 35, no. 1, pp. 221–31, 2013.

[25] P. Pullen and W. Seffens, “Machine Learning Gesture Analysis of Yoga for Exergame Development,” pp. 1–6.

[26] W. Li and O. M. Way, “Action Recognition Based on A Bag of 3D Points,” pp. 9–14, 2010.

[27] X. Yang, C. Zhang, and Y. Tian, “Recognizing actions using depth motion maps-based histograms of oriented gradients,” *Proc. 20th ACM Int. Conf. Multimed. - MM ’12*, p. 1057, 2012.

[28] “MSR Action Recognition Datasets and Codes.” [Online]. Available: https://www.microsoft.com/en-us/research/people/zliu/?from=https%3A%2F%2Fresearch.microsoft.com%2Fen-us%2Fum%2Fpeople%2Fzliu%2Factionrecorsrc%2F.

[29] X. Yang and Y. L. Tian, “EigenJoints-based action recognition using Naive-Bayes-Nearest-Neighbor,” *2012 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 14–19, 2012.

[30] L. Xia, C.-C. Chen, and J. K. Aggarwal, “View Invarient Human Action Recognition Using Histograms of 3D Joints,” *Work. Hum. Act. Underst. from 3D Data*, pp. 20–27, 2012.

[31] H. Zhang and L. E. Parker, “4-Dimensional local spatio-temporal features for human activity recognition,” *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 2044–2049, 2011.

[32] A. Wang, J. Citation Wang, J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Title Mining actionlet ensemble for action recognition with depth cameras Mining Actionlet Ensemble for Action Recognition with Depth Cameras,” pp. 1290–1297, 2012.

[33] G. Ballin, M. Munaro, and E. Menegatti, “Human Action Recognition from RGB-D Frames Based on Real-Time 3D Optical Flow Estimation,” *Biol. Inspired Cogn. Archit. 2012*, vol. 196, pp. 65–74, 2013.

[34] M. Quintana, J. M. Menendez, F. Alvarez, and J. P. Lopez, “Improving retail efficiency through sensing technologies: A survey,” *Pattern Recognit. Lett.*, vol. 81, pp. 3–10, 2016.

[35] D. Liciotti and E. Frontoni, “Video Analytics for Audience Measurement,” vol. 8811, no.

December, 2014.

[36] "Depth_Sensors_Comparison @ wiki.ipisoft.com." [Online]. Available: http://wiki.ipisoft.com/Depth_Sensors_Comparison.

[37] E. Frontoni and P. Raspa, "New Trends in Image Analysis and Processing – ICIAP 2013," vol. 8158, no. March 2016, 2013.

[38] A. Mancini, E. Frontoni, P. Zingaretti, and V. Placidi, "Smart Vision System for Shelf Analysis in Intelligent Retail Environments," *Vol. 4 18th Des. Manuf. Life Cycle Conf. 2013 ASME/IEEE Int. Conf. Mechatron. Embed. Syst. Appl.*, no. August, p. V004T08A045, 2013.

[39] D. Liciotti, E. Frontoni, A. Mancini, and P. Zingaretti, "Pervasive system for consumer behaviour analysis in retail environments," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10165 LNCS, pp. 12–23, 2017.

[40] C. Migniot and F. Ababsa, "3D Human Tracking from Depth Cue in a Buying Behavior Analysis Context," *Comput. Anal. Images Patterns*, no. Caip, 2013.

[41] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[42] D. Minnen, T. L. Westeyn, T. Starner, J. a Ward, and P. Lukowicz, "Performance Metrics and Evaluation Issues for Continuous Activity Recognition," *Proc. Int. Work. Perform. Metrics Intell. Syst.*, pp. 141–148, 2006.

[43] "Map-Mean-Average-Precision-for-Object-Detection-45C121a31173 @ Medium.Com." [Online]. Available: https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173.

[44] "understanding-the-map-evaluation-metric-for-object-detection-a07fe6962cf3 @ medium.com." [Online]. Available: https://medium.com/@timothycarlen/understanding-the-map-evaluation-metric-for-object-detection-a07fe6962cf3.

[45] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2016.

[46] "yolo-you-only-look-once-real-time-object-detection-explained-492dc9230006 @ towardsdatascience.com." [Online]. Available: https://towardsdatascience.com/yolo-you-

only-look-once-real-time-object-detection-explained-492dc9230006.

[47] J. Redmon, A. Farhadi, and C. Ap, "YOLOv3 : An Incremental Improvement," *Tech Rep.*, 2018.

[48] "YOLO: Real-Time Object Detection@ pjreddie.com." [Online]. Available: https://pjreddie.com/darknet/yolo/.

[49] "What's new in YOLO v3? @ towardsdatascience.com." [Online]. Available: https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b.

[50] X. O. Xdk and K. F. Windows, "Visual Gesture Builder : A Data-Driven Solution to Gesture Detection," pp. 1–17, 2014.

[51] "Custom-Gestures-End-to-End-with-Kinect-and-Visual-Gesture-Builder @ channel9.msdn.com." [Online]. Available: https://channel9.msdn.com/Blogs/k4wdev/Custom-Gestures-End-to-End-with-Kinect-and-Visual-Gesture-Builder.

[52] "Custom-Gestures-End-to-End-with-Kinect-and-Visual-Gesture-Builder-part-2- @ channel9.msdn.com." [Online]. Available: https://channel9.msdn.com/Blogs/k4wdev/Custom-Gestures-End-to-End-with-Kinect-and-Visual-Gesture-Builder-part-2-.

[53] "Gesture Recognition Toolkit @ Www.Nickgillian.Com." [Online]. Available: http://www.nickgillian.com/wiki/pmwiki.php?n=GRT.GestureRecognitionToolkit.

[54] "OpenframeworksKinectExample @ www.nickgillian.com." [Online]. Available: http://www.nickgillian.com/wiki/pmwiki.php/GRT/OpenframeworksKinectExample.

[55] "virtual-machines @ developer.microsoft.com." [Online]. Available: https://developer.microsoft.com/en-us/windows/downloads/virtual-machines.

[56] "VMware WorkStation 14 Player." [Online]. Available: https://my.vmware.com/en/web/vmware/free#desktop_end_user_computing/vmware_workstation_player/14_0.

[57] "NtKinect: Kinect V2 C++ Programming with OpenCV on Windows10 @ nw.tsuda.ac.jp." [Online]. Available: http://nw.tsuda.ac.jp/lec/kinect2/index-en.html.

[58] "darknet @ github.com." [Online]. Available: https://github.com/AlexeyAB/darknet.

[59] "AdaBoostTrigger - Input Parameters." [Online]. Available: https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn785522%28v%3Dieb.10%29.

[60] "Yolo_mark @ github.com." [Online]. Available: https://github.com/AlexeyAB/Yolo_mark.

[61] "Hardware Guide: Neural Networks on GPUs." [Online]. Available: https://pjreddie.com/darknet/hardware-guide/.

[62] "AdaBoostTrigger @ docs.microsoft.com." [Online]. Available: https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn785522%28v%3Dieb.10%29.

Anexos
Anexo A

Nome	Valor	Tipo	Descrição	Racional
Accuracy Level	0.95	FLOAT	Valor decimal que controla a precisão dos resultados e com impacto no tempo de treino. Quanto maior a precisão, maior o tempo de treino e devem ser usados os valores de referência: • *Retail Build* – 0.98. • *Release Build* – 0.95. • *Quick Experiment* – 0.8.	Para efeitos de desenvolvimento será usado o valor 0.95 e 0.98 na construção final do modelo.
Number of Weak Classifiers at Runtime	1000	INT	O algoritmo pode potencialmente gerar dezenas de milhares de classificadores fracos. O uso de todos estes aumentarão a precisão, mas com um maior custo de processamento. O custo de processamento para 1000 classificadores fracos é de apenas 25 microssegundos e os resultados têm uma precisão mais que adequada.	Testes preliminares em 3.4
Filter Results	True	BOOL	Os resultados do algoritmo são calculados por trama e não por gesto. Como tal, é necessário aplicar um filtro aos resultados brutos por trama. O *AdaBoostTrigger* oferece um filtro simples de baixa latência, no entanto é possível desativar este filtro e aplicar filtros customizados O filtro usado consiste numa janela deslizante simples de N tramas, que soma os resultados e compara-os com um valor limite. O número de tramas pode ser visto como uma frequência e o limite como a amplitude.	O uso do filtro de resultados padrão é adequado.
Auto Find Best Filtering Params	True	BOOL	Quando a filtragem de resultados está ativada, é possível que a fase de treino encontre automaticamente os melhores parâmetros de filtragem, o que minimizará a taxa de falsos positivos e falsos negativos.	Optou-se por beneficiar da otimização automática de parâmetros.
Weight Of False Positives During Auto Find	0.5	FLOAT	Um valor no intervalo de [0..1] que é usado ao encontrar automaticamente os melhores parâmetros de filtragem. Se for mais importante reduzir os falsos positivos, recomenda-se um valor mais alto. Se for mais importante reduzir os falsos negativos, deve ser usado um valor menor.	Testes preliminares em 3.4

Nome	Valor	Tipo	Descrição	Racional
Manual Filter Params: Num Tramas To Filter	5	INT	Um valor no intervalo de [1..10]. Define o número de tramas para filtrar caso não se opte pela filtragem automática.	Optou-se por beneficiar da otimização automática de parâmetros.
Manual Filter Params: Threshold	0.001	FLOAT	Um valor no intervalo de [0..1]. Valores mais baixos podem aumentar os verdadeiros positivos, com o risco de aumentar os falsos positivos. Valores mais altos podem diminuir os falsos positivos, com o risco de diminuir os verdadeiros positivos.	Optou-se por beneficiar da otimização automática de parâmetros-
Duplicate And Mirror Data During Training	False	BOOL	Se definido como **verdadeiro** (valor padrão), todos os dados do gesto serão duplicados e espelhados durante a fase de treino. Se houver dados específicos de um lado do corpo (Esquerdo vs. Direito), esses dados também serão espelhados e usados durante a fase de treino. Define-se como **falso** se o seu gesto exigir movimento de ambos os lados do corpo (Body Side = Any) e o gesto for assimétrico. Por exemplo, se o gesto exigir que o braço direito esteja levantado enquanto o braço esquerdo estiver para baixo.	Não aplicável
% CPU for Training	95	INT	Um valor no intervalo de [0..100] que indica a percentagem de recursos do CPU a ser usados na fase de treino.	Dado que a fase de treino é relativamente rápida manteve-se o valor padrão de 95%.
Use Hands Data	False	BOOL	Se definido como **verdadeiro**, os estados de mão suportados (aberto / fechado / laço) serão usados durante a fase de treino e a deteção. No entanto, esta opção não deve ser usada se a aplicação suportar mais de dois sujeitos simultâneos.	Definido como falso visto que não é possível a sua utilização com mais de dois sujeitos e porque existe alguma ineficácia na deteção do estado das mãos o que confundirá o detetor.
Ignore Left Arm	False	BOOL	Se definido como **verdadeiro**, todas as articulações do braço esquerdo serão ignoradas durante o treino e a detecção. Define-se como verdadeiro se o seu gesto não depender do movimento / posição do braço esquerdo (Murro_Direito, Empurrar_Direita, Pontapé, etc.).	As articulações do braço esquerdo são fundamentais para os gestos que se pretende captar.

Nome	Valor	Tipo	Descrição	Racional
Ignore Right Arm	False	BOOL	Se definido como **verdadeiro**, todas as articulações do braço direito serão ignoradas durante o treino e a detecção. Define-se como verdadeiro se o seu gesto não depender do movimento / posição do braço direito (Murro_Esquerdo, Empurrar_Esquerda, Pontapé, etc.).	As articulações do braço direito são fundamentais para os gestos que se pretende captar.
Ignore Lower Body	False	BOOL	Se definido como **verdadeiro**, todas as articulações abaixo da região do quadril serão ignoradas durante o treino e a deteção. Define-se como **falso** se o seu gesto não depender do movimento / posição da parte inferior do corpo.	Testes preliminares em 3.4

Tabela 32 – Parâmetros de entrada usados na construção do detetor de gestos com VGB [62].

Nome	Racional
Nível de precisão	Será usado o valor 0.99 visto que o tempo de treino é menos importante do que a precisão dos resultados
Número de classificadores fracos em tempo de execução	Testes preliminares em 3.4
Filtragem de resultados	O uso do filtro de resultados padrão é adequado dada a baixa latência.
Encontrar melhores parâmetros de filtragem automaticamente	Optou-se por beneficiar da otimização automática de parâmetros dado minimizar as taxas de falsos positivos/negativos.
Peso dos Falsos Positivos durante filtragem	Testes preliminares em 3.4
Parâmetros de filtragem manuais: Nº Tramas a filtrar	Optou-se por beneficiar da otimização automática de parâmetros para filtragem pelo que este campo é ignorado.
Parâmetros de filtragem manuais: Thresholda	Optou-se por beneficiar da otimização automática de parâmetros para filtragem pelo que este campo é ignorado.
Duplicar e espelhar dados durante treino	Visto que não é relevante distinguir extensões/flexões de braço com a mão esquerda/direita os dados de treino não serão duplicados e espelhados.
% CPU para treino	Dado que a fase de treino é relativamente rápida manteve-se o valor padrão de 95%.
Usar dados das mãos	Definido como falso visto que não é possível a sua utilização com mais de dois sujeitos e porque existe alguma ineficácia na deteção do estado das mãos o que confundirá o detetor.
Ignorar braço esquerdo	As articulações do braço esquerdo são fundamentais para os gestos que se pretende captar.
Ignorar braço direito	As articulações do braço direito são fundamentais para os gestos que se pretende captar.
Ignorar parte inferior do corpo	Testes preliminares em 3.4

Tabela 33 – Decisão tomada por parâmetro de configuração relativo ao processo de treino (VGB).

Apêndices

Apêndice A

Guião A: Interações de um sujeito

Durante aproximadamente **1 minuto** deverão ser desempenhadas de forma **alternada** as seguintes interações com qualquer uma das duas prateleiras:

a. Pegar no produto e meter no cesto (interação positiva)

b. Pegar no produto e voltar a colocar na prateleira ou não (interação negativa com/sem novo produto)

c. Estender a mão e não pegar nada (interação neutra)

Nota: Deve pegar em produtos posicionados em diferentes andares da prateleira.
Pretende-se captar as várias formas de desempenhar uma mesma ação (i.e., agarrar um produto agachando, inclinando o corpo ou permanecendo numa postura vertical)

Guião B: Cenários de Oclusão Simples– Interações de dois sujeitos

Durante aproximadamente **1 minuto** deverão ser desempenhadas as mesmas interações com a prateleira referidas no guião A.

Deve interagir com apenas com a prateleira à sua frente e não trocar de posição com o outro sujeito.

Guião C: Cenários de Oclusão Complexo – Interações de dois sujeitos

Pretende-se avaliar a capacidade para reconhecer as interações com as prateleiras quando dois sujeitos disputam o acesso às mesmas prateleiras

Durante aproximadamente **1 minuto** deverão ser desempenhadas as mesmas interações descritas no guião A.

Deve retirar produtos da prateleira à sua frente ou da prateleira do sujeito ao seu lado.

Após aproximadamente 30 segundos irá ser solicitado que troque de posição com o outro sujeito e continue a interagir com ambas as prateleiras.

Apêndice B

Todos os cenários e esqueletos

		Atual					
		Nula	**Positiva**	**Negativa c/ Novo Produto**	**Negativa s/ Novo Produto**	**Neutra**	**Total instâncias Detetadas**
Detetado	**Nula**	**551**	308	77	59	56	1051
	Positiva	83	**77**	9	7	8	184
	Negativa c/ Novo Produto	55	53	**5**	5	9	127
	Negativa s/ Novo Produto	48	9	0	**18**	5	80
	Neutra	461	210	57	65	**43**	836
	Total Instâncias Reais	643	385	82	77	99	-

Tabela 34 - Matriz confusão com resultados ao nível da interação (Todos os cenários e esqueletos)

Tipo de interação	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Nula	52.43	85.69	65.05
Positiva	41.85	20.00	27.07
Negativa c/ Novo Produto	3.94	6.10	4.78
Negativa s/ Novo Produto	22.50	23.38	22.93
Neutra	5.14	43.43	9.20

Tabela 35 - Métricas de desempenho precision e recall por tipo de interação (Todos os cenários e esqueletos)

Cenário/Guião A

		Atual					
		Nula	**Positiva**	**Negativa c/ Novo Produto**	**Negativa s/ Novo Produto**	**Neutra**	**Total instâncias Detetadas**
Detetado	**Nula**	**145**	79	2	23	7	256
	Positiva	16	**19**	1	4	2	42
	Negativa c/ Novo Produto	8	23	**1**	5	2	39
	Negativa s/ Novo Produto	21	5	0	**9**	1	36
	Neutra	92	50	4	33	**14**	193
	Total Instâncias Reais	154	98	3	32	21	-

Tabela 36 - Matriz confusão com resultados ao nível da interação para o cenário A.

Tipo de interação	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Nula	56.64	94.16	70.73
Positiva	45.24	19.39	27.14
Negativa c/ Novo Produto	2.56	33.33	4.76
Negativa s/ Novo Produto	25.00	28.13	26.47
Neutra	7.25	66.67	13.08

Tabela 37 - Métricas de desempenho precision, recall e F1-Score por tipo de interação para o cenário A.

Cenário/Guião B

		Atual					
		Nula	**Positiva**	**Negativa c/ Novo Produto**	**Negativa s/ Novo Produto**	**Neutra**	**Total instâncias Detetadas**
Detetado	**Nula**	**213**	116	41	18	28	416
	Positiva	35	**23**	5	2	2	67
	Negativa c/ Novo Produto	15	15	**2**	0	5	37
	Negativa s/ Novo Produto	14	2	0	**6**	4	26
	Neutra	165	76	31	19	**9**	300
	Total Instâncias Reais	243	139	43	24	37	-

Tabela 38 - Matriz confusão com resultados ao nível da interação para o cenário B e ambos os esqueletos.

Tipo de interação	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Nula	51.20%	87.65%	64.64%
Positiva	34.33%	16.55%	22.33%
Negativa c/ Novo Produto	5.41%	4.65%	5.00%
Negativa s/ Novo Produto	23.08%	25.00%	24.00%
Neutra	3.00%	24.32%	5.34%

Tabela 39 - Métricas de desempenho precision e recall por tipo de interação para o cenário B e ambos os esqueletos.

Cenário/Guião C

		Atual					
		Nula	**Positiva**	**Negativa c/ Novo Produto**	**Negativa s/ Novo Produto**	**Neutra**	**Total instâncias Detetadas**
Detetado	**Nula**	**193**	113	34	18	21	379
	Positiva	32	**35**	3	1	4	75
	Negativa c/ Novo Produto	32	15	**2**	0	2	51
	Negativa s/ Novo Produto	13	2	0	**3**	0	18
	Neutra	204	84	22	13	**20**	343
	Total Instâncias Reais	246	148	36	21	41	-

Tabela 40 - Matriz confusão com resultados ao nível da interação para o cenário C e ambos os esqueletos.

Tipo de interação	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Nula	50.92%	78.46%	61.76%
Positiva	46.67%	23.65%	31.39%
Negativa c/ Novo Produto	3.92%	5.56%	4.60%
Negativa s/ Novo Produto	16.67%	14.29%	15.38%
Neutra	5.83%	48.78%	10.42%

Tabela 41 - Métricas de desempenho precision e recall por tipo de interação para o cenário C e ambos os esqueletos.

Esqueleto A

		Atual					
		Nula	**Positiva**	**Negativa c/ Novo Produto**	**Negativa s/ Novo Produto**	**Neutra**	**Total instâncias Detetadas**
Detetado	**Nula**	**338**	190	16	59	34	637
	Positiva	40	**48**	4	7	6	105
	Negativa c/ Novo Produto	16	39	**1**	5	3	64
	Negativa s/ Novo Produto	48	9	0	**18**	5	80
	Neutra	277	126	12	65	**25**	505
	Total Instâncias Reais	391	238	17	77	59	-

Tabela 42 - Matriz confusão agregada dos cenários A, B e C para o esqueleto A dos resultados ao nível da interação.

Tipo de interação	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Nula	53.06	86.45	65.76
Positiva	45.71	20.17	27.99
Negativa c/ Novo Produto	1.56	5.88	2.47
Negativa s/ Novo Produto	22.50	23.38	22.93
Neutra	4.95	42.37	8.87

Tabela 43 - Métricas de desempenho precision e recall dos cenários A, B e C para o esqueleto A por tipo de interação.

Esqueleto B

		Atual					
		Nula	**Positiva**	**Negativa c/ Novo Produto**	**Negativa s/ Novo Produto**	**Neutra**	**Total instâncias Detetadas**
Detetado	**Nula**	**213**	118	61	0	22	414
	Positiva	43	**29**	5	0	2	79
	Negativa c/ Novo Produto	39	14	**4**	0	6	63
	Negativa s/ Novo Produto	0	0	0	**0**	0	0
	Neutra	184	84	45	0	**18**	331
	Total Instâncias Reais	252	147	65	0	40	

Tabela 44 - Matriz confusão agregada dos cenários B e C para o esqueleto B dos resultados ao nível da interação.

Tipo de interação	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Nula	51.45	84.52	63.96
Positiva	36.71	19.73	25.66
Negativa c/ Novo Produto	6.35	6.15	6.25
Negativa s/ Novo Produto	-	-	-
Neutra	5.44	45.00	9.70

Tabela 45 – Métricas de desempenho precision e recall dos cenários B e C para o esqueleto B, por tipo de interação.

Apêndice C

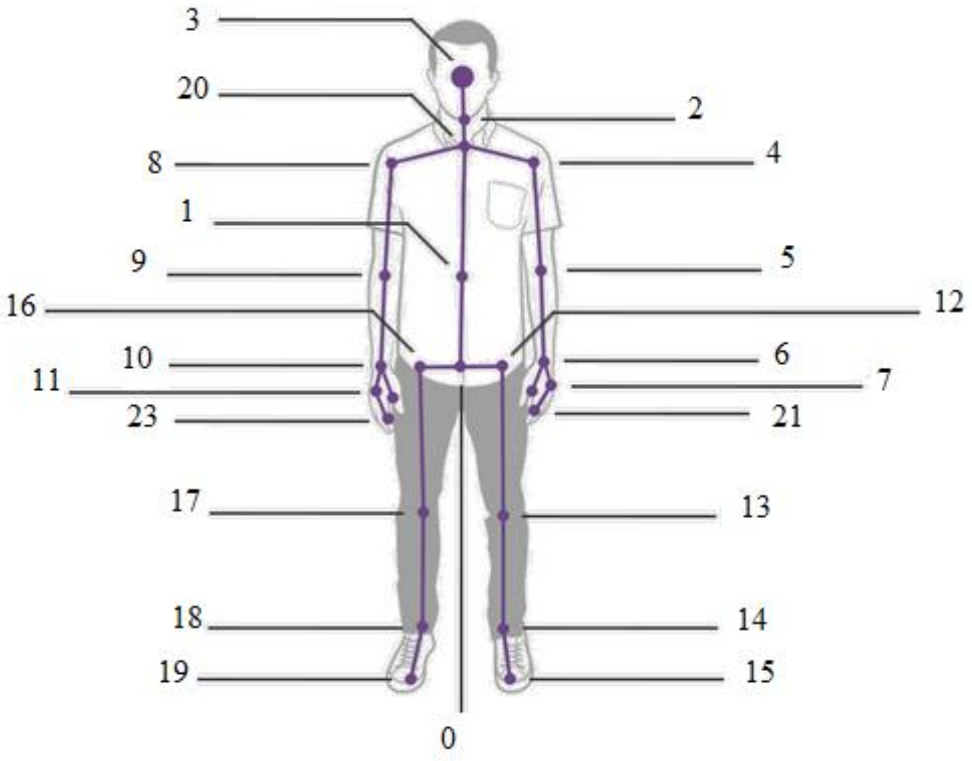


Figura 25 - Esqueleto captado via Microsoft Kinect v2

Identificador Articulação	**Nome**	**Identificador Articulação**	**Nome**
0	Base da Espinha	13	Joelho Esquerdo
1	Centro da Espinha	14	Tornozelo Esquerdo
2	Pescoço	15	Pé Esquerdo
3	Cabeça	16	Anca Direita
4	Ombro Esquerdo	17	Joelho Direito
5	Cotovelo Esquerdo	18	Tornozelo Direito
6	Pulso Esquerdo	19	Pé Direito
7	Mão Esquerda	20	Espinha do ombro
8	Ombro Direito	21	Ponta da Mão Esquerda
9	Cotovelo Direito	22	Polegar Esquerdo
10	Pulso Direito	23	Ponta da Mão Direita
11	Mão Direita	24	Polegar Direito
12	Anca esquerda		

Tabela 46 - Correpondência entre identificador de articulação e respetivo nome

Apêndice D
Todos os cenários e esqueletos

	Verdadeiros Positivos						**Verdadeiros Negativos**						**Falsos Positivos**						**Falsos Negativos**					
Gesto	Extensão			Flexão			Extensão			Flexão			Extensão			Flexão			Extensão			Flexão		
Causa	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*
Total p/ Causa	-	2	574	-	1	580	-	-	220	-	-	211	-	2	429	-	1	432	2	9	64	1	12	50
Total p/ gesto	576			581			220			211			431			433			75			63		
Total p/ componente Matriz Confusão	1157						431						864						138					
Total *C1*	-						-						-						3					
Total *C2*	3						-						3						21					
Total *C3*	1154						431						861						114					

Tabela 47 - Matriz confusão de resultados ao nível do gesto (Cenários e esqueletos agregados).

Cenário/Guião A

	Verdadeiros Positivos						**Verdadeiros Negativos**						**Falsos Positivos**						**Falsos Negativos**					
Gesto	Extensão			Flexão			Extensão			Flexão			Extensão			Flexão			Extensão			Flexão		
Causa	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*
Total p/ Causa	-	-	142	-	-	142	-	-	59	-	-	59	-	-	95	-	-	95	-	5	7	-	4	8
Total p/ gesto	142			142			59			59			95			95			12			12		
Total p/ componente Matriz Confusão	284						118						190						24					
Total *C1*	-						-						-						-					
Total *C2*	-						-						-						9					
Total *C3*	284						118						190						15					

Tabela 48 - Matriz confusão de resultados ao nível do gesto (Cenário A).

Cenário/Guião B

	Verdadeiros Positivos						**Verdadeiros Negativos**						**Falsos Positivos**						**Falsos Negativos**					
Gesto	Extensão			Flexão			Extensão			Flexão			Extensão			Flexão			Extensão			Flexão		
Causa	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*
Total p/ Causa	-	-	219	-	-	218	-	-	83	-	-	92	-	1	162	-	-	151	-	4	23	-	5	20
Total p/ gesto	219			218			83			92			163			151			27			25		
Total p/ componente Matriz Confusão	437						175						314						52					
Total *C1*	-						-						-						-					
Total *C2*	-						-						1						9					
Total *C3*	437						175						313						43					

Tabela 49 - Matriz confusão de resultados ao nível do gesto (Cenário B).

Cenário/Guião C

	Verdadeiros Positivos						**Verdadeiros Negativos**						**Falsos Positivos**						**Falsos Negativos**					
Gesto	Extensão			Flexão			Extensão			Flexão			Extensão			Flexão			Extensão			Flexão		
Causa	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*
Total p/ Causa	-	2	213	-	1	220	-	-	78	-	-	60	-	1	172	-	1	186	2	-	34	1	3	22
Total p/ gesto	215			221			78			60			173			187			36			26		
Total p/ componente Matriz Confusão	436						138						360						62					
Total *C1*	-						-						-						3					
Total *C2*	3						-						2						3					
Total *C3*	433						138						358						56					

Tabela 50 - Matriz confusão de resultados ao nível do gesto (Cenário C).

Esqueleto A

	Verdadeiros Positivos						**Verdadeiros Negativos**						**Falsos Positivos**						**Falsos Negativos**					
Gesto	Extensão			Flexão			Extensão			Flexão			Extensão			Flexão			Extensão			Flexão		
Causa	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*
Total p/ Causa	-	2	359	-	1	367	-	-	133	-	-	138	-	2	258	-	1	253	-	5	27	-	6	18
Total p/ gesto	361			368			133			138			260			254			32			24		
Total p/ componente Matriz Confusão	729						271						514						56					
Total *C1*	-						-						-						-					
Total *C2*	3						-						3						11					
Total *C3*	726						271						510						45					

Tabela 51 - Matriz confusão de resultados ao nível do gesto (Esqueleto A).

Esqueleto B

	Verdadeiros Positivos						**Verdadeiros Negativos**						**Falsos Positivos**						**Falsos Negativos**					
Gesto	Extensão			Flexão			Extensão			Flexão			Extensão			Flexão			Extensão			Flexão		
Causa	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*
Total p/ Causa	-	-	215	-	-	213	-	-	87	-	-	73	-	-	171	-	-	179	2	4	37	1	6	32
Total p/ gesto	215			213			87			73			171			179			43			39		
Total p/ componente Matriz Confusão	428						160						350						82					
Total *C1*	-						-						-						3					
Total *C2*	-						-						-						10					
Total *C3*	428						160						350						69					

Tabela 52 - Matriz confusão de resultados ao nível do gesto (Esqueleto B).

Apêndice E

Resultados	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Globais	61.91%	57.27%	91.01%	70.30%
Guião A	66.23%	59.92%	94.98%	73.48%
Guião B	63.22%	58.27%	91.04%	71.06%
Guião C	57.97%	54.74%	88.55%	67.66%
Esq. A	64.24%	58.74%	94.16%	72.35%
Esq. B	58.39%	55.01%	86.12%	67.14%
Extensão	61.69%	57.23%	89.97%	69.96%
Flexão	62.14%	57.31%	91.99%	70.62%

Tabela 53 - Métricas de desempenho ao nível do gesto excluindo causas C1 e C2, por perspetiva de análise.

Apêndice F

Tipo de Produto (Classes p/ classificação)	Característica diferenciadora	Exemplo
Caixas de cereais	*Nestum*	
	Kellogg's	
Livros	Capa vermelha, branca e cinza	
	Capa azul e cinza	
	Capa azul	

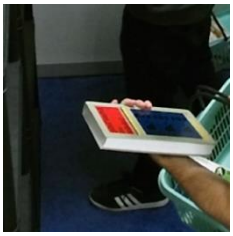
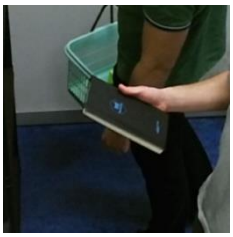
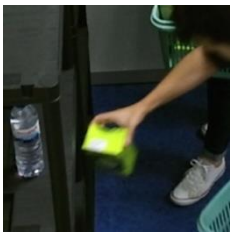
Tipo de Produto (Classes p/ classificação)	Característica diferenciadora	Exemplo
	Capa vermelha, azul e cinza	
	Capa preta	
Embalagens de Leite	-	
Garrafas de água	-	

Tabela 54 - Exemplos de imagens usadas para treinar a rede Darknet Yolo v3.

Apêndice G

Weight File	*Precision*	*Recall*	*F1 Score*	*Avg IoU*	*mAP*	*AVG(mAP,IoU)*
6000	90.00%	78.00%	84.00%	68.21%	81.34%	74.78%
6100	79.00%	75.00%	77.00%	58.21%	76.98%	67.60%
6200	74.00%	74.00%	74.00%	54.35%	74.42%	64.39%
6300	88.00%	70.00%	78.00%	68.66%	76.85%	72.76%
6400	71.00%	69.00%	70.00%	51.23%	67.87%	59.55%
6500	94.00%	68.00%	79.00%	74.69%	77.22%	75.96%
6600	88.00%	76.00%	81.00%	65.73%	81.17%	73.45%
6700	94.00%	74.00%	82.00%	72.09%	79.44%	75.77%
6800	90.00%	74.00%	81.00%	69.32%	80.24%	74.78%
6900	92.00%	71.00%	80.00%	71.93%	78.84%	75.39%
7000	89.00%	77.00%	93.00%	69.79%	80.44%	75.12%
7100	86.00%	70.00%	77.00%	63.35%	77.42%	70.39%
7200	90.00%	78.00%	83.00%	68.75%	80.89%	74.82%
7300	81.00%	80.00%	81.00%	59.97%	78.51%	69.24%
7400	83.00%	67.00%	74.00%	61.37%	74.95%	68.16%
7500	94.00%	68.00%	79.00%	74.17%	78.44%	76.31%
7600	82.00%	75.00%	79.00%	61.07%	80.77%	70.92%
7700	92.00%	75.00%	83.00%	72.48%	80.44%	76.46%
7800	92.00%	75.00%	83.00%	73.01%	80.83%	76.92%
7900	67.00%	80.00%	73.00%	47.01%	73.45%	60.23%
8000	70.00%	77.00%	73.00%	50.83%	73.05%	61.94%
8100	78.00%	78.00%	78.00%	57.91%	76.90%	67.41%
8200	91.00%	75.00%	82.00%	69.75%	79.21%	74.48%
8300	84.00%	78.00%	81.00%	63.87%	81.77%	72.82%
8400	92.00%	78.00%	84.00%	71.99%	79.50%	75.75%
8500	74.00%	75.00%	74.00%	53.65%	71.12%	62.39%
8600	94.00%	73.00%	82.00%	74.31%	79.39%	76.85%
8700	93.00%	73.00%	82.00%	74.39%	79.08%	76.74%
8800	82.00%	72.00%	77.00%	60.54%	75.85%	68.20%
8900	92.00%	78.00%	84.00%	70.50%	81.54%	76.02%
9000	93.00%	76.00%	84.00%	73.69%	79.43%	76.56%

Weight File	*Precision*	*Recall*	*F1 Score*	*Avg IoU*	*mAP*	*AVG(mAP,IoU)*
9100	76.00%	76.00%	76.00%	56.96%	78.22%	67.59%
9200	88.00%	67.00%	76.00%	66.30%	75.84%	71.07%
9300	94.00%	73.00%	82.00%	74.01%	81.32%	77.67%
9400	**94.00%**	**76.00%**	**84.00%**	**75.00%**	**81.53%**	**78.27%**
9500	93.00%	80.00%	86.00%	72.75%	81.81%	77.28%
9600	93.00%	70.00%	80.00%	73.13%	75.05%	74.09%
9700	90.00%	74.00%	81.00%	70.03%	79.34%	74.69%
9800	93.00%	72.00%	81.00%	73.84%	78.27%	76.06%
9900	88.00%	76.00%	81.00%	67.74%	77.88%	72.81%
10000	85.00%	78.00%	81.00%	64.10%	79.05%	71.58%
10100	72.00%	82.00%	77.00%	52.40%	78.38%	65.39%
10200	79.00%	74.00%	77.00%	59.43%	78.52%	68.98%

Tabela 55 - Resultados do processo de treino da rede para deteção de objetos por estado de rede.

Apêndice H

	Água			Livro			Caixa de Cereais			Leite			Mão Vazia		
Resultados	***Prec. (%)***	***Rec. (%)***	***F1-S. (%)***	***Prec. (%)***	***Rec. (%)***	***F1-S. (%)***	***Prec. (%)***	***Rec. (%)***	***F1-S. (%)***	***Prec. (%)***	***Rec. (%)***	***F1-S. (%)***	***Prec. (%)***	***Rec. (%)***	***F1-S. (%)***
Globais	80.8	15.4	25.9	73.2	53.6	61.9	69.2	40.1	50.8	87.5	23.7	37.3	64.5	94.2	76.5
Guião A	57.1	14.3	22.9	90	78.3	83.7	75	40.9	52.9	84.6	36.7	51.2	67.3	93.1	78.1
Guião B	85.7	11.8	20.7	59	46.9	52.3	62.2	39.7	48.4	83.3	23.8	37	64	92.4	75.7
Guião C	91.7	19.3	31.9	82.6	47.5	60.3	73.3	40	51.8	100	15.2	26.4	63.2	96.6	76.4
Esq. A	72.7	11	19.1	77.6	55.9	65	75.4	45.4	56.7	88.2	19.5	31.9	65	94.8	77.2
Esq. B	86.7	20.6	33.3	66.7	50	57.1	53.9	28.6	37.3	86.7	31.7	46.4	63.4	93.1	75.4
Extensão	50	19.2	27.8	67.7	53.9	60	52	28.9	37.1	80	26.7	40	82.6	94.7	88.3
Flexão	100	14.6	25.4	76.5	53.4	62.9	75.8	44.6	56.2	90.9	22.7	36.4	43.2	92.9	59

Tabela 56 - Agregação de perspetivas de análise de deteções de objetos por tipo de produto.

Todos os cenários e esqueletos

		Atual					
		Água	**Livro**	**Caixa de Cereais**	**Leite**	**Mão Vazia**	**Total instâncias Detetadas**
Detetado	**Água**	**21**	0	1	0	4	26
	Livro	3	**60**	4	3	12	82
	Caixa de Cereais	6	2	**63**	2	18	91
	Leite	1	0	0	**28**	3	32
	Mão Vazia	105	50	89	85	**597**	926
	Total instâncias Reais	136	112	157	118	634	-

Tabela 57 - Matriz confusão da deteção dos cinco tipos de produtos (Todos os cenários e esqueletos).

Tipo de Produto	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Água	80.77	15.44	25.93
Livro	73.17	53.57	61.86
Caixa de Cereais	69.23	40.13	50.81
Leite	87.50	23.73	37.33
Mão Vazia	64.47	94.16	76.54

Tabela 58 – Métricas de desempenho resultantes da matriz confusão (Tabela 57).

Cenário/Guião A

		Atual					
		Água	**Livro**	**Caixa de Cereais**	**Leite**	**Mão Vazia**	**Total instâncias Detetadas**
Detetado	**Água**	**4**	0	1	0	2	7
	Livro	0	**18**	0	0	2	20
	Caixa de Cereais	0	0	**18**	0	6	24
	Leite	1	0	0	**11**	1	13
	Mão Vazia	23	5	25	19	**148**	220
	Total instâncias Reais	28	23	44	30	159	-

Tabela 59 - Matriz confusão da deteção dos cinco tipos de produtos (Cenário A).

Tipo de Produto	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Água	57.14	14.29	22.86
Livro	90.00	78.26	83.72
Caixa de Cereais	75.00	40.91	52.94
Leite	84.62	36.67	51.16
Mão Vazia	67.27	93.08	78.10

Tabela 60 – Métricas de desempenho resultantes da matriz confusão (Tabela 59).

Cenário/Guião B

		Atual					
		Água	**Livro**	**Caixa de Cereais**	**Leite**	**Mão Vazia**	**Total instâncias Detetadas**
Detetado	**Água**	**6**	0	0	0	1	7
	Livro	3	**23**	3	2	8	39
	Caixa de Cereais	3	2	**23**	2	7	37
	Leite	0	0	0	**10**	2	12
	Mão Vazia	39	24	32	28	**219**	342
	Total instâncias Reais	51	49	58	42	237	-

Tabela 61 - Matriz confusão da deteção dos cinco tipos de produtos (Cenário B).

Tipo de Produto	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Água	85.71	11.76	20.69
Livro	58.97	46.94	52.27
Caixa de Cereais	62.16	39.66	48.42
Leite	83.33	23.81	37.04
Mão Vazia	64.04	92.41	75.65

Tabela 62 – Métricas de desempenho resultantes da matriz confusão (Tabela 61).

Cenário/Guião C

		Atual					
		Água	**Livro**	**Caixa de Cereais**	**Leite**	**Mão Vazia**	**Total instâncias Detetadas**
Detetado	**Água**	**11**	0	0	0	1	12
	Livro	0	**19**	1	1	2	23
	Caixa de Cereais	3	0	**22**	0	5	30
	Leite	0	0	0	**7**	0	7
	Mão Vazia	43	21	32	38	**230**	364
	Total instâncias Reais	57	40	55	46	238	-

Tabela 63 - Matriz confusão da deteção dos cinco tipos de produtos (Cenário C).

Tipo de Produto	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Água	91.67	19.30	31.88
Livro	82.61	47.50	60.32
Caixa de Cereais	73.33	40.00	51.76
Leite	100.00	15.22	26.42
Mão Vazia	63.19	96.64	76.41

Tabela 64 – Métricas de desempenho resultantes da matriz confusão (Tabela 63).

Esqueleto A

		Atual					
		Água	**Livro**	**Caixa de Cereais**	**Leite**	**Mão Vazia**	**Total instâncias Detetadas**
Detetado	**Água**	**8**	0	1	0	2	11
	Livro	1	**38**	3	2	5	49
	Caixa de Cereais	1	1	**49**	1	13	65
	Leite	1	0	0	**15**	1	17
	Mão Vazia	62	29	55	59	**382**	587
	Total instâncias Reais	73	68	108	77	403	-

Tabela 65 - Matriz confusão da deteção dos cinco tipos de produtos (Esqueleto A).

Tipo de Produto	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Água	72.73	10.96	19.05
Livro	77.55	55.88	64.96
Caixa de Cereais	75.38	45.37	56.65
Leite	88.24	19.48	31.91
Mão Vazia	65.08	94.79	77.17

Tabela 66 – Métricas de desempenho resultantes da matriz confusão (Tabela 65).

Esqueleto B

		Atual					
		Água	**Livro**	**Caixa de Cereais**	**Leite**	**Mão Vazia**	**Total instâncias Detetadas**
Detetado	**Água**	**13**	0	0	0	2	15
	Livro	2	**22**	1	1	7	33
	Caixa de Cereais	5	1	**14**	1	5	26
	Leite	0	0	0	**13**	2	15
	Mão Vazia	43	21	34	26	**215**	339
	Total instâncias Reais	63	44	49	41	231	-

Tabela 67 - Matriz confusão da deteção dos cinco tipos de produtos (Esqueleto B).

Tipo de Produto	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Água	86.67	20.63	33.33
Livro	66.67	50.00	57.14
Caixa de Cereais	53.85	28.57	37.33
Leite	86.67	31.71	46.43
Mão Vazia	63.42	93.07	75.44

Tabela 68 – Métricas de desempenho resultantes da matriz confusão (Tabela 67).

Gesto Extensão do Braço

		Atual					
		Água	**Livro**	**Caixa de Cereais**	**Leite**	**Mão Vazia**	**Total instâncias Detetadas**
Detetado	**Água**	**5**	0	1	0	4	10
	Livro	0	**21**	1	1	8	31
	Caixa de Cereais	1	1	**13**	1	9	25
	Leite	0	0	0	**8**	2	10
	Mão Vazia	20	17	30	20	**413**	500
	Total instâncias Reais	26	39	45	30	436	-

Tabela 69 - Matriz confusão da deteção dos cinco tipos de produtos (Gesto Extensão).

Tipo de Produto	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Água	50.00	19.23	27.78
Livro	67.74	53.85	60.00
Caixa de Cereais	52.00	28.89	37.14
Leite	80.00	26.67	40.00
Mão Vazia	82.60	94.72	88.25

Tabela 70 – Métricas de desempenho resultantes da matriz confusão (Tabela 69).

Gesto Flexão do Braço

		Atual					
		Água	**Livro**	**Caixa de Cereais**	**Leite**	**Mão Vazia**	**Total instâncias Detetadas**
Detetado	**Água**	**16**	0	0	0	0	16
	Livro	3	**39**	3	2	4	51
	Caixa de Cereais	5	1	**50**	1	9	66
	Leite	1	0	0	**20**	1	22
	Mão Vazia	85	33	59	65	**184**	426
	Total instâncias Reais	110	73	112	88	198	-

Tabela 71 - Matriz confusão da deteção dos cinco tipos de produtos (Gesto Flexão).

Tipo de Produto	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
Água	100.00	14.55	25.40
Livro	76.47	53.42	62.90
Caixa de Cereais	75.76	44.64	56.18
Leite	90.91	22.73	36.36
Mão Vazia	43.19	92.93	58.97

Tabela 72 – Métricas de desempenho resultantes da matriz confusão (Tabela 71).