Universidade do Minho
Escola de Engenharia

André Filipe Proença e Silva

Apreensão e Discretização de Ambientes Tangíveis em Sistemas de Realidade Aumentada

Outubro de 2018

Universidade do Minho
Escola de Engenharia

André Filipe Proença e Silva

Apreensão e Discretização de Ambientes Tangíveis em Sistemas de Realidade Aumentada

Dissertação de Mestrado

Mestrado Integrado em Engenharia Informática

Trabalho efetuado sob a orientação dos

Professor Doutor Luís Gonzaga Mendes Magalhães

Professor Doutor Miguel Angel Guevara Lopez

Outubro de 2018

AGRADECIMENTOS

Esta secção de agradecimentos poderá parecer um tanto genérica e muito semelhante a outras que surgem em trabalhos de colegas, estudantes e demais investigadores. Embora assim o pareça, espero que as palavras que aqui vou deixar, não percam, de forma alguma, o seu valor.

Gostaria de apontar que, na vida, e nesta situação em particular, acabamos sempre por agradecer e contar com as mesmas pessoas. Não pretendo que esta afirmação seja conotada no sentido negativo, pelo contrário, pretendo sim realçar a importância e o peso destas pessoas nas nossas vidas e no quanto podemos contar com elas para todas as ocasiões.

Sendo assim, gostaria de guardar em poucas palavras, neste pequeno espaço, exatamente aqui, no início e fim desta pequena e grande etapa, toda a minha enorme gratidão pelo suporte que recebi da minha família, incluindo pais (Aida e Jorge), irmãos (Soraia e Gabriel) e avós (António e Judite), da minha namorada (Márcia) e amigos próximos que estiveram presentes e me apoiaram sempre nas decisões mais difíceis que surgem durante o percurso universitário e também na vida em geral. Com tudo isto, devo reforçar e não deixar esquecer que a família é tudo o que temos, é impossível definir o valor que lhe damos, tudo o que fazemos, fazemo-lo por nós próprios, mas também por ela.

Não poderia também deixar de apontar o papel importante que os professores que nos acompanham têm no desenrolar da nossa educação. Por pena, ainda desprezados pela nossa sociedade, não nos podemos esquecer que são eles que apoiam os nossos pais na nossa educação, que nos dão asas para podermos voar, que nos abrem a porta para um mundo de oportunidades. Sem eles, não saberíamos nada, não teríamos nada, seríamos ninguéns.

Graças a todos eles, consegui ultrapassar vários obstáculos que foram surgindo pelo meu caminho. Sei que sozinho provavelmente teria a alma e a força para o fazer, mas não seria a mesma coisa. Se assim o tivesse feito, sei que no fim, as coisas que obtemos, todos os objetivos que atingimos não têm o mesmo valor. Tudo o que realizamos ganha mais importância e um sentido de maior grandeza e objetivo quando o fazemos acompanhados pelas pessoas que mais gostamos.

RESUMO

A Realidade Aumentada (RA) caracteriza-se pela mistura de elementos virtuais no mundo real de forma interativa e em tempo real. O conceito de RA levanta uma ampla variedade de questões quanto à coerência visual entre os objetos reais e virtuais num ambiente. De forma a melhorar o processo de inclusão destes elementos no meio físico foram criadas várias técnicas e algoritmos de visão por computador que através do mapeamento de espaços físicos, extração de características e marcadores fiduciais de objetos, verificação, deteção, identificação, classificação, entre outros, permitem analisar e estruturar o conteúdo de uma cena.

O maior desafio que se coloca com a realização desta proposta de dissertação encontra-se associado à forma como é extraída e processada a informação que conseguimos obter a partir dos sensores que complementam os dispositivos de RA hoje em dia, a fim de representar e compreender, da melhor forma possível, os ambientes que nos rodeiam e preparar um espaço apto para a introdução e apresentação de conteúdo virtual com a maior harmonia.

Neste documento é possível encontrar o estado da arte relativo aos temas previamente citados a fim de explorar, melhorar e desenvolver novas técnicas e paradigmas para, a partir da informação dos sensores mais genéricos encontrados em muitas das tecnologias móveis e óculos de realidade aumentada mais atuais, extrair várias características do cenário e objetos envolventes em tempo real. O processamento e tratamento desta informação tem como objetivo final realizar o reconhecimento e compreensão da cena e objetos que se encontram no espaço que rodeia estes sensores.

Em paralelo à realização desta proposta de dissertação, foi desenvolvida uma framework denominada "*Tangible Environments in Augmented Reality Systems* (TEARS)" com o objetivo de demonstrar tudo o que é discutido neste documento não só como algo para fins de investigação científica, mas também para utilização e apoio num projeto e protótipo realizado no âmbito da unidade curricular do 5°Ano do Mestrado Integrado em Engenharia Informática (MIEI) de Projeto em Engenharia Informática (PEI) e que apresenta o título: "Assistência Remota com Realidade Mista (ARRM)".

Palavras-Chave: Visão por Computador, Realidade Aumentada, Análise de Cena, Contexto, Objetos, Reconhecimento

ABSTRACT

Augmented Reality (AR) is described as the mixing of virtual elements in the real world in an interactive way and in real time. The concept of AR raises many questions about the visual coherence between real and virtual objects in an environment. In order to improve the process of inclusion of these elements in the physical environment, a number of techniques and algorithms of computer vision have been created, which, through spatial mapping, extraction of characteristics and fiducial markers of objects, verification, detection, identification, classification, among others, allow us to analyse and structure the content of a scene.

The greatest challenge with this dissertation proposal is associated to how information, that we can get from the sensors that complement the AR devices today, is extracted and processed to better represent and understand our surroundings and prepare a suitable space that allows the introduction and presentation of virtual content with the greatest harmony.

In this document it is possible to find the state of art related to the before mentioned themes in order to explore, improve and develop new techniques and paradigms in a way that, from the information of the most generic sensors found in many of the most current mobile technologies and augmented reality glasses, we can extract various features of the scene and surrounding objects in real time. The stage of processing and treat this information has as its final goal the recognition and understanding of the scene and objects that are in the space that surrounds these sensors.

In parallel to this dissertation proposal, a framework called "Tangible Environments in Augmented Reality Systems (TEARS)" was developed with the intention of demonstrating everything that is discussed in this document not only for scientific research purposes, but also for use and support in a project and prototype carried out within the scope of the curricular unit of the 5th year of the Integrated Master's in Informatics Engineering (IMIE) named Informatics Engineering Project (IEP) and is titled: "Remote Assistance with Mixed Reality (RAMR)".

Key Words: Computer Vision, Augmented Reality, Scene Understanding, Context, Objects, Recognition

ÍNDICE

ANEXOS

LISTA DE FIGURAS

LISTA DE TABELAS

Lista de Abreviaturas, Siglas e Acrónimos

AIO	All-In-One
API	Application Programming Interface
AR	Augmented Reality
ARRM	Assistência Remota com Realidade Mista
CAC	Campo Aleatório Condicional
CAM	Campo Aleatório de Markov
CGI	Computer-Generated Imagery
FPS	Frames Per Second
HMD	Head Mounted Display
HPU	Holographic Processing Unit
HTTP	Hypertext Transfer Protocol
IEP	Informatics Engineering Project
IMIE	Integrated Master's in Informatics Engineering
IOU	Intersection Of Union
MAP	Mean Average Precision
MBD	Monitor Based Display
MIEI	Mestrado Integrado em Engenharia Informática
MSV	Máquinas de Suporte Vetorial
PEI	Projeto em Engenharia Informática
PR	Precision-Recall
RA	Realidade Aumentada
RAMR	Remote Assistance with Mixed Reality
REST	Representational State Transfer
RM	Realidade Mista
RNC	Redes Neuronais Convolucionais
RV	Realidade Virtual
SAAC	Sistemas Avançados de Assistência ao Condutor
SIT	Sistemas Inteligentes de Transporte
SMT	Sistemas de Monitorização de Tráfego

ST	See Through
TCP	Transmission Control Protocol
TEARS	Tangible Environments in Augmented Reality Systems
UC	Unidade de Controlo
ULA	Unidade Lógica e Aritmética
VA	Virtualidade Aumentada
WoW	Window On the World
YOLO	You Only Look Once

1. INTRODUÇÃO

1.1 Motivação e Objetivos

Em 1994, Paul Milgram, definiu uma taxonomia para ambientes virtuais e ambientes reais onde propôs o que ele designou de "*Virtuality Continuum*", ou "Contínuo de Virtualidade" [2]. Esta teoria designa um contínuo cujos extremos são o ambiente completamente virtual, a Realidade Virtual (RV), e o ambiente completamente real. Entre os dois extremos encontra- se o que ele denominou como sendo a Realidade Mista (RM), que define-se como sendo a combinação do mundo virtual com o mundo real, no qual interagem objetos físicos, pessoas reais e possíveis objetos virtuais onde a quantidade de elementos que são incluídos nesta "mistura" poderá variar.

Neste contexto, a RA caracteriza-se pelo predomínio do real sobre o virtual, ou seja, o utilizador visualiza o mundo real onde são inseridos objetos virtuais. No "Contínuo de Virtualidade" pode considerar-se que este tipo de RM se encontra mais próximo do extremo referente ao mundo real. A RA tem sido utilizada com enorme sucesso em diversas áreas, incluindo a medicina, o entretenimento, o comércio, a indústria de produção, a arqueologia, o turismo e a educação.

Esta filosofia de funcionamento tecnológico caracteriza-se pela adição de informação no mundo real, através da inclusão de objetos virtuais numa cena e ambiente real. A fidelidade dos modelos geométricos utilizados, incluindo a sua posição, movimento e outros possíveis métodos de interação com este "mundo virtual" são sem dúvida importantes para a obtenção de uma visualização e lógica fidedigna desses objetos virtuais. Em diversas áreas de aplicação, a coerência visual do espaço e ambiente encontrado em RA é extremamente importante, visto que os objetos virtuais devem surgir integrados na cena real com o maior realismo possível.

Um outro ponto que surge no cerne de todas estas novas tecnologias é a questão da deteção e reconhecimento de objetos, e a RA não foge deste padrão. O processo de reconhecimento define-se já como uma tecnologia no campo da visão por computador em que se pretende trabalhar com diferentes cenários e objetos que surgem não só em imagens e vídeo 2D, mas também em ambientes 3D.

O ser humano não apresenta qualquer dificuldade em detetar e verificar objetos, mesmo em situações em que os observamos em diferentes formatos e com propriedades distintas: posição, escala, perspetiva, iluminação, entre outros. Mesmo na eventual ocorrência de qualquer oclusão ou deformação do objeto, o ser humano continua a demonstrar facilidade na sua identificação. Quando traduzida e implementada em sistemas computacionais de visão, esta tarefa permanece um enorme desafio. Várias foram as técnicas e soluções criadas, testadas e reinventadas ao longo destas últimas décadas.

Com a realização desta dissertação pretende-se, pela utilização de sensores, obter e processar em tempo real, a maior quantidade de informação possível do ambiente que nos rodeia. Todos estes dados, depois de trabalhados e bem estruturados, têm como objetivo final permitir, não só, um melhor conhecimento da cena que nos rodeia através da identificação dos diferentes objetos e contexto da cena, mas também, permitir uma melhor coerência no processo de inclusão de objetos virtuais no mundo real.

1.2 Estrutura do Documento

Inicialmente será realizada uma revisão da literatura de modo a ter um conhecimento mais alargado sobre as temáticas em questão e dos trabalhos de investigação realizados na área, procurando assim identificar abordagens, métodos e técnicas que possam ser aplicados no âmbito desta dissertação e da mesma forma úteis para o desenvolvimento do protótipo final referido. Para tal, será realizada uma pesquisa em alguns portais de conteúdos científicos, incluindo "*Scopus*", "*Web of Knowledge*", "*IEEE Xplore Digital Library*", "*Springer Link*", "*Research Gate*" e "*Google Scholar*". Na pesquisa serão utilizadas palavras-chave características da área em estudo, tanto em português como em inglês. Na seleção dos artigos e trabalhos a ter em conta serão considerados aspetos como o ano, o título, os autores e o número de citações.

Tendo por base a revisão de literatura e o estudo das abordagens, métodos e técnicas relevantes, será proposta e adaptada uma metodologia para extrair, processar e estruturar as características mais percetíveis de um ambiente físico, tendo sempre por base os dados do sensor utilizado para tal efeito. No final pretende-se medir a rapidez e grau de confiança obtido na identificação da cena, contexto e dos múltiplos objetos 3D que encontramos na mesma, quando alimentamos um dado sistema para reconhecimento de imagens com estas mesmas características compostas (e.g Redes Neuronais Convolucionais (RNC), Máquinas de Suporte Vetorial (MSV), etc.).

Para demonstrar esta metodologia, será implementada uma plataforma de RM com o título "*Tangible Environments in Augmented Reality Systems (TEARS)*" de modo a expor as técnicas científicas que serão discutidas, juntamente com uma biblioteca para interação, processamento e acesso aos dados brutos provenientes dos sensores do *hardware* em questão, *HoloLens* da *Microsoft*.

Por fim será realizada uma prova da metodologia desenvolvida e poder de processamento deste *hardware*. Caso se conclua que este tipo de *hardware* não se encontra pronto para realizar tais tarefas de aprendizagem máquina, ou surja uma outra qualquer razão que o justifique, será desenvolvido uma arquitetura cliente-servidor em que se apresente um serviço de reconhecimento que irá ser disponibilizado via *web*, utilizando uma máquina que apresente as especificações mínimas necessárias para correr estes trabalhos de forma mais rápida e eficiente. Ou seja, a obtenção de dados continuará a ser realizada por parte do dispositivo do *HoloLens* (Cliente), mas o processo de interpretação, processamento e reconhecimento será realizado num outro sistema remoto (Servidor).

A plataforma *Unity 3D* será a utilizada como base de programação e motor gráfico deste projeto, sempre apoiada por vários *scripts* e processos que irão correr nativamente em ambientes de linguagem C, C++ e *Python*, juntamente com algumas bibliotecas de RA e de processamento de imagem, que inclua algoritmos e algumas técnicas genéricas de visão por computador (e.g *OpenCV*). Adicionalmente será utilizado um outro conjunto de ferramentas com o objetivo de desenvolver uma plataforma de comunicações via web através da disponibilização de uma interface do estilo *RESTful* e/ou *WebSockets* (e.g *Node Javascript*).

2. ESTADO DE ARTE

2.1 Visão por Computador

Não existe certamente uma definição única para o que realmente é a visão por computador, mas numa primeira tentativa foi possível desvendar o seguinte:

"*Visão por computador, compreensão de imagem ou análise de cena é uma combinação de processamento de imagem, reconhecimento de padrões e tecnologias de inteligência artificial que concentra-se na análise computacional de uma ou mais imagens, obtidas, individualmente ou em sequência no tempo, pela utilização de um sensor individual ou multiespectral. O processo de análise reconhece, localiza a posição e orientação, e apresenta um reconhecimento ou descrição simbólica o suficientemente detalhado para as imagens de objetos que despertam interesse no espaço tridimensional. O processo de visão por computador utiliza maioritariamente modelação geométrica e representações complexas de conhecimento numa metodologia baseada em expetativa, correspondência ou procura. O processo de pesquisa poderá utilizar estratégicas hierárquicas ou heterárquicas: "bottom up", "top down", "blackboard", etc.*" [3]

Esta foi a definição dada por Robert M. Haralick[1] no ano de 1990, atual professor da disciplina de Ciências de Computação na Universidade da Cidade de Nova Iorque e, na altura da redação deste documento, o último condecorado com o prémio de distinção *King-Sun Fu* na Conferência Internacional de Reconhecimento de Padrões em Cancun, México, no dia 4 de dezembro de 2016[2]. As suas principais contribuições aplicam-se na área de análise de imagem e incluem deteção remota, análise de texturas, morfologia matemática, classificação consistente e avaliação do desempenho de sistemas.

Esta definição em especial deve ser considerada, tendo em conta o prestígio apresentado por este mesmo professor e pelo facto de ser citado inúmeras vezes noutros livros e artigos pesquisados [4] [5] [6] [7] que concentram o seu trabalho nas mesmas áreas de interesse.

---

[1] http://haralick.org/

[2] https://www.gc.cuny.edu/Page-Elements/Academics-Research-Centers-Initiatives/Doctoral-Programs/Computer-Science/Program-News/Detail?id=37244

Facilmente encontramos inúmeras definições para a temática que se encontra a ser discutida nesta secção, mas antes de prosseguirmos com este processo de definição e exploração dos vários campos de estudo e aplicação desta área científica, uma questão deve ser colocada: "*Porquê estudar a visão por computador/visão artificial?*".

A nossa visão é provavelmente o sentido mais poderoso que o ser humano dispõe e que permite obter informação sobre o meio circundante: determinar a posição dos objetos, a relação entre os objetos numa dada cena, identificar os objetos e de certa forma interatuar de forma inteligente com o mundo, sem sequer haver a necessidade de contacto físico direto. A compreensão dos sistemas de visão biológicos é ainda muito limitada, existindo assim a necessidade de criar e explorar todos estes campos e vertentes de estudo. Todos eles apresentam o objetivo de compreender o processo de perceção que integra o ato da visão.

Segundo o paradigma definido por David Marr [8], em 1982, para conseguirmos compreender este processo e todas as metodologias que o suportam, é necessário realizar uma análise em três níveis distintos:

1. Teoria Computacional

Surge como a etapa e ponto mais importante desta série e descreve-se como sendo uma ciência da matemática e computação que tem como objetivo determinar a quantidade de problemas que podem ser apurados através da definição matemática de um modelo que correlacione os dados observados (Imagens): O problema colocado tem uma solução? É possível definir uma função por meio de um dado algoritmo? É única?

2. Algoritmos e Estruturas de Dados

Etapa que compreende o desenho e definição dos algoritmos e estruturas de dados que, quando aplicados a um dado conjunto de elementos de entrada, produzem, no melhor dos casos, a saída desejada. Deve ser sempre considerada a lógica, a estabilidade, segurança e robustez dos sistemas que irão ser desenvolvidos durante a etapa final.

3. Implementação

Por último, surge a implementação do modelo, algoritmos, estruturas de memória e outras quaisquer arquiteturas definidas durante todo o processo. Quando executados numa máquina, deverão ser levados em conta todas as noções de eficiência computacional e distribuída, que incluem por exemplo, a escolha de processos em série ou paralelos, de forma a obtermos um sistema completo em todos os aspetos.

Para Marr, os objetivos da visão deverão ser sempre os seguintes: "*Partindo de uma imagem ou sequência de imagens de um objeto móvel ou estático ou de uma cena, adquiridas por um observador monocular ou multiocular, estático ou em movimento, a visão por computador pretende compreender o objeto em causa e as suas propriedades tridimensionais*".

Seria possível continuar a explorar e até mesmo discutir uma enorme variedade de definições, características e objetivos que são possíveis encontrar à volta da tão complexa temática que é a visão por computador, mas todas elas iriam ser semelhantes na sua forma e ser. De modo a manter atual esta dissertação à data em que foi realizada é possível encontrar de seguida uma pequena síntese pessoal de tudo aquilo que foi possível descobrir ao longo das várias pesquisas realizadas:

"*A visão por computador é deste modo uma área das ciências de computação que visa introduzir o poder da visão nos computadores, através de câmaras e outros sensores, de forma a poder analisar e processar imagens do mundo real tal como a visão humana assim o permite* [9]. *Esta disciplina estuda problemas metodológicos, algorítmicos e tópicos relacionados com a definição e implementação de soluções que vão de encontro com as necessidades apresentadas pelos vários setores da indústria e que adaptam-se na sua maior conformidade com o surgimento de novas tecnologias.*

No campo da visão por computador procura-se analisar, de uma certa forma exaustiva, a essência da imagem e obter todo um conjunto de características que não seriam facilmente alcançadas pelo ser humano, tudo com o intuito de responder a várias questões, como por exemplo: calcular a profundidade em imagens cujos dados se encontram estruturados e representados num mundo com apenas duas dimensões através de técnicas de estereoscopia, identificar o contexto e local onde uma data fotografia é retirada, contar o número de objetos numa cena, reconhecer uma face, uma pessoa ou

um objeto, descobrir qual a relação de um dado objeto ou apenas uma parte dele com os demais, detetar defeitos na produção de componentes eletrónicos, seguir o movimento e pose de uma multidão, monitorizar o tráfego automóvel, e muito mais.

A sua área de aplicação tem vindo a expandir nos últimos anos de uma forma drástica tendo em conta o progresso realizado nos sistemas de computação e o aparecimento de novos sensores, mas também pelos avanços que ocorreram nos conteúdos básicos e mais teóricos que alimentam as metodologias da visão por computador"

2.1.2 A Velha Realidade

Tendo em conta a definição apresentada, é de esperar que na área da visão por computador seja discutido uma variedade de assuntos que passam por todos os setores e áreas da sociedade em geral.

Na tentativa de encontrar uma resposta à questão: "*Onde se encontra a visão por computador hoje em dia?*" decidiu-se definir como ponto de partida a análise realizada por William T. Freeman no início de 2011, sobre algumas das discussões que foi realizando com colegas, professores e investigadores durante as conferências [10] em que participou e foi possível verificar a enorme necessidade que existe em utilizar a visão por computador para responder a questões que, ainda hoje, são colocadas no âmbito da investigação científica em geral.

Embora iniciemos com a exploração de documentos e tópicos já um pouco desatualizados, é importante apontar que o objetivo desta secção é exatamente compreender qual foi a evolução e alterações que ocorreram, nesta última década, nos conteúdos que são discutidos pelos grupos científicos na área da visão por computador.

1. Procura do vizinho mais próximo em dados de grandes dimensões

A tarefa de reconhecimento de objetos em visão por computador pode ser formulada como um problema de procura do vizinho mais próximo, no qual se pretende encontrar o conjunto de classificações de treino que melhor correspondem às características extraídas duma imagem. Os elementos que se

procuram corresponder podem ser características, partes da imagem ou até mesmo imagens completas, mas são, em quase todos os casos, dados de grandes dimensões.

Embora existam métodos que apresentem soluções para resolver esta questão [11] [12] [13], é sempre necessário obter resultados com maior rapidez e exatidão para o tratamento deste tipo de estruturas, já que, tal feito, dá um impulso direto em muitos dos algoritmos utilizados em visão por computador.

A maioria dos investigadores com que William T. Freeman teve a oportunidade de discutir, procuravam soluções que conseguissem responder ao problema introduzido pela existência de dados de grande dimensão que são precisos processar para a realização de muitas das tarefas de reconhecimento apresentadas na área de visão. Uma das razões que contribui para a existência destas estruturas de dados complexas, deve-se ao facto de que a maior parte da informação utilizada para estudos e testes surge de conteúdo bruto obtido de várias fontes na internet. Este problema surge da necessidade inerente em utilizar estruturas mais pesadas ao nível do seu alojamento e que adicionam informação vital para a utilização de alguns algoritmos em específico, mas por vezes, pode também ser devido à falta da própria informação, surgindo assim a necessidade de fazer algum tipo de preparação e tratamento prévio. A introdução de qualquer tipo de atraso nestes processos impede muitas vezes de realizar as tarefas em tempo real.

William T. Freeman deixou assim o alerta de que existe a necessidade de escalar as soluções existentes, em especial, procurar generalizar [14] muitas das estruturas de classificação e categorização de cenas e objetos utilizadas hoje em dia com o intuito de reconhecer milhares ou talvez mesmo dezenas de milhares de categorias com maior eficiência [15]. Muitos problemas da visão por computador resultam na criação de aplicações informáticas que tratam modelos de programação inteira, linear e não linear quadrática para questões que apresentam estruturas de dados de grandes dimensões. Já naquele período, as soluções tomadas como padrão já não funcionavam e continua a ser necessário explorar novas soluções que adaptem-se e tenham em conta o surgimento deste tipo de estruturas com informação um tanto vasta e esparsa.

2. Averiguar o contexto de uma cena

De uma forma geral, o processamento realizado sobre imagens visa identificar e classificar os objetos que se encontram numa dada cena, mas por vezes é de interesse procurar contextualizar e categorizar a cena em questão. Para a realização desta tarefa surgem vários artigos científicos, projetos e ferramentas desenvolvidas que referem a utilização de Campos Aleatórios de Markov (CAM) como uma solução, sendo também uma área de estudo com um elevado crescimento nas áreas de ciências da computação.

Em imagens do mundo real, as diferentes regiões que podem existir são normalmente homogéneas e os pixéis vizinhos de um qualquer ponto aleatório apresentam, de uma forma um tanto regular, as mesmas características e propriedades. Um CAM é um modelo probabilístico normalmente utilizado para descrever imagens, tendo em conta que vai de encontro com estas mesmas condições e apresenta uma descrição simples e modular. É de notar que um CAM também é utilizado dentro de outras áreas e contextos. Como o trabalho realizado na área de visão por computador utiliza a imagem como unidade base, na sua forma mais básica, a estrutura utilizada consiste numa grelha ou matriz.

Se x_i for um estado desconhecido no nodo i e y_i uma hipótese apresentada para o mesmo, a probabilidade conjunta sobre o CAM é

$$P(\vec{x}, \vec{y}) = \prod_i \Phi(x_i, y_i) \prod_{i,j} \Psi(x_i, x_j)$$

onde o segundo produto é realizado nos nodos vizinhos, i e j. $\Phi(x_i, y_i)$ deve ser visto como um termo para prova e teste local. Quando as funções de compatibilidade, descritas aqui como funções de união $\Psi(x_i, x_j)$ também dependem das hipóteses y, entramos na definição de Campo Aleatório Condicional (CAC). Normalmente procuramos saber qual o conjunto de estados x_i em cada nodo i que maximizam a probabilidade conjunta $P(\vec{x}, \vec{y})$ para um dado conjunto de hipóteses $\vec{y}$, ou equivalente, que minimizem uma dada função das variáveis discretas $\vec{x}$.

Grandes progressos foram feitos [16] [17] [18] [19] [20] [21], mas o que é possível fazer no momento é limitado pelas técnicas existentes para resolver este problema e, segundo o autor, avanços

na área seriam utilizados de forma quase imediata pela comunidade. São necessários algoritmos eficientes para o processo de minimização das subfunções não modulares que nos assistem na procura e construção de novas soluções, uma vez que estas assistem na procura e cálculo dos limites da solução que é possível obter. Existe um benefício real em lidar com grupos de ordem superior nos CAM e CAC de forma a modelar melhor os problemas em vez de cairmos na tentação de simplesmente agregar técnicas e outras tarefas com inúmeras iterações e ciclos de processamento. William T. Freeman reforça que é mesmo necessário encontrar melhores formas de o fazer.

Como as restrições topológicas são frequentemente relevantes quando se trabalha com uma imagem, é necessário executar uma otimização discreta sobre a função apresentada acima, tendo em conta estas mesmas restrições. Por exemplo, poderemos necessitar de especificar que todos os estados que tomam uma classificação específica, dentro de certos limites de uma área da imagem, devem ser conectados e relacionados ou que uma certa sub-região da imagem, especificada pelo utilizador, toma algumas propriedades que a ajudam a aproximar-se com um outro conjunto de classificações [19] [22].

Faltam formas eficientes que permitam realizar esta tarefa ou até mesmo conseguir perceber quais são os limites do desempenho das ferramentas e soluções atuais. Na realização de modelos gráficos, trabalha-se geralmente com dois tipos de restrições: restrições de estrutura, como a planaridade ou a largura da árvore, e restrições de linguagem, como o nível de modulação ou convexão. Muitos investigadores reforçam a utilidade em juntar estes dois elementos.

3. Modelos de Estatística para Imagens

Os CAM são modelos probabilísticos realizados sobre a imagem e apresentam uma ampla aplicação no processo de interpretação e aprimoramento da mesma. Embora muitas descobertas já tenham sido realizadas sobre este mesmo campo, investigadores encontram-se agora estagnados e um tanto inquietos em querer desenvolver novos modelos e representações que se apresentem úteis para o estudo da imagem. Outros modelos de estatística sobre imagens foram desenvolvidos [23] [24] [25], mas avaliando a sua utilidade no que diz respeito ao processo de síntese de novas imagens, os melhores modelos continuam a ser os algoritmos de sintetização de texturas que não necessitam de quaisquer parâmetros [26] [27].

Uma vez mais, foram levantadas questões por parte dos investigadores, no qual pretendem perceber se existe a possibilidade de dar algum salto nesta área de estudo, se é mesmo necessário realizar a amostragem de imagens já existentes para conseguir criar uma nova, se é possível definir uma estrutura tal que permita controlar o processo de sintetização destes algoritmos não parametrizados, etc. Muitos dos progressos realizados neste campo devem–se às características e descritores de imagem, mas muito mais é preciso fazer.

2.1.3 A Nova Atualidade

Tendo em conta os vários problemas delineados na subsecção anterior, falta ainda compreender, de forma um tanto resumida, qual a evolução sentida nestes últimos anos, com principal foco na área da visão por computador. Igualmente, tentaremos ainda perceber o quanto dessa "velha" realidade ainda é discutida nos dias de hoje e qual a nova moda que se tornou tendência nas temáticas e conteúdos explorados pelos grupos científicos e os seus investigadores.

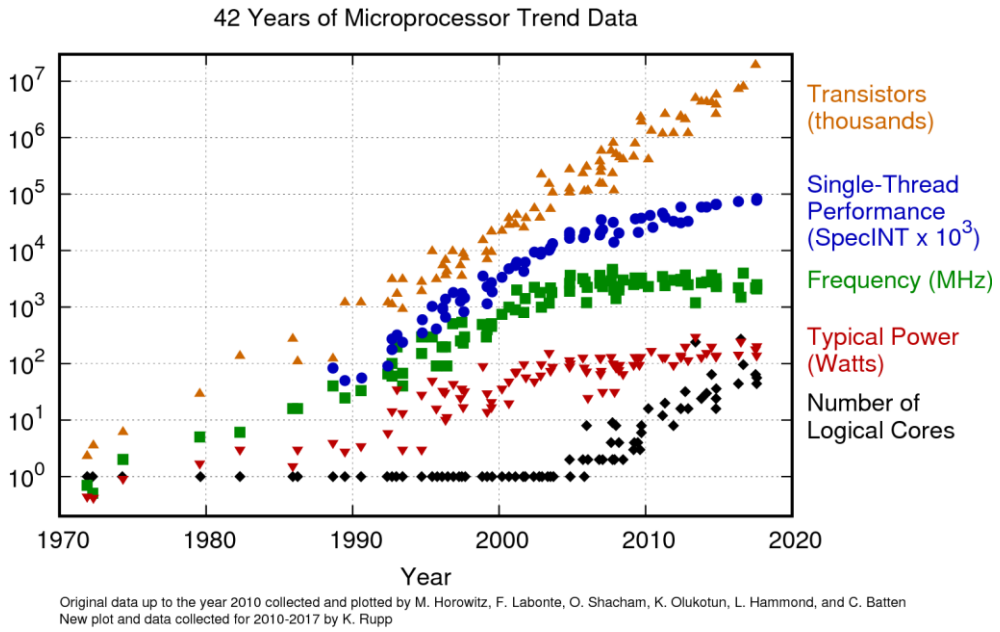

Figura 1 - 42 anos de dados estatísticos na quantidade de transístores, desempenho, frequência, consumo energético e número de *cores* lógicos de vários microprocessadores existentes no mercado[3].

[3] https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/

Provavelmente um dos pontos mais importantes e revolucionadores dos últimos anos será aquela que se refere à própria evolução tecnológica sentida devido ao surgimento de novos sistemas, arquiteturas e *hardware* cada vez mais eficiente e robusto ao nível do seu funcionamento e poder de processamento que daí é possível extrair, tal como é possível concluir das estatísticas e valores apresentados na Figura 1. Tal como já seria de prever, o crescimento e a melhoria a um nível mais técnico, físico e metodológico, dos componentes (e.g Unidade Lógica e Aritmética (ULA), Unidade de Controlo (UC), Registos em memória, etc.) que constituem as principais unidades de processamento distribuídas que conhecemos (e.g CPU, GPU, etc.), continua a sentir-se de forma praticamente exponencial.

Adicionalmente, o desenvolvimento e a utilização de novos materiais, a própria capacidade e quantidade da produção e o grau de facilidade no acesso a estes produtos por parte do mercado consumista, são melhores e maiores, em todos os sentidos. Tudo isto para referir que temos acesso a um maior poder de processamento em troca de algo que ocupa cada vez menos espaço, consome menos energia, é mais resistente e que no fim, custa menos para o consumidor final.

Por esta mesma razão, tem sido possível aos grandes cientistas e inventores da nossa geração construir estruturas cada vez mais simples e de uma forma mais direta e bruta pelo aumento, em quantidade, destas pequenas unidades, mas ao mesmo tempo tem vindo a tornar-se uma tarefa mais complexa devido à necessidade de modelar, definir e introduzir novas arquiteturas estruturais e protocolos de comunicação que necessitam de ser mais inteligentes, eficientes e seguros para permitir uma melhor troca de mensagens, distribuição de processamento e partilha de memória entre estes pequenos "*trabalhadores*".

Esta melhoria ao nível do *hardware* e acesso a um maior poder de processamento leva a que muitos dos algoritmos e técnicas de visão já conhecidas ganhem, de forma quase grátis, uma maior escalabilidade na sua aplicabilidade e eficiência. Hoje em dia, processos de computação de alto nível que necessitam de percorrer estruturas de grandes dimensões para a extração ou transformação de informação para a obtenção de dados específicos e certas características, encontra-se assim, desta forma, a ser resolvida com os avanços tecnológicos que tem vindo a sentir-se.

A realidade que é apresentada na subsecção anterior tem já 7 anos de idade, e embora muitos dos seus conteúdos continuem a ser discutidos na atualidade, os problemas de hoje fixam-se principalmente nas áreas da visão por computador que têm como objetivo estudar os mecanismos por detrás do processamento e essência da própria imagem a um nível muito mais baixo. Estas áreas mais específicas da visão por computador preocupam-se assim em procurar definir e estudar novas características que sejam possíveis retirar de uma imagem e delas constituir novos descritores de imagens (Normalmente representados por imagens também) [28].

O tipo de análise que aqui se pretende realizar, normalmente, não tem acesso a nenhuma informação sobre o contexto ou objetos que se encontram presentes numa dada cena, nem sequer onde a mesma se encontra em relação ao próprio observador. Neste momento é possível ter acesso a múltiplos descritores, totalmente independentes uns dos outros, entre os quais, fragmentos de linhas e contornos, manchas, reflexões, etc.

Um exemplo para melhor compreender o conceito seria o seguinte: considere-se uma cena em que um utilizador olha na direção de uma caneca de café que se encontra disposta na sua secretária, os descritores de baixo nível possibilitam a definição explicita de onde os contornos e forma dessa caneca se encontram, onde as luzes especulares melhor se definem nesse corpo, quais as cores presentes na sua textura, etc.

É importante referir que, como os descritores se encontram diretamente interligados com a imagem a que pertencem como um todo, estes podem ser aplicados a qualquer área e elementos que encontram-se presentes na imagem, não estando apenas limitados à caneca, como é demonstrado no seguinte exemplo pela Figura 2.

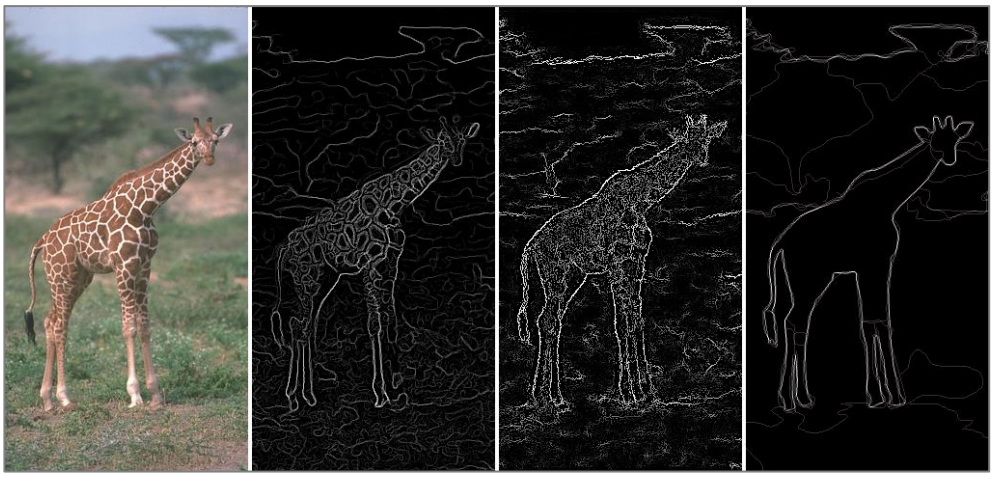

Figura 2 - Exemplo de análise realizada sobre a imagem de uma girafa para deteção de limites e contornos por utilização do operador e filtro denominado "*Canny Edge Detection*", juntamente com algumas melhorias introduzidas pelo algoritmo "*Sketch Token*" [4].

Para concluir esta secção, gostaria ainda de nomear outras tecnologias e sistemas inteligentes que se encontram na moda. Encontram-se definidos dentro dos tópicos de *machine* (*deep) learning* e têm sido alvo e foco de vários estudos e artigos no meio científico, sendo eles também utilizados como ferramentas na análise de imagem e surgem em várias etapas, processos e tarefas do reconhecimento de imagem, nomeadamente: *Convolutional Neural Networks* (CNNs) e *Support Vector Machines* (SVMs), os quais despertam algum interesse para o desenrolar desta dissertação e surgem como matéria para estudo e apoio no desenvolvimento do projeto proposto no capítulo 3-TEARS-Tangible Environments in Augmented Reality Systems, presente neste mesmo documento.

2.1.4 Alguns Problemas por Resolver

Mesmo com todo o sucesso que surge das descobertas realizadas nas áreas das ciências de computação e, querendo ser um pouco mais em específico, no campo da visão por computador, é sempre necessário procurar evoluir e melhorar algo. O ser humano é, de certa forma, um ser insaciável.

É possível encontrar várias falhas e a necessidade de uma qualquer melhoria até mesmo nos melhores algoritmos e técnicas de visão que existem no momento. Por exemplo, o grau de confiança que obtém-se no reconhecimento da face de uma pessoa numa dada imagem cai de forma exponencial no momento em que verificamos alterações na sua pose, introduzimos faces não frontais ou até mesmo

[4] http://cs.brown.edu/courses/cs143/2013/proj5/

quando surgem alterações repentinas nas condições da luz. Quando colocados num ambiente imperfeito, retratado aqui pelo mundo real, algoritmos estéreo podem facilmente falhar na reconstrução de mapas de profundidade em certas regiões de uma imagem se estas mesmas apresentarem espaços amplos e de grandes dimensões, tudo devido à existência de correspondências ambíguas pela falta de padrões e texturas que suportem as decisões que surgem durante o processamento das técnicas aqui utlizadas.

Pensa-se na possibilidade do ser humano conseguir reconhecer milhares de categorias de objetos [15], mas os computadores conseguem apenas reconhecer algumas delas com elevada confiança. É possível que um dos grandes obstáculos da visão por computador seja a variabilidade, para o qual existem demasiadas fontes de informação e questões que precisam de uma resposta. É o sonho de qualquer cientista e investigador criar um único algoritmo que seja perfeito e que consiga responder a tudo o que seja necessário resolver.

Variações na iluminação, perspetiva e a possível existência de oclusões alteram por completo a imagem que é observada. Ou seja, é possível ter várias versões do mesmo objeto e para isso basta alterar e variar uma das propriedades que pode apresentar: posição, tamanho, rotação, cor, etc. Indiferentemente da forma como o objeto se apresenta e dispõe, o ser humano consegue reconhecer-lo, ao contrário dos computadores, que apresentam várias dificuldades. Da mesma forma que distinguimos diferentes objetos, também apresentamos a mesma facilidade no que toca ao reconhecimento dos diferentes materiais e texturas que os complementam. O ser humano consegue ignorar qualquer variação que ocorra nas condições do ambiente que o rodeia e perceber com elevada confiança as propriedades subjacentes a um qualquer material. De notar que esta capacidade, embora seja muito superior nos humanos quando comparada com as máquinas, também apresenta um certo limite.

2.1.5 Reconhecimento de Imagem

Uma forma um pouco mais simples, diferente e de certa forma um quanto mais criativa de introduzir alguém à temática do reconhecimento de imagem será iniciar a discussão com a seguinte questão: "*Consegues encontrar o Wally na seguinte imagem (Figura 3)?*".

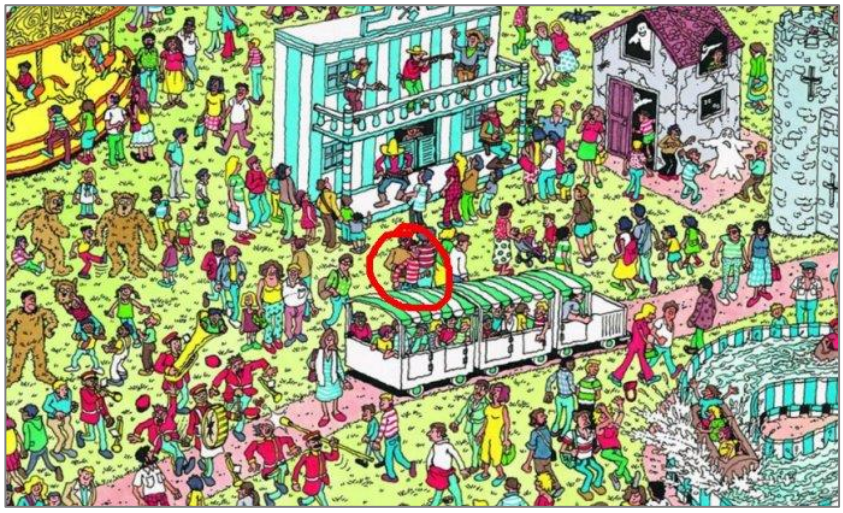

Figura 3 - Referência à séria de livros de caráter infanto-juvenil "*Where's Wally?*" criada pelo ilustrador britânico Martin Handford.

Num primeiro contacto com esta questão pode parecer uma pergunta um pouco descabida para o problema do reconhecimento de imagem, mas o simples processo de detetar a personagem do Wally nesta imagem retrata diretamente e de forma muito simplificada, uma das tarefas mais importantes que a visão por computador procura resolver[5].

Antes mesmo de procurar responder a esta questão é preciso compreender primeiro todos os mecanismos que se encontram por detrás do seu funcionamento. Um dos pontos que deve ser discutido retrata a hierarquia da cena e em perceber como ela deve ser estruturada, tendo em conta todos os objetos, elementos e pequenas características que possam surgir nela. Ou seja, necessitamos de separar e organizar toda a informação que extraímos diretamente de uma imagem, realizar um processo de classificação e depois de categorização, tudo com a finalidade de conseguirmos processar e pesquisar nesta estrutura composta e complexa, de forma a obter todo o tipo de respostas através de um processo mais rápido e eficiente.

[5] http://www.sciencealert.com/a-computer-algorithm-can-now-solve-where-s-wally-faster-than-you-can

Um outro ponto que deve ser explorado passa por perceber exatamente quais as técnicas e metodologias por detrás da magia que permite realizar tal feito. Precisamos de perceber, "ao pixel", como é resolvida esta questão, quais as diferentes soluções e técnicas que existem hoje em dia e que são alvo de novos avanços na área de visão por computador.

No final de tudo, é importante perceber que nem tudo são objetos, a cena também é um tópico muito importante a retratar. Muitos investigadores colocam demasiado peso e poder sobre o objeto e à sua própria existência numa dada cena, mas talvez seja importante anteceder o próprio conhecimento do contexto da cena antes de realizar o processo de reconhecimento dos objetos que aí se encontram [29].

Este tipo de questões são exatamente um dos alvos da pesquisa e estado de arte realizado nesta dissertação. São apresentadas algumas discussões e exemplos de aplicação ao longo das várias secções deste documento.

Sendo assim, para finalizar e resumir esta secção introdutória, ficam aqui as tarefas e objetivos que devem ser sempre considerados durante o processo de reconhecimento de imagem.

1. Verificação - "Aquilo é um carro?"

Figura 4 - Verificação realizada sobre uma imagem. Local: Avenida da Liberdade em Braga.

2. Deteção - “Existem pessoas neste espaço?”

Figura 5 - Deteção realizada sobre uma imagem. Local: Universidade do Minho em Braga.

3. Identificação - “Aquela é a Sé Catedral na Guarda?”

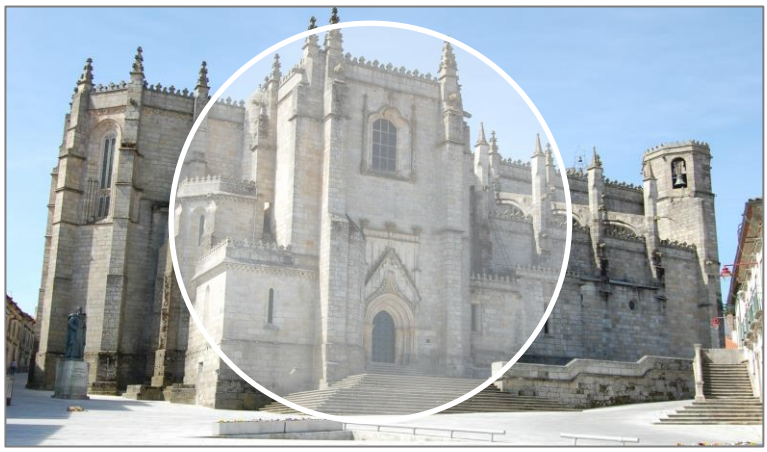

Figura 6 - Identificação realizada numa imagem. Local: Sé Catedral na Guarda.

4. Categorização de Objetos

Figura 7 - Categorização de objetos numa imagem. Local: Paços do Concelho no Porto.

5. Categorização do Contexto e da Cena

Figura 8 - Categorização do contexto e cena numa imagem. Local: Serra da Estrela na Guarda.

6. Atividade numa Cena

Figura 9 - Listagem de ações que se encontram a decorrer numa imagem. Local: Praça do Comércio em Lisboa.

Após visualizarmos todos estes exemplos, chegamos a um ponto em que também nos questionamos: "*Mas se já conseguimos fazer isto tudo, que problemas faltam tratar?*"

7. Múltiplas Perspetivas

Figura 10 - O problema das múltiplas perspetivas retratado em esculturas realizadas por Michelangelo.

8. Escala

Figura 11 - O problema da escala representado por duas maças de tamanho diferente.

9. Iluminação

Figura 12 - O problema das variações de iluminação retratado numa sessão de fotos intitulada "*Perception is Reality*"[6].

[6] http://designtaxi.com/news/368238/Portraits-Taken-In-Different-Lighting-Reveal-Striking-Differences

10. Oclusão

Figura 13 - O problema da oclusão de objetos retratado num quadro de Rene Magritte denominado "*Carte Lanche*" em 1965.

11. Deformação

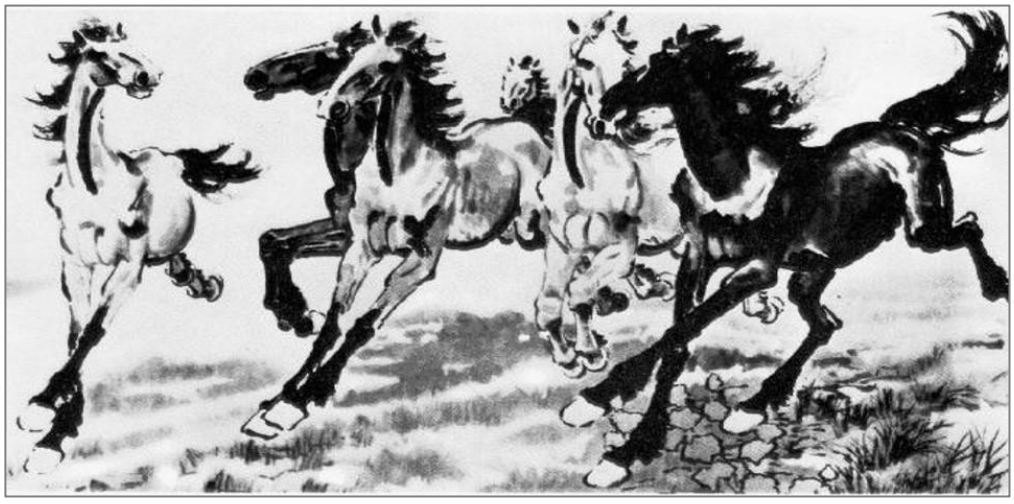

Figura 14 - O problema da deformação de objetos retratado numa pintura de Xu Beihong denominado "*Six Galloping Horses*".

12. Desordem

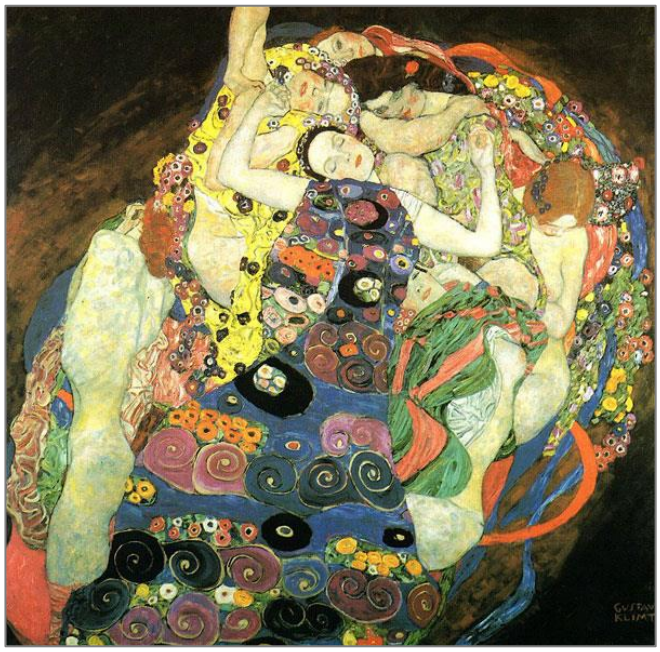

Figura 15 - O problema da desordem de objetos retratado numa pintura de Gustav Klimt denominado "*The Virgins*" em 1913.

13. Variação de Objetos da mesma Classe

Figura 16 - O problema da variação de objetos retratado por vários tipos de cadeiras.

2.1.6 Casos de Aplicação

1. Deteção de Faces

Num ambiente com boas condições existem aplicações informáticas que conseguem encontrar e detetar faces numa imagem tal como uma pessoa o conseguiria fazer. Aliás, câmaras mais modernas utilizam esta mesma capacidade de forma a controlar valores numéricos de exposição e foco da imagem [30] [31] [32] [33] [34] [35] [36].

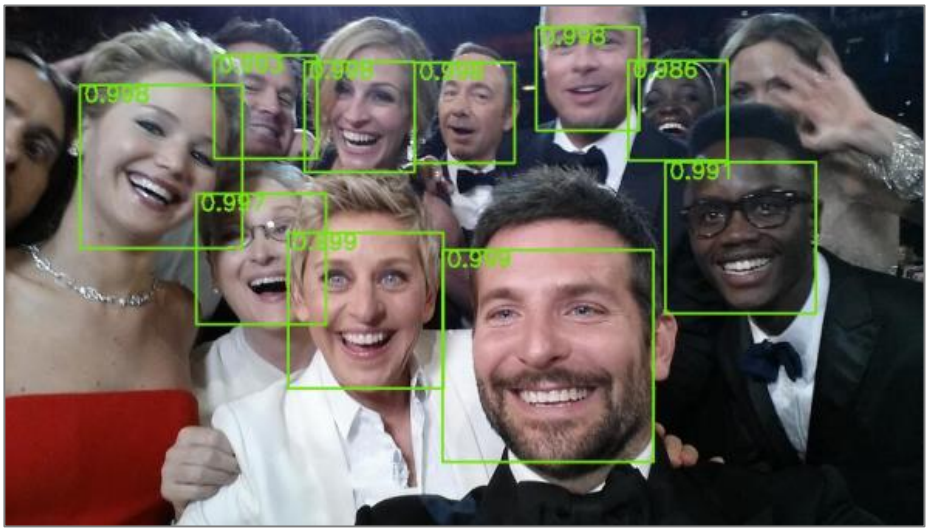

Figura 17 - Deteção de Faces - Fotografia tirada durante a entrega de Oscars do ano de 2014. Esta imagem contém várias faces que apresentam diferentes poses, iluminação e oclusões. As caixas e valores que são visíveis correspondem aos resultados obtidos pelo detetor desenvolvido por [30].

2. Sistemas Avançados de Assistência ao Condutor

Um outro exemplo muito atual encontra-se ligado à indústria automóvel que procura inovar na área da assistência e automação em viagem. Desde o início deste novo milénio que muitas empresas automóveis demonstram o seu interesse e foco na investigação e desenvolvimento de Sistemas Avançados de Assistência ao Condutor (SAAC) [37]. As suas funcionalidades passam por incluir sistemas de paragem de emergência que aplicam conceitos de previsão de trajetória e dispositivos de controlo motores e independentes em cada uma das rodas, alerta e notificações que permitem o controlo e permanência do veículo dentro da faixa de rodagem, piloto automático e condução autónoma, parqueamento autónomo, reconhecimento de sinalização na estrada, prevenção de colisão dentro de 360 graus de visão, deteção de peões, entre outros.

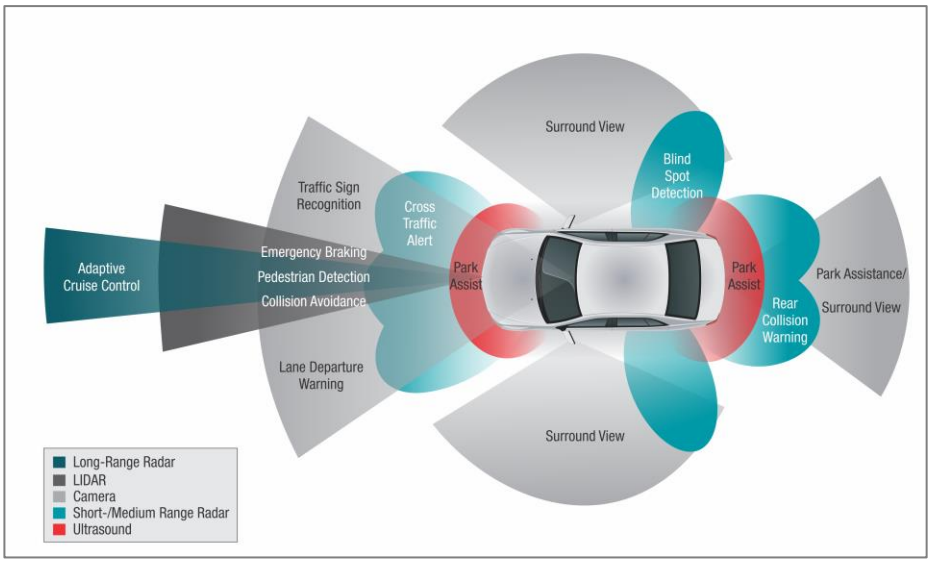

Figura 18 - Sistemas Avançados de Assistência ao Condutor - Descrição dos vários sensores e características apresentadas por esta tecnologia[7].

3. Deteção de Defeitos em Circuitos Eletrónicos

Outro tópico e possível área de aplicação do reconhecimento de imagem, na área de visão por computador, será a deteção e análise de irregularidades que podem ocorrer devido a deficiências no processo de montagem de circuitos eletrónicos. Este tipo de tecnologia e a habilidade de realizar a deteção e classificação de imperfeições nestes componentes tem vindo a ser utilizado pela indústria na produção em grande escala de peças e aparelhos elétricos.

Tendo em conta que uma linha de produção poderá produzir centenas, ou até mesmo, dezenas de milhares de *chips* num único dia, é importante garantir que este processo de reconhecimento seja efetuado em tempo real e de uma forma praticamente instantânea. Métodos convencionais utilizados para realizar esta tarefa conseguem ser lentos e pouco exatos na classificação do tipo de defeito que é verificado. Têm surgido nos últimos anos novos avanços e estudos que procuram colmatar a necessidade de criar sistemas mais rápidos e que apresentem maior grau de confiança na identificação de defeitos neste tipo de matéria [38] [39].

[7] http://aga.uk.net/advanced-driver-assistance-systems/

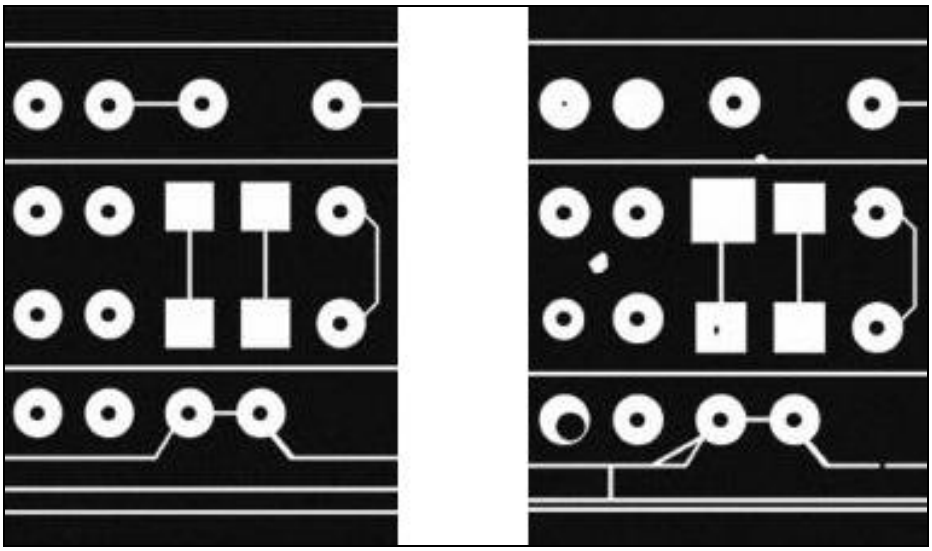

Figura 19 - Deteção de Defeitos em Circuitos Eletrónicos - Imagem binária de um circuito no seu estado perfeito (Esquerda) e com algumas imperfeições (Direita).

4. Monitorização Inteligente do Tráfego nas Estradas

Um último exemplo que gostaria de deixar nesta pequena introdução, relativamente a diferentes aplicações do reconhecimento de imagem, encontra-se mais uma vez interligado com os Sistemas Inteligentes de Transporte (SIT). Um dos elementos mais importantes deste grupo de tecnologias são os Sistemas de Monitorização de Tráfego (SMT) que são uma visão mais global daquilo que os SAAC fazem, adotando uma análise mais forte sobre o estado geral do tráfego e não num único elemento em específico [40].

A utilidade deste tipo de sistemas reflete-se, por exemplo, na extração de dados estatísticos sobre o estado das estradas, classe de veículos e outro tipo de anotações, através da adoção de sistemas inteligentes de controlo de semáforos, monitorização de eventos, trabalhos realizados para manutenção das estradas, congestionamentos, acidentes, etc.

Figura 20 - Monitorização Inteligente do Tráfego nas Estradas– Anotações sobre os respetivos intervenientes que surgem no fluxo do tráfego nas estradas[8].

5. Deteção de Objetos

A criação de modelos de aprendizagem que apresentam um elevado grau de confiança e eficiência no processo de classificação de múltiplos objetos numa única imagem ou até mesmo em vídeo ou em tempo real continua a ser um dos principais desafios da visão por computador. Um bom exemplo que pode ser aproveitado para demonstrar os avanços realizados dentro desta temática é a *framework* e *API* de Deteção de Objetos que foi criada tendo como base a ferramenta e plataforma *TensorFlow*[9], construída pela *Google* para fins de investigação científica em *machine learning*.

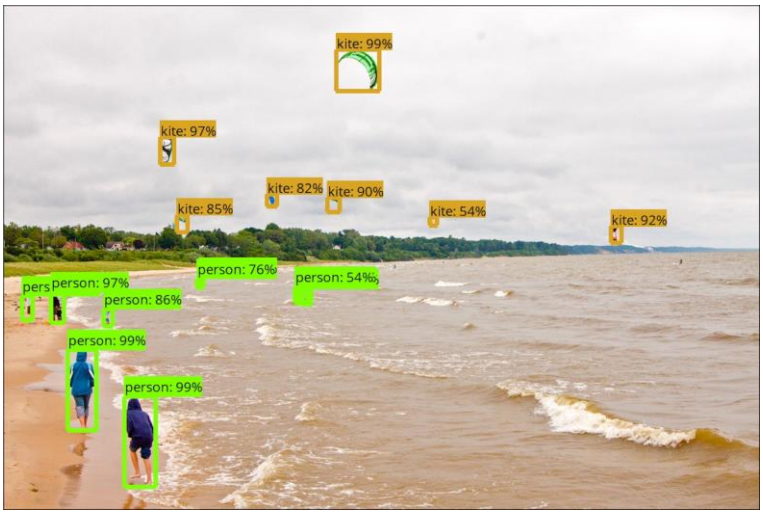

Figura 21 - Deteção de pessoas e *kites* utilizando a *API* de Deteção de Objetos da *Google* num cenário aberto de uma praia.

[8] https://www.linkedin.com/pulse/intelligent-traffic-management-applying-analytics-internet-praboo/
[9] https://www.tensorflow.org/

2.1.7 O que o futuro reserva?

Já é notável o facto dos termos "Visão por Computador" e "Reconhecimento de Imagem" serem vistos quase como sinónimos, mas a área de visão é muito mais do que apenas a imagem. O processo de visualizar algo é muito mais do que apenas extrair dados, é um processo cognitivo que envolve o poder da perceção, associação, memória, raciocínio, juízo, pensamento e linguagem, juntamente com toda a análise discutida nesta última secção. O ser humano utiliza aproximadamente dois terços do seu cérebro para realizar o processamento visual [41], por essa mesma razão, não seria de estranhar que fosse necessário algo mais do que apenas o reconhecimento de imagem para os computadores conseguirem simular perfeitamente o nosso sistema de visão.

Claro que, o reconhecimento de imagem por sí só, não deve ser desvalorizado por razão nenhuma. Este processo é uma parte fulcral para o funcionamento de todos os mecanismos da visão por computador que envolva o reconhecimento de tudo, desde texto, objetos e pessoas até ao contexto e lógica da atividade ocorrente numa cena. Mas, tal como o cientista Serge Belongie referiu durante a *LDV Vision Summit* em 2016:

"*Hoje em dia, no seu núcleo, o reconhecimento de imagem consegue apenas identificar objetos básicos como uma banana ou uma bicicleta numa imagem. Até as crianças conseguem fazer isso, mas o potencial da visão por computador deve ser sobre-humano: poder ver claramente no escuro, através das paredes, a longas distâncias, e processar todos esses dados de grandes volumes o mais rápido possível*"

Atualmente, tudo o que é explorado pelas áreas de visão por computador é completamente integrado na nossa vida, seja no nosso dia-a-dia ou em negócios, sempre com a finalidade de realizar todo o tipo de funções, incluindo por exemplo: alerta de pessoas ou animais na estrada para auxílio dos condutores, identificação de doenças em imagens de pacientes hospitalares, identificação de produtos e onde poder comprar-los, apresentação de anúncios inteligentes pela deteção do contexto em que são inseridos, entre outros. A tecnologia é complexa e, tal como todas os tarefas acima mencionadas, requer mais do que apenas o reconhecimento de imagem, é necessário ter em conta toda uma análise semântica e adaptar ou criar novos algoritmos para o processamento de grandes estruturas de dados impostas pela nova era do *big data*.

Tendo em conta esta pequena discussão, para além do reconhecimento de imagem, o que é que se passa atualmente na área de visão por computador? Para que mais poderá ela ser utilizada? Aqui estão alguns exemplos e as tecnologias que o permitem:

1. Utilização de Imagem Térmica

Figura 22 - Exemplo de visualização de uma imagem térmica de uma bicicleta em corrida retirada com um sistema móvel FLIR[10].

O ser humano não consegue ver o calor ou qualquer tipo de gás. Situações mais específicas onde se possa despoletar um fogo, surgir predadores ou existir qualquer fuga de gás, é da preocupação do ser humano conseguir visualizar estes perigos antes mesmo de os conseguir cheirar ou palpar. Avanços realizados na área de imagem térmica demonstram que esta capacidade já foi introduzida não apenas em câmaras portáteis com a finalidade de serem utilizadas nos vários sectores da indústria, mas também em *smartphones* pessoais, conforme demonstrado pelo *Cat S60*[11], o primeiro *smartphone* a introduzir este tipo de tecnologia. Eventualmente, é possível que este tipo de capacidade venha a ser integrada em todos os telemóveis.

Mas este tipo de perigos não são as únicas preocupações do ser humano, nem são os únicos exemplos onde a imagem térmica pode apresentar utilidade. Este tipo de avanço tecnológico poderá

[10] https://www.flir.com/
[11] https://www.catphones.com/en_us/cat-s60-smartphone.html

ajudar a manter o desporto um pouco mais honesto, tal como é evidenciado pelas câmaras de imagem térmica que foram utilizadas para detetar o *doping* mecânico no *Tour de France* do ano de 2016[12].

2. Sensores Inteligentes

Figura 23 - Arquitetura simplificada com algumas funcionalidades das tecnologias *Beacon* e como elas estão a conquistar o mercado[13].

Sensores que detetam a temperatura, luminosidade, qualidade do ar, gás e movimento são apenas uma amostra de tudo aquilo que a visão por computador utiliza para identificar exatamente o que se encontra lá fora no mundo real.

Nos dias de hoje, por exemplo, casas e edifícios inteligentes utilizam sensores incorporados nos seus sistemas de iluminação e controlo de temperatura para detetar o movimento das pessoas e poderem otimizar os níveis de luz e utilização de energia, definindo assim um sistema inteligente que possa aprender os hábitos e comportamentos das pessoas ao longo do tempo.

Além disso, os sistemas de monitorização doméstica não utilizam apenas sensores de movimento e câmaras de vídeo para rastrear movimentos suspeitos, crianças ou animais domésticos, mas também combinam este mesmo sistema com sensores de temperatura e de qualidade do ar para obter uma imagem muito mais completa daquilo que está a acontecer em casa enquanto estamos longe.

---

[12] https://arstechnica.com/gadgets/2016/06/tour-de-france-unveils-measures-to-catch-cheats-with-hidden-motors-neodymium-magnets/
[13] https://www.dacgroup.com/blog/the-beacons-are-taking-over/

Nas lojas, sensores e dispositivos *beacon* são integrados com as câmaras de vídeo com o objetivo de acompanhar o movimento do consumidor e cruzar os seus comportamentos com os dados que se encontram gravados em grandes bases de dados ou outros locais de armazenamento na nuvem. A finalidade destes sistemas é apoiar os comerciantes e retalhistas a melhorar e otimizar, não apenas, o *design* dos produtos e os preços das lojas, mas também atender à necessidade real dos clientes de forma inteligente[14][15].

2.1.8 Sumário

Muita informação útil poderá ser extraída desta curta iteração realizada sobre vários conteúdos que são neste momento discutidos na área da visão por computador e sobre aquilo que poderá vir a ser melhor discutido e explorado num futuro próximo. No que toca ao interesse desta dissertação gostaria de focar-me principalmente na questão da perceção de cena e objetos presentes em qualquer tipo de ambiente. Para tal, serão desenvolvidos processos e ferramentas que permitam modelar um espaço 3D em tempo real e permitir estudar-lo, tudo com o intuito de extrair o maior volume de características possível de forma a verificar a existência, detetar a posição, identificar e categorizar os vários objetos que se encontram na cena.

Outras questões pertinentes poderão surgir do estudo do contexto da cena, nomeadamente, perceber qual a sua importância na tarefa do reconhecimento de objetos. No fim desta investigação pretende-se deixar em aberto o estudo de outro tipo de questões relacionadas com a lógica introduzida pelo comportamento e estado dos elementos que se encontram presentes: Qual a relação entre os objetos? Quais as ações que se encontram a decorrer numa dada cena?

Uma das questões mais pertinentes e que deve ser considerada durante esta investigação será qual a estrutura e representação a adotar para uma dada cena, objeto, ou ainda, pequenas partes dele. De todas as características que obtemos baseadas na posição, cor, tamanho, forma, etc., poderão existir outras que apresentem interesse e foco de estudo com o intuito de resolver os problemas inerentes do processo de reconhecimento de objetos em ambientes 3D devido à introdução de múltiplas perspetivas, oclusões, diferentes condições de iluminação, etc.

---

[14] https://www.umbel.com/blog/retail/13-retail-companies-already-using-data-revolutionize-shopping-experiences/
[15] http://www.q-better.com/

Muitos dos avanços realizados nas tecnologias e técnicas que permitem o reconhecimento de imagem 3D surgem da generalização dos problemas e conjunto de soluções apresentadas para ambientes em 2D. É necessário que mais técnicas e paradigmas sejam modelados para resolver de forma mais eficiente esta tipologia de problemas. Colocam-se assim as seguintes questões: quais os melhores descritores a utilizar em ambientes 3D? Como detetamos os casos de paralelismo? Como verificamos os casos de simetria? Como resolvemos problemas associados com objetos incompletos? Como tratamos a variação da perspetiva?

Atualmente, um dos maiores avanços sentidos na área da visão por computador passa pelo desenvolvimento e exploração de novas características numa imagem. O que realmente faz de uma característica uma boa característica? Se pegarmos em duas fotografias retiradas sobre diferentes condições de luz e perspetiva, por exemplo, a característica ideal apresentaria a mesma descrição numérica para ambos os casos.

"*Um artista consegue indicar o tronco de um elefante com apenas alguns traços de tinta, mas nenhuma característica matemática consegue chegar a esse nível tão simples de eficiência*"

É com esta citação, que surge de uma conversa entre William T. Freeman [10], Pietro Perona [42] e David G. Lowe [43], que pretendo fechar esta introdução ao mundo da visão por computador e processamento de imagem, deixando em aberto as seguintes questões:

- *"Quais as características mais úteis a extrair e memorizar de uma imagem?";*
- *"Que estrutura de armazenamento utilizar?";*
- *"Qual a melhor metodologia a seguir para realizar um reconhecimento eficiente?";*
- *"Qual a importância do contexto da cena no âmbito do reconhecimento de objetos?";*

Estas são algumas das perguntas que irei tentar explorar e responder ao longo do período de realização desta dissertação. Com todas as conclusões que forem possíveis retirar pretende-se encontrar validação e suporte para as decisões que forem tomadas durante a modelação e concetualização da arquitetura por detrás do protótipo que irá ser desenvolvido e utilizado como prova científica dos conteúdos que são retratados nesta dissertação.

2.2 Contínuo da Realidade e Virtualidade

Em 1994, Paul Milgram, definiu uma taxonomia para ambientes virtuais e ambientes reais onde propôs o que ele designou de "*Virtuality Continuum*", ou "Contínuo de Virtualidade" [2].

Este projeto surgiu exatamente pelo facto de existir uma necessidade de definir, de uma forma mais concisa, o conceito de Realidade Aumentada (RA) devido à inconsistência entre as várias definições do termo que, nesse mesmo período, contemplou-se com maior frequência no meio científico. A definição que ele propôs baseia-se naquela apresentada em "*Telemanipulator and Telepresence Technologies*" [44]:

"*Aumentar o feedback natural para o utilizador através de sinais simulados ...*".

Mas, segundo o autor, nas suas longas pesquisas que recaem sobre os trabalhos publicados neste jornal, é possível encontrar uma outra definição que deve ser igualmente considerada para o contexto em que nos encontramos:

"*É uma forma de Realidade Virtual onde o Head-Mounted Display (HMD) do utilizador é transparente, permitindo uma visão mais clara do mundo real ...*"

Face a estas duas definições, que em nada são semelhantes, o autor decidiu debater –se sobre duas questões que, na sua visão, mereciam ser consideradas para a busca de uma melhor designação para o tema em questão:

- *"Qual é a relação entre a Realidade Aumentada e a Realidade Virtual?";*
- *"O termo de Realidade Aumentada deve apenas estar limitado a HMD's?";*

Um ambiente de Realidade Virtual (RV) é aquele onde o utilizador observador se encontra totalmente imersivo num mundo sintético, não real. Este mundo poderá, ou não, imitar as propriedades do mundo real, sendo elas reais ou ficcionais, e poderá quebrar os limites da realidade como a conhecemos, ao criar um lugar onde as leis da física, que governam a gravidade, o tempo e a matéria,

não têm efeito. Em contraste, no mundo real, todos os ambientes encontram-se restritos pelas leis da física.

Paul Milgram, ao invés de separar estes dois conceitos e apresentar o Ambiente Real e Ambiente Virtual como objetos totalmente opostos um do outro, preferiu visualizar-los como os extremos de um único contínuo.

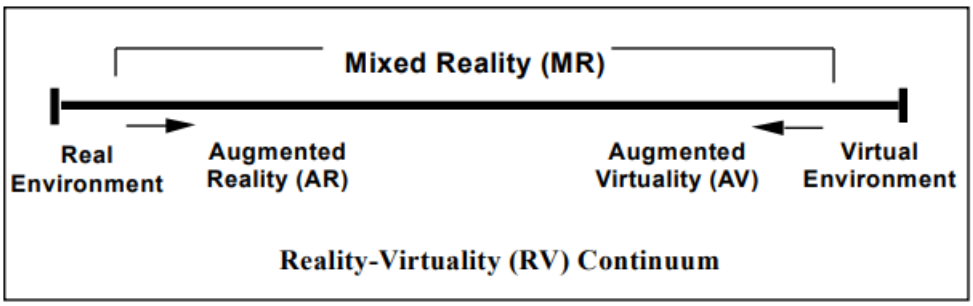

Figura 24 - Representação do Contínuo da Realidade e Virtualidade.

Desta forma, segundo o esquema apresentado, qualquer Ambiente Real consiste exclusivamente em objetos reais e todas as cenas que sejam possíveis observar ao vivo, seja esta ação realizada em pessoa, ou através de um qualquer tipo de "janela" através de um ecrã (e.g Imagem, Vídeo, etc.).

No outro oposto extremo do contínuo encontra-se o Ambiente Virtual que consiste exclusivamente em objetos virtuais, os quais poderão, por exemplo, incluir simulações gráficas obtidas por processos computacionais, para possível visualização num monitor ou ambiente mais imersivo através de outros tipos de telas que permitam a sua visualização (e.g HMD).

Neste processo, que tinha como objetivo a estruturação e correta designação do conceito de Realidade Aumentada surge, de forma adicional e praticamente grátis, a definição de um outro conceito que veio ganhar mais força apenas nos dias presentes, denominado de Realidade Mista (RM). A RM surge em qualquer ponto entre os extremos do contínuo da realidade e virtualidade e é visto como o ambiente que funde os objetos reais e virtuais no mesmo local, sendo possível observar-los através do mesmo visor.

Neste contexto, a Realidade Aumentada (RA) caracteriza-se pelo predomínio do real sobre o virtual, ou seja, o utilizador visualiza o mundo real onde são inseridos objetos virtuais. Por observação do esquema, é possível visualizar que este tipo de RM se encontra mais próximo do extremo referente ao mundo real.

Em contraste, a Virtualidade Aumentada (VA) define-se exatamente como o oposto, sendo que o utilizador, imerso num ambiente virtual, poderá visualizar componentes do mundo real. Este tipo de RM encontra-se assim mais próximo do extremo referente ao mundo virtual.

2.2.1 Categorias de Visualização

Considerando o contexto apresentado na Figura 24 - Representação do Contínuo da Realidade e Virtualidade., a primeira definição apresentada nesta secção - "*Aumentar o feedback natural para o utilizador através de sinais simulados ...*" –, é agora mais clara.

Após introdução a conceitos mais elementares, que surgem no âmbito da temática do contínuo da realidade e virtualidade, é de interesse destacar e aprofundar os diferentes tipos e ambientes de RA, sejam eles baseados em "*see-through*" HMDs ou em monitores, mas igualmente orientados e definidos por aquilo que é apresentado no diagrama que se encontra representado na Figura 24 - Representação do Contínuo da Realidade e Virtualidade.

1. "*See-through AR displays*"

Este grupo de "*displays*" é caracterizado pela possibilidade que o observador tem em visualizar diretamente elementos do mundo virtual e complementar aquilo que o rodeia no mundo real, atingindo um grau máximo na sensação de presença e imersão nestas novas realidades.

O método mais comum para a construção deste tipo de tecnologia, surge na utilização de monitores especiais ou lentes de vidro que permitem sobrepor e manipular imagens, geradas por sistemas computacionais que são alimentados por motores gráficos próprios, em cenários do mundo real.

Este tipo de técnicas e *hardware* composto são já tecnologias que surgem com algum nível de maturidade no mercado desde os anos 90, especialmente em sistemas de aviação [45] (Em maioria militar) e outros painéis semelhantes de instrumentos, mas que encontram-se neste preciso momento a ser aplicados noutras áreas que procuram também utilizar tecnologias relacionadas com a realidade virtual e/ou aumentada na sua atividade, especialmente em ambientes de produção e manutenção do produto [46] [47] [48].

Uma das áreas de maior interesse e destaque, que neste momento demonstra um especial esforço em aplicar estas técnicas de RA na sua prática, é a área da imagem médica que pretende amplificar a visualização de dados médicos, úteis para os técnicos de saúde, diretamente sobre o próprio paciente durante os processos de teste e diagnóstico ou ainda na realização de outras operações um pouco mais delicadas [49].

Após longos períodos de investigação e desenvolvimento realizado sobre os "*displays*" "*see-through*" (ST), vários problemas foram detetados no que toca ao funcionamento deste tipo de sistemas e é possível incluir: falta de precisão e baixa latência na obtenção de dados dos sensores relativos à posição e direção da cabeça e corpo do observador, existência de vários passos de calibração normalmente demasiado complexos ou demorosos, desconforto na utilização de HMD's devido ao peso que surge da complexidade e numerosa quantidade de componentes de hardware e outros periféricos que complementam o sistema móvel, em particular, a bateria utilizada para alimentar esta categoria e classe de sistemas computacionais.

Muitos dos problemas apresentados anteriormente afetam, na sua maioria, a experiência de interação e usabilidade do observador utilizador. Outro grupo de possíveis obstáculos que podemos encontrar durante a definição e utilização deste tipo de tecnologias e sistemas informáticos integrados passa pela existência de problemas de carácter mais natural.

Um exemplo disso será a variação da perspetiva na visão do utilizador sobre o conteúdo virtual gerado no ambiente real que pode, em vários casos, levar à criação de ambiguidades e outros artefactos pela existência de oclusões entre estes mesmos elementos e os objetos do mundo real, que se podem sobrepor no decorrer da cena que rodeia o observador.

Nos dias de hoje é possível encontrar muitas soluções que ajudam a aliviar, em grande maioria, os problemas que foram aqui apresentados. É de certa forma interessante apontar que já em 1994, Paul Milgram, previa a solução mais utilizada atualmente na indústria e outros meios científicos, através da introdução que fez ao conceito de sistemas de vídeo ST que passa por um processo de integração de câmaras de vários formatos de imagem (e.g RGB, Profundidade, etc.) nestes sistemas compostos, tudo com a finalidade de tentar reproduzir e sincronizar o próprio sistema de visão da máquina com aquilo que o observador consegue ver com os próprios olhos no preciso momento da sua utilização, de forma quase instantânea e em tempo real.

2. "*Monitor based AR displays*"

Quando inserimos o termo RA em conjunto com "*monitor-based*" ou "*window-on-the-world*" (WoW) na mesma frase, estamos a referir-nos a sistemas de visualização não imersivos onde imagens, geradas por sistemas computacionais, são introduzidas analogicamente ou digitalmente sobre outras imagens ou vídeos alojados num qualquer dispositivo, ou mesmo, noutro tipo de multimédia que se encontre a ser reproduzido ou transmitido em tempo real. Embora este tipo de tecnologia já seja conhecido há muito tempo, através de técnicas baseadas geralmente em processos de "*chroma-keying*"-técnica que consiste em colocar uma imagem sobre outra através do anulamento de uma cor padrão, como por exemplo o verde ou o azul (e.g Utilização de *green-screen*)-é possível encontrar uma maior aplicabilidade deste tipo de tecnologias e técnicas quando introduzidas em sistemas estereoscópicos.

Um exemplo dado pelo próprio autor seria o projeto "*ARGOS - Augmented Reality through Graphic Overlays on Stereovided*", desenvolvido nos laboratórios "*ATR Communication Systems Research Laboratories*", que permitiu, na década dos anos 90, despoletar um novo interesse por parte da indústria no desenvolvimento deste tipo de ambientes de RA. Vários estudos foram realizados para perceber qual a sua aplicabilidade prática, entre outros processos e tecnologias que surgiram, como por exemplo: "*virtual landmarks*", "*virtual tethers*" ou até mesmo "*virtual control*".

Já naquele período era possível encontrar este tipo de tecnologias aplicadas em sistemas altamente sensíveis e exatos, utilizados na sua maioria para a visualização de imagens médicas, com o objetivo de prestar apoio aos assistentes e técnicos de saúde para a anotação direta e medição de

distâncias entre pontos em modelos 3D virtuais, através de simples cliques no ecrã. Estes elementos resultavam de testes médicos que eram realizados pelos departamentos de imagiologia (e.g Exame raio-x, TAC, etc.).

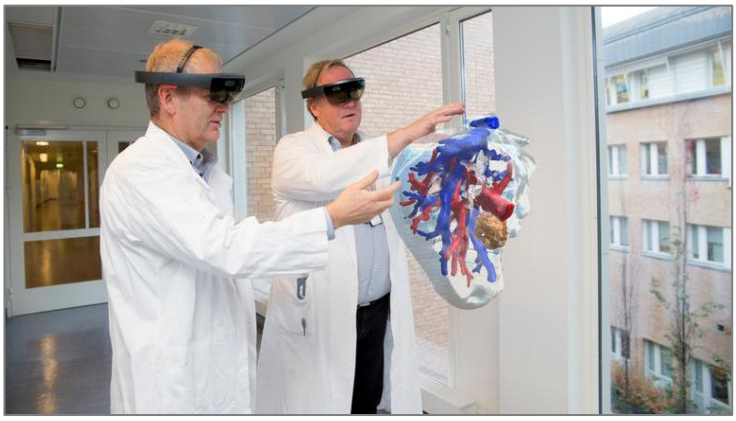

Figura 25 - Cirurgiões utilizam o Hololens da Microsoft para visualizar e explorar o modelo 3D virtual de um órgão antes e durante a operação a realizar[16].

Outro tipo de aplicação mais atual e igualmente interessante incide sobre a utilização de ecrãs gigantes onde, pela introdução de tecnologias que permitem realizar uma análise inteligente do contexto que os rodeiam, é possível criar uma interação mais realista e satisfatória entre o utilizador e quaisquer elementos e objetos virtuais que habitam nessa tela[17].

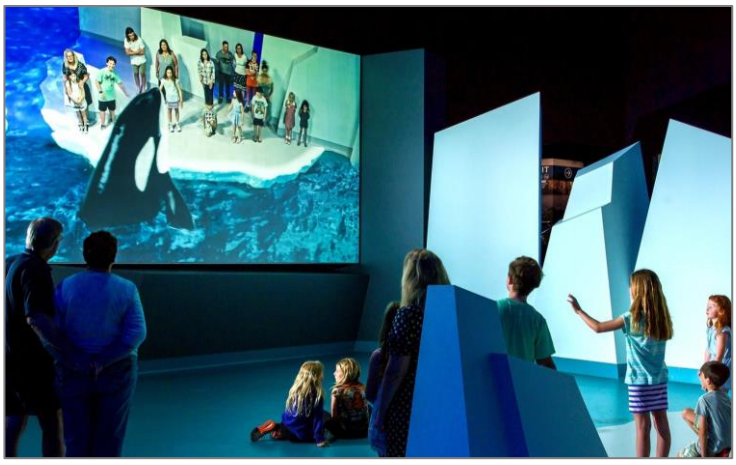

Figura 26 - Experiência de RA cinemática em qualquer ecrã com a tecnologia *Broadcast AR*.

[16] https://www.zdnet.com/article/microsofts-hololens-these-surgeons-are-using-ar-to-explore-organsmicrosofts-hololens-how-these/

[17] http://www.indestry.com/broadcast-ar/

2.2.2 Ambientes de Realidade Mista

Até ao momento surgiu a oportunidade de definir o conceito de RA dentro do contexto da temática do contínuo da realidade e virtualidade e ainda definir e ilustrar duas subclasses muito particulares de "*displays*" de RA. Nesta subsecção pretende-se explorar a possibilidade da RA poder relacionar-se com outras classes de "*displays*" de RM.

Tendo em conta a breve discussão realizada na subsecção anterior, deve ser agora evidente que os fatores principais que distinguem os sistemas "*see-through*" (ST) e "*monitor based*" vão para além do simples facto do próprio "*display*" ser "*head mounted*" ou "*monitor based*", respetivamente. Esta será mesmo a razão principal que governa a grande metáfora por detrás do comportamento apresentado por um utilizador que espera sentir-se completamente imerso num mundo plenamente egocêntrico ou, por outro lado, ter a possibilidade de visualizar esse mesmo mundo, de uma forma mais excêntrica e indireta.

Por outro lado, surgem ainda várias questões por detrás do conhecimento que o próprio utilizador, personagem que experiencia todos estes mundos, apresenta previamente, sobre o ambiente que o rodeia. Esta informação é essencial para uma melhor preparação e adaptação do ambiente virtual por processos de mapeamento obrigatórios e necessários pelos paradigmas utilizados para a visualização de conteúdos através dos sistemas ST, sendo este mesmo ponto muito menos crítico quando apresentado em "*displays*" que seguem a tipologia WoW.

Adicionalmente, deverão ainda ser levadas em conta outras questões, na sua maioria de caráter mais percetual, que surgem em função do grau de fidelidade que deve ser mantido neste novo mundo que tem como objetivo sobrepor-se sobre a realidade, tal como a conhecemos.

No mesmo período em que Paul Milgram escrevia o seu artigo, os sistemas óticos do tipo ST, constituídos por lentes transparentes, encontravam-se ainda pouco desenvolvidos, mostrando assim pouco potencial no papel e poder que tinham em alterar a realidade daquilo que um qualquer utilizador podia observar em seu redor. Embora não existisse ainda a tecnologia para realizar tal feito, manifestavam-se já várias discussões em âmbito científico e a introdução de modelos matemáticos e computacionais por parte dos investigadores. Através da sua aplicação em sistemas mais simplificados

de vídeo, era já possível, naquela década, saborear um pouco daquilo que seria possível fazer e tirar proveito deste tipo de tecnologias que permitem interligar o mundo virtual e real e manipular estes mesmos elementos como um só.

Embora naquele período esta tarefa fosse ainda realizada de uma forma muito crua e o conteúdo virtual gráfico, gerado computacionalmente, apresentasse ainda pouca qualidade e fidelidade quando inserido sobre o mundo real, os alicerces teóricos e metodológicos ficaram ali bem definidos, faltando apenas a evolução tecnológica ao nível do hardware e de outros componentes de sistemas inteligentes e distribuídos, para obter um resultado com um maior nível de maturidade.

Tudo isto para referir que, no momento em que este documento foi regido, tal feito já foi atingido graças ao avanço tecnológico que veio-se a sentir nestes últimos anos nas áreas da visão e sistemas de computação em geral. Este pode apenas ser o início, mas o surgimento de vários dispositivos que permitem realizar este tipo de tarefas (e.g Hololens, Magic Leap One, etc.) deve-se não só a estes avanços mais técnicos, mas também ao facto da indústria e o mercado empresarial e consumista em geral, cada vez demonstrar maior interesse e investimento neste tipo de tecnologias.

Retomando assim a discussão à volta do conceito do contínuo da realidade e virtualidade e a todo um grupo de problemas que surgem da própria definição de todo o "*substratum*" que o complementa:

- *"O ambiente que está a ser observado é maioritariamente real, ao qual foram adicionados elementos gerados computacionalmente?";*
- *"O ambiente que rodeia o utilizador é principalmente virtual, mas "aumentado" através da utilização de informação obtida pelo mapeamento daquilo que conhecemos ser real?";*

Gostaria de referir que o autor, durante as suas investigações, deixou pendente uma terceira questão que não se encontra aqui listada, mas seria aquela referente ao ponto de equilíbrio e central do contínuo apresentado no diagrama da Figura 24. Já no período dos anos 90 se previa que tal feito apenas seria possível com os avanços da tecnologia, e acreditava-se que o dia iria chegar em que o próprio

utilizador, imerso entre os dois mundos, virtual e real, teria mesmo dificuldade em compreender se o cenário principal que ali se desenrolasse perante os seus olhos seria o real ou o virtual.

Todas estas questões servem o único propósito de melhor introduzir e ao mesmo tempo definir toda uma metodologia integral daquelas que seriam as melhores regras e conceitos a seguir para classificar e compreender os diferentes tipos de "*displays*" de RM existentes.

O caso apresentado pela segunda questão retrata exatamente aquilo que já foi definido anteriormente neste documento como sendo a Virtualidade Aumentada (VA), como referência a tudo o que nos surge como sendo qualquer ambiente gráfico virtual, seja ele, completamente imersivo, parcialmente imersivo, ou ainda, ao qual tenha sido adicionado alguma parte da realidade (e.g Objeto, textura, etc.). Quando esta classe de "*displays*" é alterada e especialmente estendida para incluir situações em que objetos e outros elementos reais, tais como a mão dos utilizadores, podem ser introduzidos neste mundo virtual, que é maioritariamente representado por aquilo que consideramos ser o real ou não, com o objetivo de realizar algum tipo de ações como apontar, tocar, ou manipular qualquer elemento desta cena virtual, os problemas percetuais que surgem, especialmente para os "*displays*" estereoscópicos, tornam-se realmente desafiantes.

Com o objetivo de melhor distinguir as diferenças e semelhanças entre os vários conceitos de "*displays*" existentes e classificar-los dentro do ambíguo espetro da Realidade Mista (RM), é útil listar-los desta forma:

1. Sistemas de RA baseados em "*Monitor-based displays*" (MBDs), não imersivos, aos quais, imagens geradas computacionalmente, são sobrepostas;
2. Seguindo o mesmo que estipulado no ponto 1, mas utilizando "*displays*" baseados em HMDs e imersivos, em vez de utilizar ecrãs que seguem a tipologia WoW;
3. Sistemas de RA baseados em HMDs, incorporando sistemas óticos ST pela utilização de lentes transparentes;
4. Sistemas de RA baseados em HMDs, incorporando sistemas óticos ST pela utilização de vídeo;

5. Sistemas de VA baseados em MBDs, incorporando um mundo gráfico gerado computacionalmente, através da sobreposição da realidade em vídeo;
6. Sistemas de VA imersivos ou parcialmente imersivos, incorporando um mundo gráfico gerado computacionalmente, através da sobreposição da realidade em vídeo ou mapeamento direto de certas formas e texturas;
7. Sistemas de VA parcialmente imersivos, que permitem alguma interação entre o utilizador e objetos reais através da manipulação, utilizando as suas próprias mãos (e.g. Tocar, agarrar, mover, etc.);

É de notar que muitas mais classes poderiam ser aqui delineadas, mas a ideia apresentada por Paul Milgram foi exatamente a de limitar esta população aos principais fatores e características que predominam entre as várias classes de "*displays*" existentes para RM.

Classes do Sistema de RM	Real (R) ou CGI?	Visualização Direta (DT) ou Digitalizada (DG) do substrato?	Referência Excêntrica (EX) ou Egocêntrica (EG)?	Mapeamento (1:1) ou (1:k) ?
1	R	DG	EX	1:k
2	R	DG	EG	1:k
3	R	DT	EG	1:1
4	R	DG	EG	1:1
5	CGI	DG	EX	1:k
6	CGI	DG	EG	1:k
7	CGI	DT, DG	EG	1:1

Tabela 1 - Diferenciação entre as várias classes de "*displays*" de RM.

Um pequeno resumo de como alguns dos fatores discutidos até ao momento influenciam as múltiplas classes selecionadas cuidadosamente e que se encontram listadas acima, é apresentado na Tabela 1.

A primeira coluna refere-se àquela que será a propriedade com maior influência e peso no que toca ao processo de distinção entre as classes apresentadas, separando e distinguindo os lados opostos (Esquerdo e direito) da Figura 24 - Representação do Contínuo da Realidade e Virtualidade., procurando definir, respetivamente, se a camada que define a cena principal apresentada ao utilizador deriva da realidade (R) ou de um mundo gráfico gerado computacionalmente (CGI). Esta informação não é ainda suficiente para conseguirmos compreender qual o *hardware* utilizado pelo observador para visualizar uma dada cena. Essa distinção é reforçada pela segunda coluna, onde notamos imediatamente que não existe uma correspondência estrita com a primeira coluna.

Dentro deste contexto, uma "visualização direta" refere-se ao caso em que o mundo principal e maioritário de uma cena virtual é visualizado diretamente através de ar ou de vidro (também conhecida como realidade não mediada), enquanto que no caso oposto em que a visualização é realizada indiretamente, o mundo real deverá ser digitalizado por recurso a qualquer meio existente, seja por utilização de câmaras de vídeo, laser ou sistemas de ultrassom, etc. e depois processado e ressintetizado ou reconstruído por meio da utilização de algoritmos e outras técnicas de visão de forma a expor o resultado em formato de vídeo ou transmitindo diretamente em "*displays*" e arquiteturas prontas para tal.

A questão apresentada na quarta e última coluna, ou seja, procurar perceber se estamos perante uma situação em que é necessário intervir com um processo de mapeamento entre o mundo real e virtual, relaciona-se de uma forma muito próxima com a distinção excêntrica/egocêntrica realizada na terceira coluna. Quando estamos perante certos contextos mais específicos e situações onde é necessário realizar algum tipo de mapeamento e perceção espacial entre os dois mundos, estes casos devem ser necessariamente definidos como sistemas de imersão egocêntricos, enquanto que, num raciocínio inverso, esta conclusão já não será logicamente válida.

Provavelmente, a conclusão mais importante que devemos retirar por análise da Tabela 1 é que nenhuma linha e classe apresentada é idêntica. Embora a comparação destas características seja um pouco limitada e possibilite apenas uma simples distinção entre as classes aqui nomeadas, uma estrutura global para categorizar todas as possíveis entradas de "*displays*" de RM seria muito mais complexa. Esta pequena observação realça a importância e necessidade de definir uma taxonomia que seja integral e eficiente na classificação dos "*displays*" de RM, tanto para a identificação dos diferentes

agrupamentos e dimensões chave utilizados para, de forma mais simplificada, distinguir todos os sistemas e conceitos candidatos e servir como estrutura para descrever os problemas mais comuns que surgem pela utilização de cada umas das tipologias apresentadas de "*displays*".

2.2.3 Casos de Aplicação

1. HoloMaps

Para começar, apresenta-se uma simples solução com o intuito de demonstrar as capacidades do processador holográfico presente no dispositivo do Hololens da Microsoft. Esta aplicação foi desenvolvida pela empresa Taqtile e surge como um dos primeiros ambientes que permitiu ao utilizador navegar um modelo holográfico da representação de um mapa tridimensional enquanto, simultaneamente, deixa o mesmo adicionar e sobrepor o mapa com quaisquer anotações e outros dados em tempo real. Mas mais do que isso, esta aplicação permite realizar a partilha, no mesmo local, daquilo que encontra-se a ser gerado graficamente entre vários possíveis utilizadores que tenham um dispositivo Hololens. Apresentado pela equipa de investigadores e desenvolvedores da Microsoft, este conceito surge batizado como "*hololens shared experience*".

Figura 27 - Demonstração do poder do dispositivo Hololens para a visualização de mapas 3D dentro do ambiente da aplicação HoloMaps, desenvolvida pela empresa Taqtile[18].

[18] https://taqtile.com/2017/07/28/holomaps-enterprise-vrfocus/

2. *SketchUp Viewer for Hololens*

Um outro exemplo que também permite aos utilizadores colaborar em ambientes de realidade mista e criar novas experiências imersivas do mesmo género, surge com o nome de *SketchUp Viewer for Hololens*. Ao acedermos a esta plataforma é possível partilhar o espaço de um projeto ou planificação que o utilizador tenha criado em 3D. Ao recorrermos a um dispositivo *Hololens* é possível navegar, criar, modificar e até mesmo interagir com os modelos delineados. Juntamente com todo o poder social e comunicativo aqui presente, não existe limites para o trabalho e toda a originalidade que daqui pode surgir.

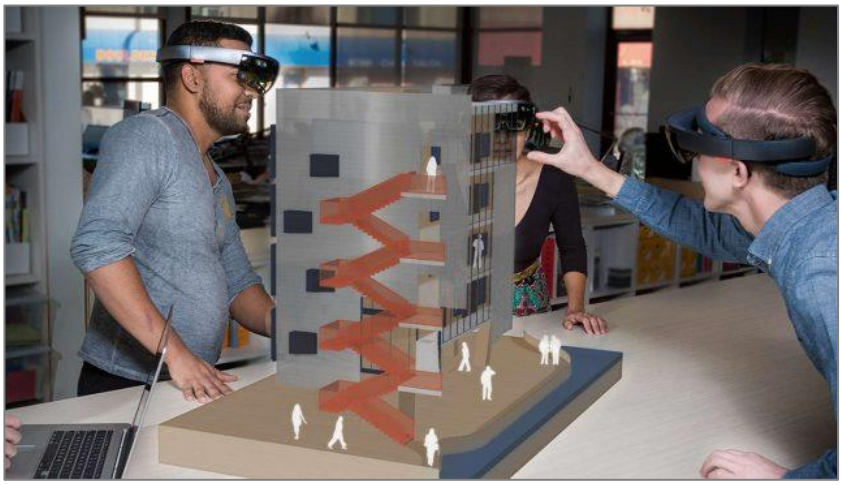

Figura 28 - Exemplo de colaboração e partilha de um espaço de trabalho pela utilização do dispositivo Hololens no ambiente *SketchUp Viewer for Hololens*, aplicação desenvolvida pela empresa Trimble[19].

3. Fragments

Este tipo de tecnologias não poderia apenas ser utilizado como ferramenta de trabalho, existe sempre lugar para o entretenimento. Surge assim um estúdio de desenvolvimento de jogos denominado AsoboStudio, muito reconhecido pelo seu trabalho realizado por detrás da criação de várias adaptações de filmes da Disney em videojogos. O utilizador tem o papel de um detetive numa cena de crime e o ambiente apresentado pelo jogo Fragments tira partido de várias técnicas de mapeamento espacial com o objetivo de colocar dinamicamente várias pistas no espaço físico onde o próprio utilizador se encontra. Adicionalmente, surgem ainda personagens virtuais que "apercebem-se" da presença do utilizador e conseguem igualmente interagir com o espaço físico que rodeia o jogador (e.g. Andar, sentar, tocar, etc.).

[19] https://www.geospatialworld.net/blogs/sketchup-viewer-for-microsoft-hololens-is-here/

Figura 29 - Experiência de jogo que é possível ter no local de crime apresentado em Fragments, título desenvolvido pelo estúdio AsoboStudio[20].

2.2.4 Sumário

Com tudo aquilo que foi possível absorver das várias pesquisas que foram realizadas até ao momento, é finalmente tempo para parar um pouco, refletir sobre a utilidade desta informação e ao mesmo tempo encaixar todas as peças para perceber de que modo será possível usufruir deste conhecimento para a construção do projeto que no fim se pretende desenvolver como demonstração e prova destes mesmos conceitos.

Todas as temáticas discutidas aqui surgem pela necessidade que existe em fundir os dois mundos da visão por computar/reconhecimento de imagem e os vários conceitos por detrás do contínuo da realidade e virtualidade.

Assim que nos introduzimos neste documento, é possível perceber de imediato que um dos objetivos para com a realização desta dissertação é exatamente a de perceber quais as possibilidades das tecnologias atuais que vão de encontro com os tópicos de visualização e imersão. Apresenta-se assim a disponibilidade de acesso a umas das tecnologias de topo e pioneiras da atualidade na área do *rendering* e visualização de hologramas em ambientes de realidade mista, o Hololens da Microsoft.

[20] http://www.asobostudio.com/games/fragments

Tendo em conta todas as classes de "*displays*" anteriormente apresentadas, facilmente coloca-se este dispositivo na categoria de RM onde surge a possibilidade de trabalhar com sistemas de RA baseados na utilização de um HMD que é constituído por um sistema ótico ST com lentes transparentes, neste caso através de um ecrã desenvolvido com tecnologia OLED.

Embora o *hardware* disponível não pareça uma solução realmente polida e final, já que, no momento da realização desta dissertação, a empresa da Microsoft apenas disponibiliza o produto através de *kits* de desenvolvimento, continua-se a estar perante um enorme e complexo pacote que junta vários módulos de *software*, sensores, tecnologias e no fim, puro conhecimento.

Com tudo dito, pretende-se explorar ao máximo o poder do dispositivo que aqui tem-se em mãos e desenvolver uma solução que tire o maior partido dos sensores presentes e, consequentemente, dos dados que são passíveis de extrair. Não podemos esquecer das possibilidades que existem ao ser possível visualizar elementos holográficos através da construção de um ambiente de realidade mista completamente imersivo que irá servir de apoio para as várias tarefas de análise de ambiente que aqui vão decorrer.

É de notar que o objetivo do projeto TEARS é retirar partido deste tipo de imersão e permitir visualizar no mundo físico que nos rodeia, o resultado de processos de análise e deteção para a classificação de múltiplas classes de objetos que rodeiam o utilizador final. O resultado esperado será a colocação de anotações virtuais e tridimensionais que vão apoiar-se nos dados das previsões obtidas para manipular a sua posição, cor, tamanho e outras propriedades da sua própria forma.

3. TEARS - TANGIBLE ENVIRONMENTS IN AUGMENTED REALITY SYSTEMS

3.1 Motivação e Objetivos

Após o processo de exploração e apreensão de uma pequena amostra daquilo que a área de visão por computador tem para oferecer, entramos agora numa fase de demonstração e aplicação de parte desse conhecimento.

Na nossa atualidade, o reconhecimento de imagem já não é uma incógnita para a indústria e investigação, existem várias soluções lá fora que nos permitem realizar tal efeito com uma enorme rapidez e eficiência, sejam elas soluções aplicadas sobre uma única imagem ou até mesmo em vídeo e em tempo real [50] [51] [52] [53].

Juntamente com esta questão, junta-se ainda a temática da RA, que surge como uma das principais tecnologias dos últimos anos. Recentemente, as empresas de topo, como por exemplo a *Microsoft*, lançaram no mercado novos óculos de RM, como o HoloLens, renovando o interesse dos vários setores para um nível sem precedentes. A utilização de sistemas de RA no apoio à realização das atividades humanas tem despertado um enorme interesse. Estes sistemas, para fornecerem o devido apoio, devem reconhecer a atividade executada pelo utilizador [54], o respetivo contexto e os objetos que o rodeiam.

A *framework* que aqui propõe-se desenvolver consiste, na sua essência, em implementar uma nova metodologia para o processo de extração e processamento dos dados que conseguimos obter do ambiente através dos sensores mais genéricos presentes nos dispositivos de RM da atualidade (e.g Câmara, Giroscópio, Sensor de Profundidade, etc.). Ou seja, procurar explorar e verificar qual a melhor aplicabilidade das características encontradas e ao mesmo tempo desvendar quais os processos de computação, modelos estruturais de organização e armazenamento de informação [55] [56] [57] [58] que permitem obter os resultados com maior qualidade no momento de identificação dos múltiplos objetos que se encontram presentes no ambiente em que se encontra o dispositivo utilizado para este mesmo efeito.

Com a realização deste projeto pretende-se obter um sistema integrado e constituído por várias camadas que, na sua essência, permitam a classificação em tempo real das múltiplas classes de objetos que rodeiam o utilizador do dispositivo. Para cumprir com estes objetivos, é necessário desenvolver três grandes módulos que, embora apresentem tarefas totalmente independentes, encontram-se em harmonia na hora de realizar o reconhecimento de objetos num ambiente de realidade mista:

- WebTEARS - Servidor *All-In-One* (AIO) para comunicações, alojamento de dados e que permite realizar o processo de deteção, localização e classificação de objetos como um serviço;
- DeskTEARS - Plataforma para gestão dos dados necessários para o funcionamento do sistema;
- UnityTEARS - Plataforma imersiva em realidade mista para análise em tempo real do ambiente;

De notar que um dos objetivos com a realização deste trabalho é a de executar todo o processo de reconhecimento num servidor remoto e que este deve ser disponibilizado como um serviço numa máquina que se encontre melhor preparada e que seja mais apropriada para este tipo de tarefas. Esta mesma decisão deve-se pelo simples facto do dispositivo utilizado para esta demonstração, o *Hololens* da Microsoft, não apresentar as condições necessárias a nível de *hardware* para permitir a criação de um ambiente local que permita explorar os recursos de aprendizagem máquina na tarefa de detetar e classificar múltiplos objetos de forma simultânea e em tempo real. Embora, fica a informação de que uma segunda geração deste dispositivo encontra-se prestes a ser lançada em 2019 e foi prometido por parte da Microsoft um novo processador holográfico (HPU) que apresenta integrado um novo componente acelerador preparado especialmente para módulos e sistemas de aprendizagem em *deep learning*.

Sendo assim, para este mesmo efeito, surgem quatro fases de desenvolvimento associadas a este projeto que serão discutidas ao longo desta secção. Expresse-se o facto de que as seguintes etapas deste trabalho encontram-se ordenadas de forma cronológica e pela necessidade com que foram surgindo ao longo deste projeto. Era possível listar de forma muito direta os requisitos e as funcionalidades escolhidas especialmente para o desenvolvimento das várias plataformas que vão ser apresentadas de seguida,

mas desta forma será mais simples e intuitivo perceber toda a lógica e as decisões que foram realizadas ao longo da construção desta arquitetura e também, de certa forma, conseguir que o leitor acompanhe melhor todo o raciocínio e que nele desperte maior interesse em tudo aquilo que aqui será apresentado.

3.2 Serviços e Comunicações

3.2.1 Disponibilização de uma Base de Dados Relacional

Era fácil de prever que, neste contexto, a necessidade de recorrer a uma estrutura organizada para o armazenamento dos dados fosse mesmo inevitável. Embora as entidades intervenientes nesta prova e demonstração sejam poucas em número, torna-se muito mais fácil manipular e correlacionar os dados entre sí quando integra-se uma base de dados no modelo de arquitetura do sistema que se pretende aqui construir.

Para cumprir com os objetivos enunciados para este projeto, é necessário começar por criar uma estrutura relacional onde seja possível guardar os dados relativos a cada utilizador que irá utilizar o sistema, a lista de objetos/modelos que cada um pretende reconhecer de forma única e independente, as imagens que vão ser carregadas e associadas a cada um dos objetos e as anotações manuais que vão ser feitas para cada uma delas.

Ilustra-se de seguida as diferentes tabelas de dados que existem no sistema:

UTILIZADOR			
CHAVE	ATRIBUTO	TIPO	DESCRIÇÃO
PRIMARY	Email	STRING	Email para autenticação
-	Password	STRING	Password para autenticação
-	CreatedAt	TIMESTAMP	Momento de criação da entidade
-	UpdatedAt	TIMESTAMP	Momento da última atualização da entidade

Tabela 2 - Esquema da tabela de Utilizadores na Base de Dados.

MODELO			
CHAVE	ATRIBUTO	TIPO	DESCRIÇÃO
PRIMARY	Id	UUID	Valor para identificação única do objeto
-	Label	STRING	Classificação definida pelo utilizador proprietário
-	Description	STRING	Descrição do objeto
FOREIGN	UserEmail	STRING	Chave primária do utilizador proprietário
-	CreatedAt	TIMESTAMP	Momento de criação da entidade
-	UpdatedAt	TIMESTAMP	Momento da última atualização da entidade

Tabela 3 - Esquema da tabela de Modelos na Base de Dados.

IMAGEM			
CHAVE	ATRIBUTO	TIPO	DESCRIÇÃO
PRIMARY	Id	UUID	Valor para identificação única do objeto
-	OriginalPath	STRING	Caminho do ficheiro da imagem original
-	Original	STRING	Conteúdo da imagem em formato *Base64*
-	ThumbnailPath	STRING	Caminho do ficheiro da imagem miniatura
-	Thumbnail	STRING	Conteúdo da imagem miniatura em formato *Base64*
FOREIGN	ModelId	UUID	Chave primária do modelo proprietário
-	CreatedAt	TIMESTAMP	Momento de criação da entidade
-	UpdatedAt	TIMESTAMP	Momento da última atualização da entidade

Tabela 4 - Esquema da tabela de Imagens na Base de Dados.

ANOTAÇÃO			
CHAVE	ATRIBUTO	TIPO	DESCRIÇÃO
PRIMARY	Id	UUID	Valor para identificação única do objeto
-	CroppedPath	STRING	Caminho do ficheiro da imagem recortada
-	Cropped	STRING	Conteúdo da imagem recortada em formato *Base64*
-	TopLeftX	FLOAT	Posição X do *pixel* superior esquerdo do limite
-	TopLeftY	FLOAT	Posição Y do *pixel* superior esquerdo do limite
-	BottomRightX	FLOAT	Posição X do *pixel* inferior direito do limite
-	BottomRightY	FLOAT	Posição Y do *pixel* inferior direito do limite
FOREIGN	PhotoId	UUID	Chave primária da imagem proprietária
-	CreatedAt	TIMESTAMP	Momento de criação da entidade
-	UpdatedAt	TIMESTAMP	Momento da última atualização da entidade

Tabela 5 - Esquema da tabela de Anotações na Base de Dados.

Através da simples visualização e leitura das tabelas apresentadas anteriormente, facilmente apercebemos-nos quais são as relações existentes entre as diferentes entidades apresentadas. Mas, de qualquer forma, exemplifica-se de seguida o esquema resumido dessas mesmas associações:

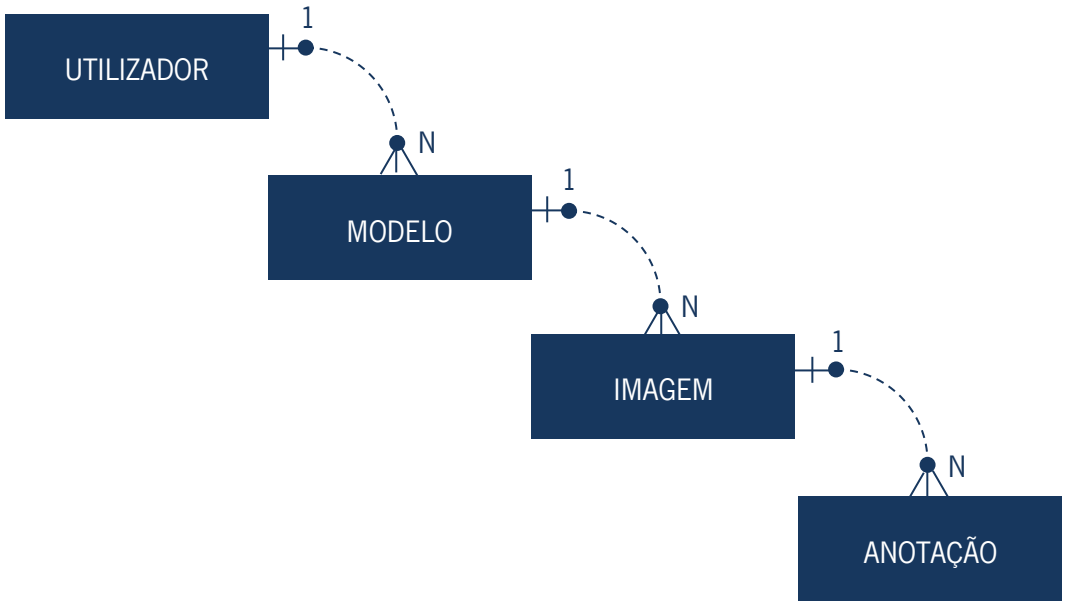

Figura 30 - Esquema de relações entre as diferentes tabelas da Base de Dados.

3.2.2 Manipulação de Dados através de uma *REST API*

Tendo adicionado a possibilidade de o sistema ter acesso a uma fonte com todos os dados primários que irão fluir neste sistema integrado, surge agora a necessidade de aceder diretamente a eles e de certa forma conseguir gerir-los remotamente. Com o intuito de realizar este mesmo propósito, tomou-se a decisão de criar uma interface de comunicações *web* orientada pelas várias propriedades e regras estabelecidas pela arquitetura *Representational State Transfer* (REST), baseadas no protocolo de comunicações *Hypertext Transfer Protocol* (HTTP).

Com a adição deste módulo aos serviços que aqui se pretendem estabelecer, ganhamos acesso a funcionalidades que nos permitem requisitar, organizar e manipular os dados apartir de qualquer lugar através da utilização de pequenas e simples rotinas (e.g GET, POST, PUT, DELETE, etc.). A decisão por detrás da utilização deste protocolo parte do princípio que a utilização do mesmo tornou-se um padrão na atualidade e encontra-se na base da construção de vários sistemas de comunicação que seguem esta mesma tipologia de forma a permitir o acesso aos dados de forma rápida, segura e totalmente independente e abstrata da plataforma e linguagem de programação utilizada para o efeito.

Desta forma, tendo em conta as diferentes entidades existentes no sistema e diversos serviços que se pretendem aqui disponibilizar, foram criados para este servidor HTTP diferentes *routers* e pontos de acesso com a simples finalidade de melhor estruturar e organizar o acesso aos recursos e às diferentes funcionalidades que são facultadas pelo mesmo:

UTILIZADOR		
PROTEGIDO	PONTO DE ACESSO	DESCRIÇÃO
NÃO	GET /api/users/auth	Autenticar ou adicionar utilizador
NÃO	POST /api/users/getUsers	Listar todos os utilizadores
SIM	GET /api/users/getUser/email	Obter informação do utilizador

Tabela 6 - Esquema da interface *RESTful* para a entidade Utilizador.

MODELO		
PROTEGIDO	PONTO DE ACESSO	DESCRIÇÃO
SIM	POST /api/models/getModels	Listar todos os modelos
SIM	GET /api/models/getModel/id	Obter informação do modelo
SIM	POST /api/models/addModel	Adicionar modelo
SIM	PUT /api/models/updateModel/id	Atualizar dados do modelo
SIM	DELETE /api/models/removeModel/id	Remover modelo

Tabela 7 - Esquema da interface *RESTful* para a entidade Modelo.

IMAGEM		
PROTEGIDO	PONTO DE ACESSO	DESCRIÇÃO
SIM	POST /api/photos/getPhotos	Listar todas as imagens
SIM	GET /api/photos/getPhoto/id	Obter informação da imagem
SIM	PUT /api/photos/updatePhoto/id	Atualizar dados da imagem
SIM	DELETE /api/photos/removePhoto/id	Remover imagem
SIM	POST /api/photos/addModelPhoto/id	Adicionar imagem ao modelo
SIM	POST /api/photos/getModelThumbnails/id	Devolver miniaturas do modelo

Tabela 8 - Esquema da interface *RESTful* para a entidade Imagem.

ANOTAÇÃO		
PROTEGIDO	PONTO DE ACESSO	DESCRIÇÃO
SIM	POST /api/annotations/getAnnotations	Listar todas as anotações
SIM	GET /api/annotations/getAnnotation/id	Obter informação da anotação
SIM	PUT /api/annotations/updateAnnotation/id	Atualizar dados da anotação
SIM	DELETE /api/annotations/removeAnnotation/id	Remover Anotação
SIM	POST /api/annotations/addPhotoAnnotation/id	Adicionar anotação à imagem
SIM	POST /api/annotations/getPhotoAnnotations/id	Devolver anotações da imagem

Tabela 9 - Esquema da interface *RESTful* para a entidade Anotação.

APRENDIZAGEM		
PROTEGIDO	PONTO DE ACESSO	DESCRIÇÃO
SIM	GET /api/learning/train	Obter estado do processo de treino
SIM	POST /api/learning/train	Desencadear processo de treino
SIM	POST /api/learning/predict	Prever classe de objetos numa imagem

Tabela 10 - Esquema da interface *RESTful* para o módulo de Aprendizagem.

3.2.3 Comunicações em Tempo Real por utilização de *WebSockets*

Nesta fase de desenvolvimento, chegamos a um ponto em que é possível realizar todo o tipo de pedidos e gerir toda a informação incidente sobre as múltiplas entidades existentes no sistema através dos serviços HTTP que podem ser prestados por um qualquer servidor remoto (WebTEARS). Mesmo assim, surge ainda a questão de qual a velocidade obtida e tempo que é necessário para realizar estas trocas de mensagens e se seria realmente possível obter respostas em tempo real para as várias petições que serão realizadas ao longo do tempo.

Apenas surge esta questão pois é de referir que um dos principais objetivos deste projeto é disponibilizar a deteção, localização e classificação de múltiplas classes de objetos como um serviço e conseguir realizar todo o processamento necessário em tempo real.

O protocolo de comunicações HTTP apresenta uma subcamada ao nível do protocolo de transporte de mensagens baseado no *Transmission Control Protocol* (TCP). Apenas este facto surge logo como um atraso no sistema que aqui pretende-se construir. Este tipo de protocolo de transporte rege-se por princípios de segurança, verificação e ordenação dos pacotes de dados o que enfatiza diretamente um aumento da latência no processo de comunicação de mensagens. Juntamente com estes factos e para piorar de forma quase exponencial a eficiência na tarefa de realizar um pedido e resposta entre dois pontos, o protocolo de comunicações HTTP obriga à transação de uma grande quantidade de metadados no cabeçalho de cada uma das mensagens expedidas. Adicionalmente, versões desatualizadas deste protocolo de comunicações adicionam ainda a necessidade de estabelecer uma nova ligação ponto a ponto de cada vez que se realiza um novo pedido, o que representa um atraso demasiado elevado para as comunicações que aqui se pretendem realizar.

Após esta pequena discussão, facilmente conseguimos concluir que a utilização do protocolo HTTP não é a resposta que procuramos, tendo em conta a necessidade que aqui surge de realizar o *streaming* de dados em tempo real entre os serviços disponibilizados pelo módulo de reconhecimento e o cliente que irá utilizar o dispositivo do Hololens (UnityTEARS). De notar que este mesmo protocolo continua a ser muito útil para toda a gestão e manipulação de dados que é necessário realizar para permitir a utilização e manutenção do sistema. É preciso ter em conta que estas tarefas são realizadas por uma outra plataforma em ambiente *desktop* (DeskTEARS).

Para cumprir com os objetivos definidos para o desenvolvimento deste sistema integrado, a solução encontrada para o problema enunciado baseia-se na utilização de um outro tipo de protocolo de transporte de mensagens denominado *WebSocket*.

O protocolo de transporte de mensagens aqui apresentado foi construído em cima do já existente TCP e utiliza um processo por HTTP de forma a realizar uma única transação inicial (*upgrade handshake*) para estabelecer a ligação. Entenda-se que, após estabelecida a ligação ponto a ponto, deixa de existir a necessidade de sobrecarregar as comunicações com transações inúteis de pacotes de controlo de lógica e semântica, nem de enviar em cada uma das mensagens longos cabeçalhos de metadados. Este tipo de comunicações permite assim uma ligação *full-duplex* e ainda bidirecional entre os intervenientes. Por fim, ao contrário do TCP, que é um protocolo de transporte baseado em *streaming* de pacotes de dados, o *WebSocket* é baseado em *streaming* de mensagens, que embora repartidas durante a transmissão, são totalmente reconstruídas antes de chegarem ao seu destino.

Desta forma, e tendo em conta as necessidades que aqui foram delineadas, foi adicionado paralelamente ao servidor HTTP já existente um módulo para realizar o tratamento deste tipo de conexões que se pretendam realizar através da utilização do protocolo de transporte *WebSocket*, tudo com o intuito de responder aos utilizadores em tempo real.

SOCKET.IO		
PROTEGIDO	EVENTO	DESCRIÇÃO
NÃO	"connection"	Conectar novo utilizador
NÃO	"disconnect"	Desconectar utilizador
NÃO	"ping"	Teste de conexão *ping-pong*
NÃO	"auth"	Autenticar utilizador e preparar novo processo para previsão
SIM	"frame"	Prever classe de objetos numa imagem

Tabela 11 - Esquema da interface Socket.Io para tratamento de mensagens através da utilização do protocolo de transporte *WebSocket*.

De forma a resumir e concluir esta subsecção relativamente à estrutura do servidor, protocolos de comunicações e serviços prestados pelo módulo desenvolvido para este mesmo efeito no desenrolar deste projeto (WebTEARS), encontra-se abaixo um simples diagrama e esquema ilustrativo dos vários componentes participantes e das interações realizadas entre os mesmos:

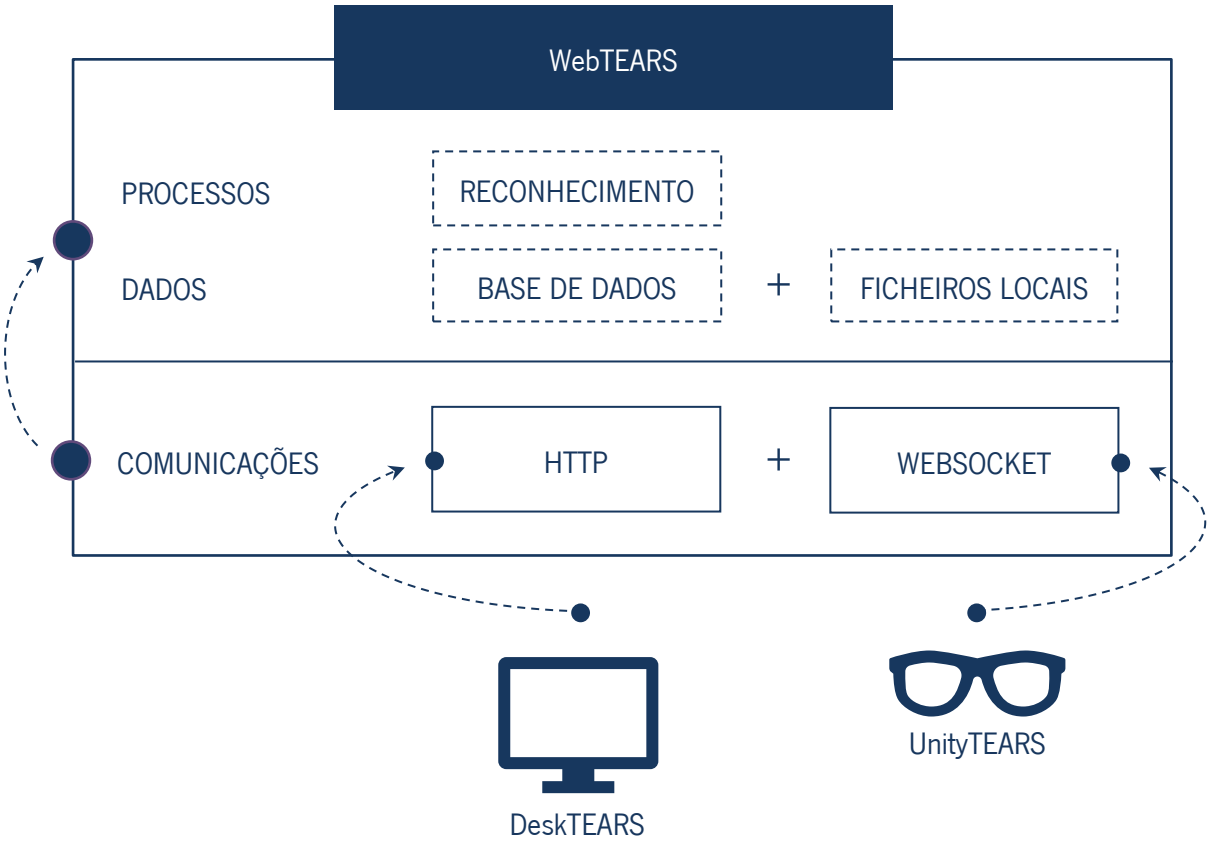

Figura 31 - Esquema dos componentes participantes e suas interações no sistema de serviços e comunicações.

3.3 Repositório de Dados

3.3.1 Armazenamento de Imagens

Embora exista uma estrutura organizada constituída por um serviço de comunicações e uma base de dados pronta para albergar e relacionar todos os dados que vão ser gerados pelos módulos integrantes deste sistema, falta ainda perceber qual a relação existente entre os utilizadores das plataformas e os mecanismos que permitem gerir os ficheiros das imagens e anotações que persistem no serviço de alojamento, tal como é possível visualizar no esquema da Figura 31.

Um dos objetivos com o desenvolvimento deste sistema é permitir distinguir os diferentes utilizadores, separando os modelos, imagens e anotações que são gerados por cada um deles. Ou seja, idealiza-se a existência de um ambiente único e totalmente independente para cada utilizador do sistema.

Para permitir tal feito, considera-se a seguinte estrutura de ficheiros para cada utilizador:

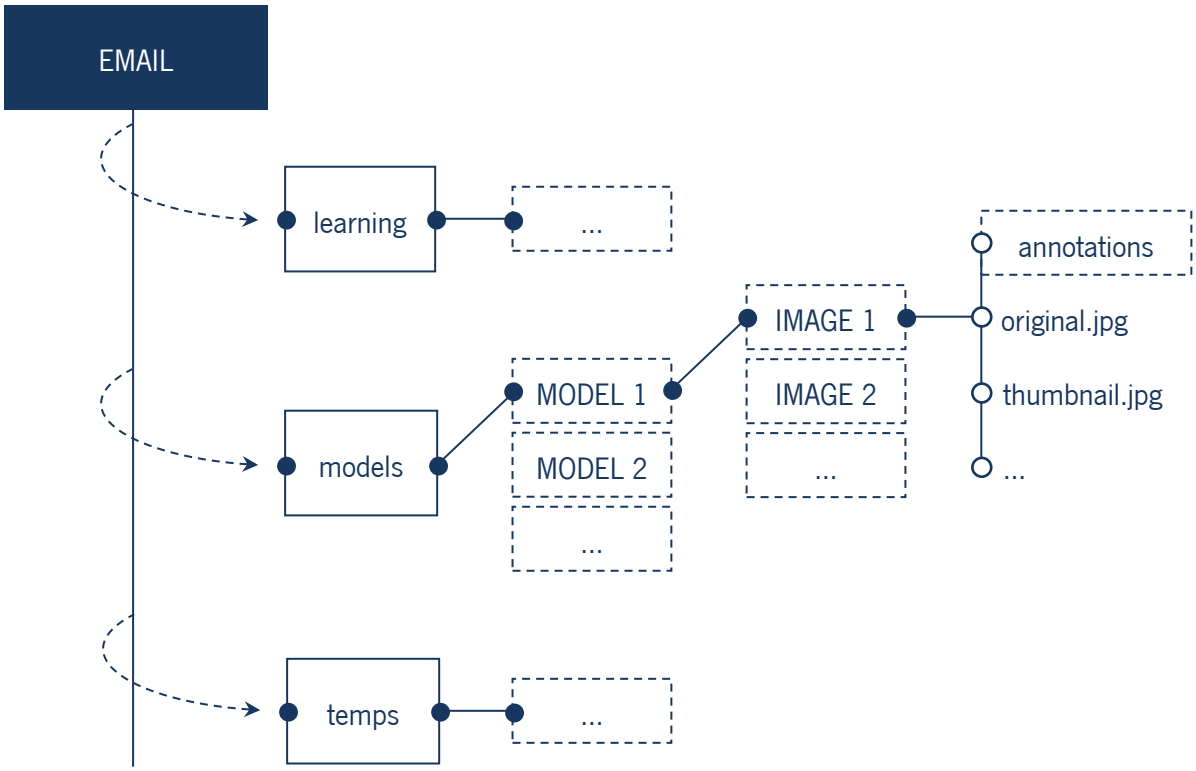

Figura 32 - Estrutura do banco de dados e ambiente de alojamento de ficheiros para cada utilizador do sistema.

Tal como é possível visualizar na Figura 32, cada utilizador apresenta uma pasta que por sí só contém mais três pastas, *"learning"*, *"models"* e *"temps"*, cada uma delas com uma finalidade diferente. Nesta subsecção apenas iremos discutir a organização e estrutura que foram definidas para as pastas *"models"* e *"temps"*, sendo que o objetivo da pasta aqui denominada por *"learning"*, e os seus conteúdos, irão ser explorados na subsecção seguinte em 3.4 - Sistema de Aprendizagem.

Uma das principais funcionalidades apresentadas ao utilizador deste sistema passa pela possibilidade de adicionar novos modelos/objetos na base de dados. Para isso, é necessário aceder aos serviços disponibilizados remotamente no servidor onde se encontra a ser executado o módulo WebTEARS e para isso basta autenticar-se na plataforma denominada DeskTEARS que foi criada exatamente com o objetivo de gerir os dados primários que fluem na arquitetura definida para este projeto, tal como já foi introduzido e discutido no inicio desta secção. Durante o processo de inserção de um novo modelo é necessário definir a *label* do próprio objeto e deixar uma qualquer descrição informativa sobre o que pareça ser pertinente para o utilizador. Assim que termina-se este processo, é necessário complementar esta tarefa partindo para o carregamento de várias imagens que, no melhor dos casos, contenham o objeto. Finalizado o processo, será depois necessário percorrer e visualizar individualmente cada uma das imagens associadas ao modelo, e que agora se encontram alojadas no servidor, tudo com a finalidade de manualmente delimitar e definir as regiões onde seja possível detetar esse mesmo objeto.

Durante o processo de carregamento de uma única imagem para o servidor, numa primeira instância, é criada uma cópia temporária do ficheiro que se pretende guardar, com o único propósito de permitir que o serviço remoto verifique se realmente se trata de uma imagem válida. Justifica-se desta forma a criação da pasta denominada *"temps"* no ambiente de cada um dos utilizadores, cuja existência responde diretamente à necessidade de gerir os ficheiros temporários. Após realizar-se esta verificação, o ficheiro temporário é eliminado e caso seja realmente válido, o sistema gera um novo identificador único que vai ser utilizado para rotular a nova imagem. Esta mesma referência é colocada na base de dados e permite nomear a pasta que é de seguida criada e colocada dentro da pasta do modelo ao qual pretendemos associar a nova imagem. É de notar que, durante esta primeira fase, são armazenados dois ficheiros por cada imagem adicionada: uma cópia original e uma versão redimensionada da mesma.

Encontram-se no Anexo II - Plataforma DeskTEARS várias imagens de capturas de ecrã que vão de encontro com todo o processo que aqui foi descrito.

3.3.2 Classificação e Anotação Manual

Partindo do princípio que o utilizador pode agora aceder à plataforma *desktop* (DeskTEARS), aproveitar-se das funcionalidades presentes para adicionar novos modelos/objetos ao sistema e carregar várias imagens para cada um deles, falta agora complementar este processo com o poder das anotações.

Provavelmente a funcionalidade mais importante da plataforma DeskTEARS, é dada a possibilidade ao utilizador de definir várias regiões de interesse (ROI) para cada uma das imagens que encontram-se alojadas no servidor. Para este mesmo propósito, basta percorrer a galeria de imagens para cada um dos modelos e utilizar a ferramenta que foi especialmente criada com o propósito de visualizar, criar e modificar as várias anotações que o utilizador estabelece para cada uma das imagens.

Neste momento, poderá levantar-se a questão sobre qual a necessidade em criar um sistema que permita tal tarefa. Durante o período de modelação e conceptualização do projeto que aqui se pretende desenvolver, um dos pontos explorados recaiu exatamente sobre perceber quais os dados que são necessários existir para permitir a deteção, localização e classificação de objetos. Como a realização desta dissertação encontrava-se ainda numa fase de pesquisa e várias soluções poderiam, entretanto, vir a surgir, uma das decisões foi exatamente a de tentar generalizar a fase de preparação de dados para qualquer método de processamento de imagem que fosse escolhido para analisar e extrair as características de uma imagem. Nas primeiras pesquisas realizadas, rapidamente concluiu-se que um dos pontos mais comuns entre os vários algoritmos de visão existentes e outros quaisquer sistemas inteligentes que são supervisionados assenta exatamente no facto de praticamente todos eles se apoiarem na análise de imagens previamente rotuladas e organizadas manualmente.

Sendo assim, tendo em conta a necessidade de generalizar os ideais e valores do sistema aqui apresentado, decidiu-se armazenar na base de dados os valores absolutos dos limites de todos os retângulos que são definidos em cada um dos ROI's (*TopLeftX*, *TopLeftY*, *BottomRightX* e *BottomRightY*).

De forma a complementar e escalar a utilidade desta ferramenta, decidiu-se também gerar e armazenar as imagens resultantes de cada região que é selecionada. Note-se que esta afirmação acaba

por justificar a necessidade de criar uma nova pasta *"annotations"* dentro da pasta pertencente a cada uma das imagens que são carregadas e associadas aos diferentes modelos/objetos existentes no banco de dados do utilizador, tal como é possível visualizar na Figura 32. Adicionalmente, fica ainda a informação de que o mesmo identificador único que é gerado pelo sistema de forma automática para referenciar a anotação na base de dados é utilizado igualmente para nomear o ficheiro da imagem recortada pelo ROI aí definido e depois armazenada nesta mesma pasta.

É importante voltar a referir que encontram-se no Anexo II - Plataforma DeskTEARS várias imagens de capturas de ecrã que demonstram as várias funcionalidades da plataforma DeskTEARS que foram discutidas até este momento.

3.4 Sistema de Aprendizagem

3.4.1 Modelo de uma Rede Neuronal Convolucional

De forma muito resumida, sistemas de deteção para utilização geral no processo de classificação aplicam um dado modelo numa imagem e selecionam várias regiões com diferentes escalas. No fim do processo, regiões que apresentam uma elevada pontuação são consideradas deteções.

A arquitetura e o sistema integrado que aqui pretende-se desenvolver apresenta por detrás um sistema inteligente para deteção, localização e classificação de múltiplas classes de objetos baseado no trabalho *You Only Look Once* (YOLO) [1]. A equipa por detrás desta metodologia utiliza um método diferente e apenas aplica uma única rede neuronal para a área total da imagem. Esta rede divide a imagem em pequenas regiões e realiza a previsão de limites/anotações para cada uma delas. Estas pequenas anotações são pesadas e consideradas tendo em conta as probabilidades que são obtidas durante o processo de teste.

Esta forma de raciocínio apresenta várias vantagens quando comparada com outros sistemas baseados em classificadores. O modelo analisa a imagem como um todo, logo, as previsões obtidas, alimentam-se do contexto global da imagem. De notar que o ponto principal deste método incide exatamente no facto de apenas utilizar uma única rede neuronal para realizar todo o trabalho, ao invés de sistemas como a R-CNN que necessitam the milhares de redes para processar uma única imagem.

Esta decisão justifica diretamente a possibilidade que existe em obter resultados mil vezes mais rápidos que a arquitetura R-CNN e cem vezes mais rápidos que a implementação *Fast R-CNN*.

A rede que apoia a terceira e última versão apresentada pela equipa por detrás do YOLO, é denominada por Darknet-53, utilizada nomeadamente no processo de extração de características de uma imagem. O seu nome indica o número de camadas convolucionais (53) que constituem este modelo.

	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 33 - Modelo e estrutura de camadas existentes na rede Darknet-53.

O método apresentado pelo YOLO faz com que seja possível obter resultados extremamente rápidos e exatos. Ficam de seguida alguns dados e conclusões que demonstram exatamente o poder do YOLO, desde o seu tempo de execução em GPU até ao nível de precisão obtida quando testada com o conjunto de dados COCO da Microsoft com diferentes configurações.

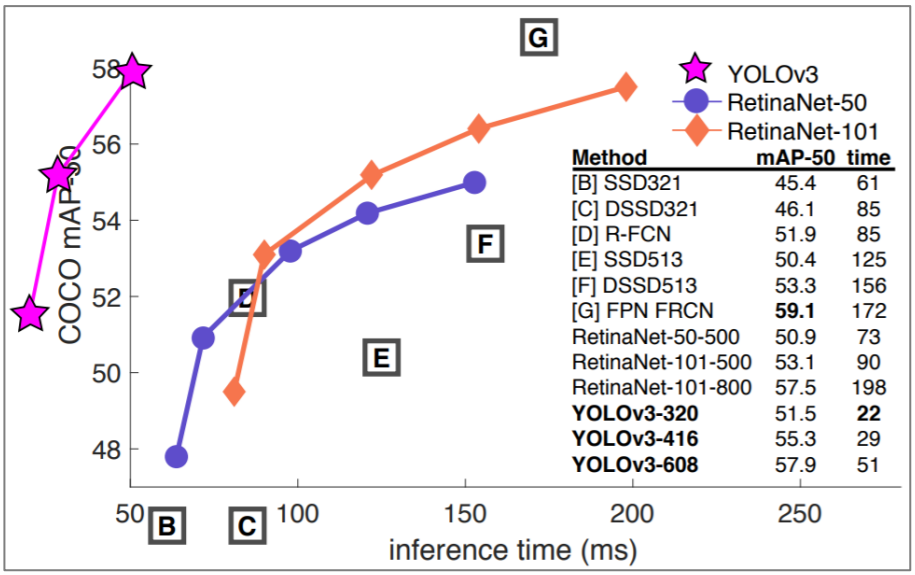

Figura 34 - Tempos de execução para uma única imagem, no processo de deteção, localização e classificação de vários objetos que se encontram distribuídos por múltiplas classes.

DESEMPENHO NO CONJUNTO DE DADOS COCO			
MODELO	mAP	FLOPS	FPS
SSD321	45.4	-	16
DSSD321	46.1	-	12
R-FCN	51.9	-	12
SSD513	50.4	-	8
DSSD513	53.3	-	6
FPN FRCN	59.1	-	6
Retinanet-50-500	50.9	-	14
Retinanet-101-500	53.1	-	11
Retinanet-101-800	57.5	-	5
YOLOv3-320	51.5	38.97 Bn	45
YOLOv3-416	**55.3**	**65.86 Bn**	**35**
YOLOv3-608	57.9	140.69 Bn	20
YOLOv3-tiny	33.1	5.56 Bn	**220**
YOLOv3-spp	**60.6**	141.45 Bn	20

Tabela 12 - Comparação no desempenho e precisão das operações efetuadas por vários modelos de classificadores.

A métrica "mAP" é uma métrica de precisão utilizada como padrão na competição PascalVOC e é exatamente a mesma métrica que encontramos na competição MS COCO onde é denominada por "AP50".

Para melhor compreender estas métricas apresenta-se de seguida um resumo das suas definições e pequenos esquemas para melhor demonstrar o seu conceito quando aplicados no problema da deteção e localização de objetos em imagens:

- **IoU** - *Intersection Of Union* - Média da interseção na união entre os objetos e deteções para um dado *threshold*;
- **mAP -** *Mean Average Precision* - Valor médio obtido pela média das precisões para cada uma das classes existentes, sendo esta última a média de 11 pontos na curva *Precision-Recall* (PR) para cada um dos *thresholds* possíveis (Cada um dos valores de probabilidade obtidos nas deteções) para a mesma classe;

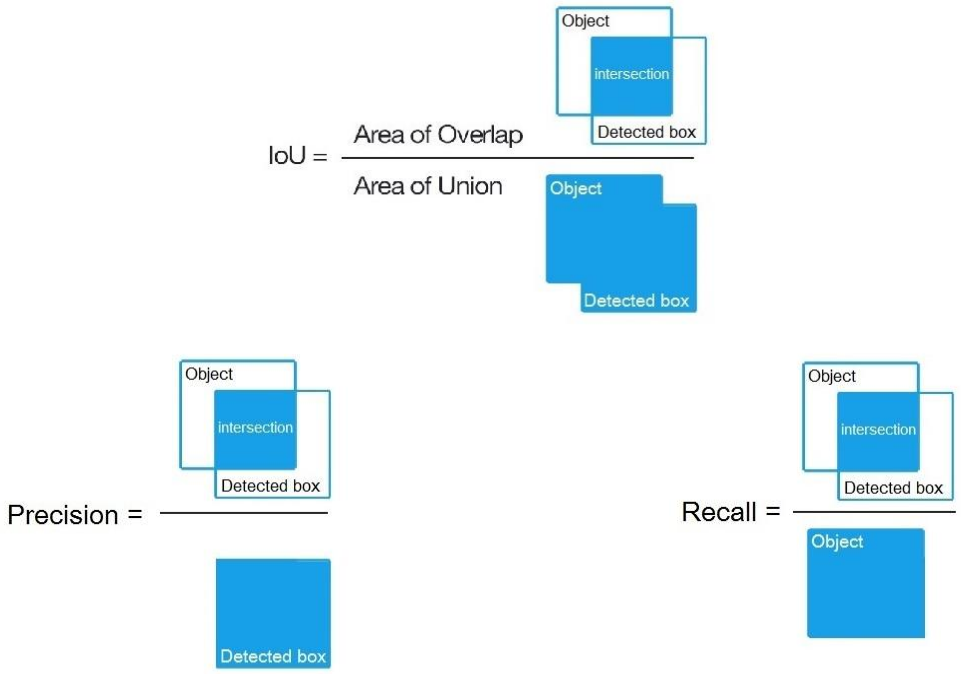

Figura 35 - Esquema de algumas métricas utilizadas para a análise do desempenho de modelos na tarefa de deteção e classificação de objetos numa imagem[21].

[21] https://github.com/AlexeyAB/darknet

Para esta subsecção apenas falta referir que o modelo e configuração selecionada como a base do sistema de deteção, localização e classificação de objetos que se pretende aqui desenvolver e disponibilizar como um serviço é o "**YOLOv3-416**". O número 416 exprime a resolução 416x416 das imagens que são utilizadas para alimentar a rede neuronal convolucional Darknet-53 durante o seu processo de treino e aprendizagem. Salvo dizer que, quanto maior a resolução da imagem, melhor será a exatidão da deteção, tornando-se assim possível localizar objetos mais pequenos e em maior pormenor em cada imagem, mas é importante apontar também que o tempo de processamento será superior e uma menor quantidade de imagens serão possíveis de analisar por segundo (FPS). Note-se que a configuração escolhida surge como um meio-termo entre a precisão e a velocidade de execução quando comparadas com as outras também apresentadas na Tabela 12.

3.4.2 Processo de Treino

Chegando a este ponto e tendo em conta a existência de um servidor HTTP que apresenta já os serviços necessários para alojar e organizar os dados necessários para o processo de deteção, localização e classificação de objetos em imagens, falta apenas compreender qual o próximo passo que é necessário realizar de forma a conseguir aceder às funcionalidades das máquinas inteligentes para previsão e colocar todo este sistema distribuído a trabalhar.

Anteriormente, no ponto 3.3 - Repositório de Dados desta secção, foi possível apresentar a estrutura do sistema de ficheiros que é utilizada para gerir os ficheiros locais do utilizador. Tal como foi também indicado, faltou demonstrar qual a necessidade de criar uma pasta denominada *"learning"* para cada um dos utilizadores do sistema. Após uma breve apresentação da metodologia e arquitetura do sistema inteligente que aqui pretende-se utilizar para resolver o problema do reconhecimento de objetos em imagens, facilmente conclui-se que esta pasta serve exatamente para gerir todos os ficheiros necessários para realizar este feito.

Um dos pontos de acesso HTTP disponíveis remotamente para o utilizador da plataforma DeskTEARS permite despoletar o processo de treino de um modelo para previsão, construído único e exclusivamente tendo em conta os diferentes modelos e anotações existentes para cada utilizador individual.

No momento em que um utilizador decide iniciar o processo de treino de um novo modelo, é necessário cumprir e verificar os seguintes passos:

1. Para cada uma das imagens carregadas pelo utilizador até ao momento, gerar dentro da pasta da própria imagem um ficheiro denominado "original.txt" que introduza em cada linha, de forma muito resumida, as anotações que aí foram realizadas. O formato utilizado para cada uma das entradas deve ser: "[indice_classe] [centro_absoluto_x] [centro_absoluto_y] [largura_absoluta] [altura_absoluta]";

2. Gerar um ficheiro "tears.data" que servirá para definir alguns metadados necessários para o processo de treino. Os atributos a definir são: número de classes, caminho para um ficheiro "tears.names" que liste as diferentes *labels* das classes existentes (o número de linha corresponde ao índice da classe = num_linha - 1), caminho para um ficheiro "train.txt" que liste os caminhos para as diferentes imagens que se pretende utilizar como imagens de treino, caminho para um ficheiro "test.txt" que liste os caminhos para as diferentes imagens que se pretende utilizar como imagens para teste e por fim o caminho para uma pasta "backup" que irá servir para armazenar os diferentes ficheiros temporários que surgem ao longo das várias fases de treino e iterações realizadas, nomeadamente os pesos definidos para cada uma das camadas de neurónios constituintes da rede neuronal;

3. Gerar os ficheiros e pasta referenciados em "tears.data": "tears.names", "train.txt", "test.txt" e "backup";

4. Gerar o ficheiro de configuração da rede neuronal a utilizar baseado no "**YOLOv3-416**". Neste ponto deverá ser tido em conta o número de classes existente de forma a atualizar o número de filtros utilizados e tamanho do *input* que é utilizado por *default*;

3.4.3 Classificação de uma única Imagem

Caso o utilizador saiba agora aceder corretamente à plataforma DeskTEARS e utilizar as funcionalidades aí presentes para introduzir novos objetos na plataforma, realizar o carregamento de imagens e fazer a introdução manual de anotações, é possível concluir e validar que este cumpre com os requisitos necessários para poder treinar um modelo, ficando assim apto para realizar a deteção, localização e classificação de objetos numa qualquer imagem.

Com o propósito de testar o sistema desenvolvido até esta fase, foi criado um ponto de acesso HTTP remoto para o utilizador da plataforma DeskTEARS, de forma a permitir que o utilizador envie uma imagem para os serviços WebTEARS e obtenha como resposta a mesma imagem, mas devidamente anotada com os diferentes objetos que foram possíveis detetar.

De notar que este ponto de acesso não permite realizar a deteção, localização e classificação de imagens em tempo real devido aos limites impostos pelo protocolo de comunicações, tal como já foi possível discutir em 3.2.3 - Comunicações em Tempo Real por utilização de WebSockets.

Para cumprir com o pedido realizado pelo utilizador, assim que os serviços WebTEARS recebem a imagem que o utilizador enviou, um novo processo é iniciado no servidor com uma máquina de previsão que é criada tendo em conta os ficheiros de metadados criados anteriormente e o ficheiro de pesos obtido durante o processo de treino. De seguida, basta alimentar esta máquina com a própria imagem e esperar pelo resultado. No momento em que obtém-se a lista de deteções, é da responsabilidade dos serviços existentes gerar uma imagem onde se sobreponha as deteções pela utilização de anotações e um sistema de cores próprio de forma a corresponder uma única cor a cada *label* de objeto existente.

Segue um exemplo em baixo pela utilização de pesos pré-treinados com o conjunto de dados disponibilizados pelo COCO da Microsoft:

Figura 36 - Quadro "*A Friend in Need*", pintura de C. M. Coolidge em 1903.

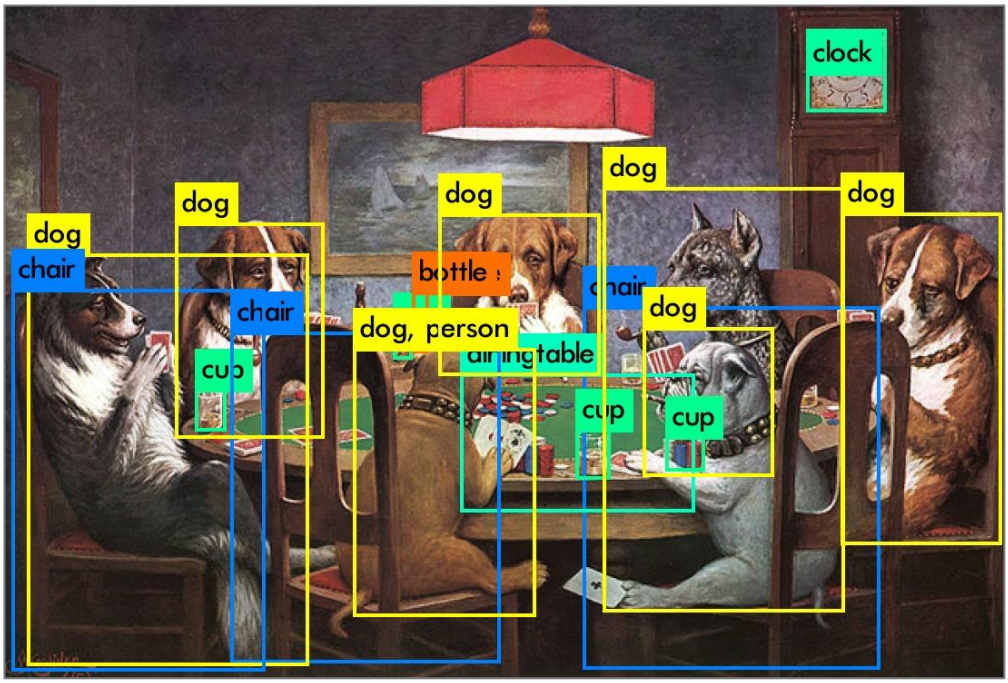

Figura 37 - Previsões obtidas por utilização do modelo YOLO e pesos pré-treinados com o conjunto de dados disponibilizados pelo COCO da Microsoft.

3.4.4 Previsão em Tempo Real

Embora já seja possível realizar a deteção, localização e classificação de objetos numa imagem por recurso a uma máquina de previsão inteligente baseada no modelo de uma única rede neuronal convolucional, falta ainda cumprir com o objetivo de o fazer em tempo real.

Uma das razões que leva à impossibilidade de realizar o processo de previsão em tempo real é precisamente o facto de os serviços de comunicações existentes apenas permitirem, até ao momento, realizar a previsão através de um ponto de acesso HTTP. De forma a contornar este problema, e tendo em conta que o módulo para permitir um tipo de comunicações em tempo real através de *WebSockets* já se encontra desenvolvido nesta fase, basta pegar na mesma lógica que é realizada para um pedido HTTP e modificar-la um pouco de forma a conseguir cumprir com o mesmo objetivo num ambiente de comunicações SOCKET.

Algumas alterações deverão ser tidas em conta. Uma delas, não tão óbvia, passa pelo facto que, num pedido HTTP, sempre que uma nova imagem é enviada pelo utilizador, é instanciada uma nova máquina para previsão num novo processo que é criado pelo servidor, e como consequência, era já possível prever que o tempo de carregamento dos ficheiros de configuração e a construção de um novo classificador fosse um tanto demorada. Uma forma de contornar este problema é exatamente a de criar uma máquina de previsão, não no momento em que chega uma nova imagem, mas assim quando o utilizador realiza o processo de autenticação através da utilização de protocolos de transporte *WebSocket* e se permaneça aí ligado. Desta forma, quando o utilizador começar a enviar as imagens que pretende prever, não será necessário criar e iniciar uma nova máquina sempre que uma *frame* dá entrada nos serviços do WebTEARS.

Uma última alteração a realizar tem a ver com o formato da resposta que é devolvida ao utilizador. Visto que o utilizador final que irá tirar partido da existência deste módulo de comunicações não tem como objetivo trabalhar com uma imagem anotada (UnityTEARS), não é necessário devolver como resposta uma imagem, mas sim os dados formatados relativos às deteções que se obtém diretamente da máquina de previsões.

3.5 Análise do Ambiente

3.5.1 Mapeamento Espacial

Um longo passo foi dado no desenvolvimento do projeto realizado para esta dissertação. Tendo agora acesso a serviços que disponibilizam os dados do utilizador, serviços que permitem realizar o processo de deteção, localização e classificação de múltiplas classes de objetos (WebTEARS) e uma plataforma para gerir e manipular os dados necessários para alimentar toda a cadeia das camadas e módulos existentes nesta arquitetura (DeskTEARS), falta agora construir a plataforma que vai permitir trazer a verdadeira utilidade ao resultado das deteções obtidas nas máquinas utilizadas para previsão.

A primeira tarefa a completar com o desenvolvimento da plataforma do UnityTEARS é decidir exatamente qual o objetivo desta plataforma e ao mesmo tempo perceber quais as ferramentas, sensores e tecnologias que aqui é possível utilizar.

Com o acesso a uma tecnologia como a do Hololens pretende-se tirar partido do poder da RM e dar vida às deteções que conseguimos obter remotamente através da utilização do serviço de reconhecimento disponibilizado. Um dos primeiros passos a realizar para conseguir atingir esse objetivo é exatamente a de obter, por utilização dos sensores disponibilizados pelo dispositivo, uma reconstrução e representação tridimensional do mundo que rodeia o utilizador, e tudo isso ser feito em tempo real.

A equipa da Microsoft disponibiliza bibliotecas que permitem a análise e processamento dos dados dos sensores que são possíveis encontrar no Hololens. Os sensores de profundidade disponíveis serão possivelmente aqueles que demonstram maior impacto e interesse para o problema que temos em mãos. Desta forma, foi criado um módulo com o único propósito de extrair e processar os dados destes sensores, nomeadamente os mapas de profundidade, e de certa forma transformar e disponibilizar-los numa estrutura mais simplificada apenas constituída pelos vários blocos de nuvens de pontos e as consequentes superfícies que daqui resultam.

Fica de seguida um exemplo de uma possível visualização que é possível obter das superfícies que rodeiam o utilizador deste dispositivo. De referir que a seguinte imagem foi obtida por utilização do ambiente disponibilizado pela plataforma UnityTEARS em modo *debug*, permitindo desta forma visualizar as *meshes* das múltiplas superfícies que são obtidas durante o processo de mapeamento do espaço.

Figura 38 - Uma simples demonstração e exemplo de um certo espaço e objetos que rodeiam o utilizador do dispositivo Hololens.

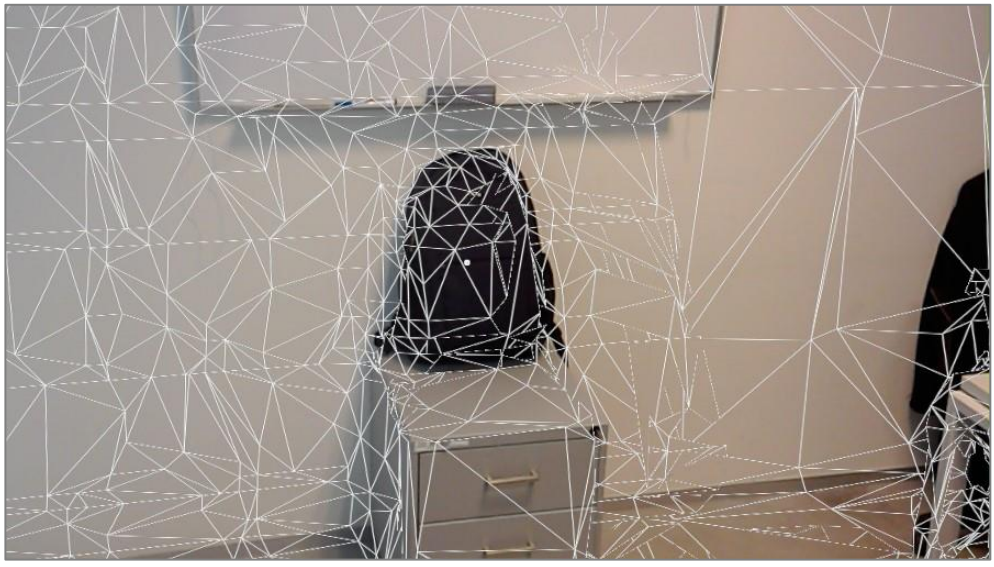

Figura 39 - Visualização das *meshes* associadas às superfícies que são obtidas através de técnicas de mapeamento espacial pela utilização do dispositivo Hololens.

3.5.2 *Streaming* de Dados

Para ser possível tirar partido dos serviços disponibilizados no servidor WebTEARS, é preciso primeiro desenvolver um cliente que vá de encontro com a lógica de utilização do protocolo de transporte por utilização de *WebSockets* e ainda realizar os passos corretos para concluir a autenticação e abrir um túnel de comunicações entre o utilizador e o serviço de deteção, localização e classificação disponível no servidor.

Tendo este facto como garantido é agora preciso criar um módulo para gerir as imagens do vídeo que é possível obter da câmara embutida no dispositivo do Hololens. De notar que a ideia é exatamente a de enviar para o servidor as imagens obtidas pela câmara do dispositivo num dado momento, quase como se os serviços aqui disponibilizados se tornassem os olhos do próprio utilizador, respondendo, o mais rápido possível, com aquilo que as máquinas de previsão conseguem "realmente ver".

Para conclusão desta subsecção foram criados dois modos de funcionamento para o módulo em questão:

- Para demonstrar o processamento em tempo real das imagens obtidas pela câmara do dispositivo do Hololens, existe um modo que permite realizar um pedido de previsão a cada 'X' segundos (e.g Obter uma previsão a cada 100ms = 10 FPS);

- Para um modo mais interativo entre o utilizador e o espaço que o rodeia, foi desenvolvido um módulo que tira partido da biblioteca da Microsoft para o reconhecimento de gestos. Desta forma, assim que um utilizador prende o seu olhar num qualquer objeto que pretenda identificar e realiza o gesto de um clique no ar, é enviado um novo pedido de previsão para os serviços de reconhecimento (e.g. Obter uma previsão a cada clique);

3.5.3 Anotação Tridimensional

Provavelmente uma das fases mais interessantes e empolgantes para com o desenvolvimento deste projeto, a possibilidade de realizar a anotação tridimensional de objetos surge exatamente com a necessidade de criar técnicas que permitam utilizar as deteções bidimensionais obtidas por recurso às

máquinas de previsão disponibilizadas pelos serviços do WebTEARS e colocar-las no mundo tridimensional que rodeia o utilizador da plataforma UnityTEARS.

Uma das bibliotecas que é disponibilizada de forma nativa pelo motor gráfico utilizado para gerar o mundo virtual que aqui é visualizado através do dispositivo do Hololens, dispõe de algoritmos e técnicas que permitem a sintetização de modelos físicos tridimensionais. Ou seja, através do disparo de raios fictícios que partem de um observador é possível definir um caminho numa dada direção que pode colidir com os vários objetos que se encontram no espaço (*raycasting*).

Juntando esta nova possibilidade ao facto de já termos acesso, em tempo real, a uma reconstrução tridimensional das superfícies físicas que rodeiam o utilizador, falta apenas compreender qual a solução que aqui foi utilizada para, de certa forma, conseguir transformar as deteções que são obtidas em imagens com duas dimensões e colocar-las no domínio tridimensional através de processos de *raycasting*.

O formato da mensagem que é devolvida ao utilizador por parte dos serviços de reconhecimento no servidor do WebTEARS e que contém a lista das previsões obtidas, é o seguinte:

```
id: "TEARS-20AMP0MUSH9WISE"
predictions: [
        {
                label: "chair",
                probability: 0.987642,
                x: 0.25,
                y: 0.33,
                width: 0.2,
                height: 0.5
        },
        ...
]
```

No momento em que um novo pedido de previsão é realizado por parte do utilizador, é necessário cumprir e verificar os seguintes passos:

1. Obter novo *frame* apartir da *stream* do vídeo que é obtido pela câmara do dispositivo do Hololens;

2. Gerar um identificador único para a *frame* obtida;

3. Guardar o **momento** do vector da posição e direção da câmara virtual que surge como a representação do utilizador do dispositivo no espaço virtual;

4. Associar o **momento** ao *frame* da câmara através do identificar único;
5. Enviar um pedido para o serviço de deteção, localização e classificação, enviando no corpo da mensagem o identificador único da *frame* obtida e o conteúdo da própria imagem em formato *base64*;

6. Assim que uma resposta é devolvida pelo servidor, processar a previsão tendo sempre como associado o **momento** em que a respetiva *frame* para esse pedido foi obtida, através do identificar único que é devolvido outra vez ao utilizador;

7. Realizar um primeiro processo de *raycasting* para calcular a superfície mais próxima do utilizador;

8. Realizar o cálculo das dimensões dos planos do *frustum* à distância a que a superfície mais próxima se encontra;

9. Tendo agora uma boa referência e um plano com que trabalhar no espaço tridimensional, partindo da posição x e y da previsão obtida, calcular o verdadeiro ponto no espaço tridimensional onde se encontra o centro do objeto detetado, realizando um novo processo de *raycasting*;

10. Partindo do tamanho *width* e *height* dos limites previstos pela localização, calcular o tamanho real do corpo do objeto detetado no espaço tridimensional;

11. Colocar o modelo de uma *label* 3D no ponto espacial obtido anteriormente;

12. Colocar uma esfera invisível neste mesmo ponto (Com raio = min(*realWidth*, *realHeight*) * 0.5) que irá servir como um objeto obstáculo onde os raios procedentes de futuros *raycastings* poderão colidir. Este elemento é muito importante já que permite que o utilizador não coloque *labels* repetidas em locais com objetos que já foram identificados anteriormente;

De seguida, fica uma pequena demonstração do resultado que é possível obter quando o utilizador realiza o gesto de um clique no ar para identificar um objeto que encontra-se à sua frente:

Figura 40 - Exemplo da colocação de uma *label* no espaço tridimensional pela utilização das funcionalidades presentes no ambiente UnityTEARS e serviços do servidor WebTEARS no dispositivo *Hololens*.

É de apontar que encontram-se no Anexo III - Plataforma UnityTEARS várias imagens de capturas de ecrã que demonstram as várias funcionalidades da plataforma UnityTEARS que foram discutidas até este momento.

3.6 Ferramentas

ATOM - Um editor de texto *hackable*, contruído especialmente para o século 21;

NODE.JS - Ambiente de execução *JavaScript* construído a partir do mecanismo *JavaScript V8* do *Chrome*;

EXPRESS - Estrutura *web* minimalista, rápida e unipinionada para o *Node.js*;

POSTGRESQL - A base de dados relacional de código aberto mais avançada do mundo;

SOCKET.IO - O mecanismo para comunicações em tempo real mais rápido e confiável do mundo;

DARKNET - A *Darknet* é uma *framework* de redes neuronais de código aberto escrita em *C* e *CUDA*. É rápida, fácil de instalar e suporta realizar qualquer tipo de computação em *CPU* e/ou *GPU*;

CUDA - É uma plataforma de computação paralela e um modelo de programação desenvolvido pela NVIDIA para realizar qualquer tipo de processamento em unidades de processamento gráfico (GPUs). Com o CUDA, os desenvolvedores podem acelerar drasticamente os aplicativos de computação, aproveitando o poder das GPUs;

CUDNN - É uma biblioteca constituída por várias primitivas que permitem acelerar o processamento das GPUs em ambientes onde se requer a utilização de redes neurais para *deep learning*. As bibliotecas cuDNN fornecem implementações altamente ajustadas e afinadas para quaisquer rotinas padrão, como por exemplo *feedforward* e *backpropagation* de camadas convolucionais, *pooling*, normalização e camadas de ativação;

UNITY - É o principal motor gráfico em tempo real do mundo e é utilizado para criar metade dos jogos conhecidos da atualidade. As ferramentas flexíveis e em tempo real oferecem incríveis possibilidades para os desenvolvedores de jogos e criadores em vários setores e aplicativos criados em 2D, 3D, RV e RA;

MIXEDREALITYTOOLKIT - É uma coleção de *scripts* e componentes destinados a acelerar o desenvolvimento de aplicativos destinados aos dispositivos *Microsoft HoloLens* e ambientes *Windows Mixed Reality*. Este projeto visa reduzir as barreiras para os desenvolvedores que desejam criar aplicativos de realidade mista e, de certa forma, contribuir de volta para a comunidade à medida que vão crescendo;

3.7 Resultados

Tal como já foi referido no início desta secção, o objetivo principal com a realização deste projeto é a de apoiar um utilizador que encontra-se imersivo em ambientes de realidade mista através do reconhecimento das múltiplas classes de objetos que rodeiam o dispositivo aqui utilizado, o *Hololens* da *Microsoft*.

Tal como foi especulado na primeira fase de desenvolvimento deste projeto, dificilmente seria possível explorar o *hardware* presente no dispositivo do Hololens para realizar a tarefa local de, a partir da análise de imagem e dados de profundidade, realizar a deteção, localização e reconhecimento de objetos, em tempo real. Foi possível concluir e realizar este tipo de julgamento logo no momento em que a desenvolvimento deste módulo iniciou-se. Para realizar este tipo de processamento localmente, o dispositivo do *Hololens* teria que simultaneamente processar os dados do vídeo da câmara, proceder à análise das imagens, extrair as deteções, realizar o reconhecimento, de algum modo armazenar toda a informação e ainda conseguir realizar o *rendering* dos vários hologramas necessários para a anotação, e fazer tudo isto em tempo real.

Como o nível de risco já era elevado o suficiente só com a simples necessidade existente em recorrer à utilização do dispositivo do Hololens, foi decidido que este processamento adicional para realizar o armazenamento dos dados e a previsão de múltiplas classes de objetos por análise de imagem em tempo real, seria disponibilizada como um serviço.

Esta decisão trouxe só vantagens, não só no sentido que ganha-se alguma folga nos recursos a que temos acesso, mas também por outro facto não tão visível à primeira vista que é a compatibilidade que ganhou-se para com outros dispositivos, que não o Hololens. Sendo assim é de esperar que estes também podem, teoricamente, ter acesso a este serviço.

Com tudo isto exposto, é importante apontar o cuidado que foi necessário ter com o planeamento deste projeto. Para fechar esta subsecção, conclui-se que todos os objetivos listados foram cumpridos. Com o desenvolvimento deste projeto disponibiliza-se assim um sistema completamente integrado e personalizado para o utilizador com o objetivo de detetar e reconhecer as múltiplas classes de objetos que o rodeiam, através da anotação tridimensional em ambientes de realidade aumentada.

No final de tudo, imagine-se agora nas possibilidades existentes com o simples acesso a dados como a posição e tamanho desses mesmos objetos. Se uma máquina conseguir interpretar o espaço físico que rodeia o utilizador, passamos a conseguir uma maior coerência na introdução de modelos virtuais 3D no mundo físico, podemos criar guias virtuais em museus e outros espaços, criar assistentes virtuais inteligentes que deslocam-se e interagirem com o mundo tal como o conhecemos, revolucionar o mundo da industria e produção e até mesmo do entretenimento, ao colocar o utilizador a interagir com aquilo que o rodeia, não existem mesmo limites para aquilo que agora é possível.

3.8 Sumário

Logo mesmo no inicio da redação da presente dissertação, era ainda muito vaga a ideia do que era pretendido realizar para demonstração e prova de conceito de tudo aquilo que viria a ser explorado e discutido ao longo deste documento.

Os primeiros avanços sentidos com o desenvolvimento do estado de arte na exploração dos campos de visão por computador, reconhecimento de imagem e os vários conceitos de RV, RA e RM permitiram concluir, juntamente com a tecnologia disponibilizada pelo dispositivo disponível do Hololens da Microsoft, a necessidade que existe em criar novas ideias e de melhorar várias soluções já existentes de forma a alimentar as possibilidades que são possíveis de atingir pela utilização deste mesmo tipo de dispositivos.

Um ponto que ajudou a encaminhar o processo de modelação e conceptualização do projeto aqui apresentado terá sido exatamente o facto de que uma nova versão dos óculos de realidade mista do *Hololens* estaria para surgir. Através da realização de algumas pesquisas, facilmente conclui-se que aquilo que a empresa da Microsoft procura melhorar nessa versão vai de encontro com a introdução de vários conceitos de visão por computador, inteligência artificial, aprendizagem máquina e *deep learning*.

Tendo em conta que também, pessoalmente, uma das áreas que mostra ser mais empolgante e interessante envolve os vários conceitos de compreensão e reconhecimento máquina daquilo que rodeia o utilizador através da análise de dados de sensores, tudo o que foi discutido até este ponto parece fazer sentido.

Apesar de ter-se atingido praticamente todos os objetivos traçados, várias foram as dificuldades que surgiram ao longo do desenvolvimento, muitas delas relacionadas diretamente com o facto de que, no período em que esta dissertação foi realizada, o dispositivo do Hololens, embora uma tecnologia pioneira na sua área, surge como uma solução vagamente incompleta. Muito ainda falta explorar e desenvolver para realmente conseguirmos tirar partido destas tecnologias. Embora muitas empresas dentro da indústria e mercado tecnológico tenham avançado com a sua utilização na criação de novas soluções e produtos, é importante ter em conta o risco elevado que ainda existe.

As bibliotecas, módulos e documentação existente é ainda um tanto escassa tendo em conta todo o *marketing* realizado pela empresa da Microsoft sobre as possibilidades que deveriam ser possíveis de extrair desta tecnologia. Este mesmo ponto levou à necessidade de estudar a mais baixo nível o *SDK* de desenvolvimento do Windows e os sensores que se encontram embutidos no dispositivo do Hololens. Juntando a este facto surge ainda a impossibilidade de reaproveitar bibliotecas e módulos para a manipulação dos vários dados dos sensores e outras possíveis funcionalidades do Hololens, visto que são de pouca qualidade ou praticamente inexistentes.

É sempre importante referir que um trabalho deste género é sempre difícil de conseguir chocar com a possibilidade de se tornar num produto completo. O projeto que foi aqui apresentado cumpriu com todos os seus objetivos, mas claro, muito ainda poderia ter sido feito.

Uma das maiores limitações deste projeto vai de encontro diretamente com o servidor *web* desenvolvido (WebTEARS) para realizar o processo de deteção, localização e classificação de múltiplas classes de objetos e disponibilizar-lo com um serviço. É preciso considerar o facto de que realizar tal processamento envolve a necessidade de uma estrutura e máquina apropriada, que pode envolver custos elevados. Uma possibilidade para um desenvolvimento futuro seria exatamente a de encontrar resposta ao facto de como seria possível criar uma melhor distribuição e utilização dos recursos nos serviços disponibilizados. Será que seria possível colocar algum peso do lado do cliente? Seria mais fácil escalar a solução com a introdução de um sistema com filas de espera? Seria interessante introduzir uma hierarquia na estrutura das máquinas existentes? Seria possível tirar um maior proveito do processamento em CPU e não sobrecarregar a utilização do GPU?

Isto tudo para referir que neste momento, o sistema demonstra ser muito pouco eficiente na utilização dos recursos de memória do GPU e isso pode limitar o número de utilizadores simultaneamente conectados aos serviços que aqui foram expostos e discutidos, e como tal, a escalabilidade da solução demonstra-se fraca. Outros pontos que poderiam igualmente ser discutidos vão de encontro com a qualidade e adição de algumas funcionalidades nas plataformas que, para esta demonstração, demonstraram não ser necessárias, mas num ambiente real de utilização deveriam existir (e.g. Gestão de conta do utilizador, menus adicionais nas interfaces, painel de ajuda e suporte, etc.).

Por fim, gostaria de fechar esta secção referindo que um possível desenvolvimento a ter no futuro seria exatamente o de complementar esta framework de forma a que fosse possível escolher e ter acesso a diferentes modelos e sistemas inteligentes, e que o seu papel não fosse apenas o de reconhecer objetos, mas o de analisar os muitos outros aspetos e elementos que habitam o ambiente que nos rodeia, como por exemplo a pose dos objetos, as ações a decorrer, etc. e não nos limitarmos apenas à imagem, mas também decorrer a outros dados como o som, a temperatura, a luz, a localização, entre outros.

4. CONCLUSÃO

Percorreu-se um longo caminho até aqui. Não existe nada melhor que a sensação de progressão, e melhor que isso só mesmo o impacto em terminar algo, conseguir concretizar aquilo que nos obrigou a despender parte do nosso valioso tempo, que nos fez suar, que nos roubou noites de sono e descanso, que nos obrigou a dizer não a muita coisa. De tudo aquilo que planeamos é simplesmente um enorme entusiasmo ver que no fim, conseguimos finalmente concluir aquilo pelo qual lutámos fazer. A conclusão serve mesmo para isto, para podermos discutir, admirar e absorver tudo aquilo que aqui foi realizado.

Tendo em conta todos os sumários que foram realizados ao longo deste documento, pouco mais existe para relatar nesta última secção. O importante a retirar daqui será mesmo referir que toda a informação que foi explorada e selecionada contribuiu sempre para o próprio crescimento pessoal.

A visão por computador é uma área que se encontra em grande expansão, tudo aquilo que acreditamos saber hoje em dia e que ainda se discute no meio científico, rapidamente poderá tornar-se inválido de um dia para o outro, mas não podemos logo conotar este tipo de acontecimentos como algo negativo, é importante perceber que estas situações permitem, de certa forma, abrir espaço para novas teorias e metodologias que possam surgir. Ou seja, permite-nos evoluir.

O mesmo se poderá dizer das várias técnicas, conceitos e quaisquer metodologias e até mesmo tecnologias que aqui foram discutidas e no fim utilizadas. Tudo isto apenas para dizer que todas as pesquisas realizadas e decisões tomadas permitiram traçar um caminho e chegar a um estado de arte atual que surge o suficientemente completo para as necessidades que aqui se apresentaram e permitiu sempre sustentar todas as decisões e objetivos traçados para o desenvolvimento do projeto *Tangible Environments in Augmented Reality Systems* (TEARS).

Para prova de conceito e demonstração final, foi formulado uma arquitetura distribuída constituída por três sistemas e plataformas distintas. Numa primeira instância foi preparada uma estrutura denominada WebTEARS pronta a realizar todo o tipo de comunicações HTTP/TCP através de uma REST API e *WebSockets*. Para armazenar os ficheiros que viriam a ser necessários para o funcionamento de toda a arquitetura aqui presente, procedeu-se à modelação de uma base de dados integral, juntamente com toda uma lógica, estrutura e hierarquia de modo a permitir o alojamento local dos ficheiros e

imagens que viriam a alimentar os sistemas inteligentes também presentes nos serviços aqui demonstrados.

De forma a gerir e manipular todos os dados aqui nomeados junta-se à coleção uma outra plataforma denominada DeskTEARS. O utilizador ganha assim o poder de gerir os objetos que gostaria de reconhecer através dos serviços prestados, tendo ele também um papel importante no sistema através da anotação manual das imagens carregadas que servem como apoio e supervisão no processo de treino e aprendizagem realizado pelo sistema inteligente existente, com o único objetivo de realizar a deteção, localização e classificação das múltiplas classes de objetos aqui presentes.

Por fim, mas não menos importante de tudo aquilo que já pôde ser enunciado, surge a plataforma UnityTEARS. Podemos olhar para esta plataforma como a cereja no topo do bolo, ou seja, o resultado obtido até ao momento já era suficiente para ir de encontro com o estudo que aqui se pretendia fazer sobre os processos de extração de dados, análise das características e essência de uma imagem e reconhecimento de objetos, mas aquilo que realmente tornou esta solução única foi a utilização de tecnologias de processamento e *rendering* holográfico. Trabalhar com este tipo de tecnologias para realizar o mapeamento espacial de superfícies físicas e poder visualizar o resultado das previsões do reconhecimento inteligente através de anotações virtuais no meio físico foi sem dúvida o ponto mais alto deste projeto.

Tudo isto serviu como um pequeno resumo de tudo aquilo que foi realizado com esta dissertação. Salvo dizer que houve vários problemas e dificuldades ao longo deste trabalho, mas acredito que estes obstáculos surgiram com o único propósito de incentivar e explorar a motivação que existia em mim.

Muitos dos atrasos existentes estão diretamente interligados com o facto de utilizar-se tecnologias e dispositivos com *hardware* que apresenta neste momento um risco elevado, no sentido que são disponibilizados de momento apenas para prototipagem e modelação de pequenas ideias e projetos. A falta de documentação e o acesso a exemplos escassos ou de pouca qualidade comprometeram um pouco o trabalho que foi realizado, mas nunca, nenhuma destas razões levaram-me a desistir.

Finalmente, gostava de concluir que tudo aquilo que foi idealizado inicialmente para com a realização desta dissertação correu dentro dos modos esperados. Sirva o projeto apresentado como

prova de tudo aquilo que é possível realizar com os avanços tecnológicos que se sentem na área da visão nos dias de hoje. Nunca nada foi impossível e cabe-nos a nós pequenos jovens cientistas e investigadores dar continuidade na melhoria e evolução de tudo aquilo que foi construído pelos nossos antecedentes.

Quero assim terminar dizendo apenas que espero continuar a fazer aquilo que sempre quis e fiz de melhor, aprender.

5. BIBLIOGRAFIA

[1] J. Redmon e A. Farhadi, "YOLOv3: An Incremental Improvement," p. 6, 8 Abril 2018.

[2] P. Milgram e F. Kishino, "A Taxonomy of Mixed Reality Visual Displays," *IEICE Transactions on Information and Systems,* Vols. %1 de %2E77-D, pp. 1321-1329, 1994.

[3] R. M. Haralick e L. G. Shapiro, "Glossary Of Computer Vision Terms," *Pattern Recognition,* vol. 24, pp. 69-93, 1991.

[4] V. Cantoni, S. Levialdi e V. Roberto, Artificial Vision: Image Description, Recognition and Communication, 1 ed., Massachusetts, Cambridge: Academic Press, 1996, p. 306.

[5] E. Guerra e J. Villalobos, "A Three-Dimensional Automated Visual Inspection System For SMT Assembly," *Computers & Industrial Engineering,* vol. 40, pp. 175-190, 2001.

[6] I. B. Gurevich e I. Koryabkina, "Comparative Analysis and Classification of Features for Image Models," *Pattern Recognition and Image Analysis,* vol. 16, p. 265–297, 2006.

[7] A. Pinz e R. Bartl, "Information fusion in image understanding," *Proceedings. 11th IAPR International Conference on Pattern Recognition. Vol.1. Conference A: Computer Vision and Applications,* pp. 366-370, 1992.

[8] D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, Cambridge, Massachusetts, EUA: MIT Press, 1982.

[9] R. Klette, Concise Computer Vision - An Introduction into Theory and Algorithms, London: Springer, 2014.

[10] W. T. Freeman, "Where computer vision needs help from computer science," *ACM-SIAM Symposium on Discrete Algorithms, SODA 2011,* pp. 814-819, 2011.

[11] M. Datar, N. Immorlica, P. Indyk e V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," em *Nearest Neighbor Methods in Learning and Vision: Theory and Practice*, Massachusetts, Cambridge: MIT Press, 2006.

[12] C. Barnes, D. B. Goldman, E. Shechtman e A. Finkelstein, "The PatchMatch Randomized Matching Algorithm for Image Manipulation," *Communications of the ACM,* vol. 54, pp. 103-110, 2011.

[13] M. Muja e D. G. Lowe, "Fast Approximate Nearest Neighboars with Automatic Algorithm Configuration," *VISAPP 2009: Lisboa, Portugal,* pp. 331-340, 2009.

[14] J. B. Tenenbaum, T. L. Griffiths e C. Kemp, "Theory-based Bayesian models of inductive learning and reasoning," *TRENDS in Cognitive Sciences,* vol. 10, p. 309–318, 2006.

[15] I. Biederman, "Recognition-by-Components: A Theory of Human Image Understanding," *Psychological Review,* vol. 94, pp. 115-147, 1987.

[16] Y. Boykov, O. Veksler e R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *Proceedings of the Seventh IEEE International Conference on Computer Vision,* vol. 1, pp. 377-384, 1999.

[17] J. S. Yedidia, W. T. Freeman e Y. Weiss, "Understanding Belief Propagation and its Generalizations," em *Exploring Artificial Intelligence in the New Millenium*, Massachusetts, Burlington: Morgan Kaufmann Publishers Inc., 2002, pp. 239-269.

[18] M. J. Wainwright, T. S. Jaakkola e A. S. Willsky, "Exact MAP estimates by (hyper)tree agreement," em *Advances in Neural Information Processing Systems 15*, Cambridge, Massachusetts: MIT Press, 2003, p. 833–840.

[19] P. Kohli, L. Ladický e P. H. S. Torr, "Robust Higher Order Potentials for Enforcing Label Consistency," *International Journal of Computer Vision,* vol. 82, p. 302–324, 2009.

[20] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen e C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, pp. 1068-1080, 2008.

[21] P. F. Felzenszwalb e D. P. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision,* vol. 70, p. 41–54, 2006.

[22] V. Lempitsky, P. Kohli, C. Rother e T. Sharp, "Image segmentation with a bounding box prior," *International Conference on Computer Vision,* pp. 277-284, 2009.

[23] J. Portilla e E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision,* vol. 40, p. 49–70, 2000.

[24] S. Roth e M. Black, "Fields of experts: A framework for learning image priors," *IEEE Computer Vision and Pattern Recognition,* vol. 2, pp. 860-867, 2005.

[25] Y. Weiss e W. T. Freeman, "What makes a good model of natural images?," *IEEE Computer Vision and Pattern Recognition,* pp. 1-8, 2007.

[26] A. A. Efros e W. T. Freeman, "Image quilting for texture synthesis and transfer," *Proceedings of the 28th annual conference on Computer graphics and interactive techniques,* pp. 341-346, 2001.

[27] A. A. Efros e T. Leung, "Texture synthesis by non-parametric sampling," *Proceedings of the Seventh IEEE International Conference on Computer Vision,* vol. 2, pp. 1033-1038, 1999.

[28] S. M. Smith e J. M. Brady, "SUSAN - A New Approach to Low Level Image Processing," *International Journal of Computer Vision,* vol. 23, pp. 45-78, 1997.

[29] I. Biederman, R. J.Mezzanotte e J. C. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cognitive Psychology,* vol. 14, pp. 143-177, 1982.

[30] S. S. Farfade, M. J. Saberian e L.-J. Li, "Multi-view Face Detection Using Deep Convolutional Neural Networks," *ICMR '15 Proceedings of the 5th ACM on International Conference on Multimedia Retrieval,* pp. 643-650, 2015.

[31] M. Fischler e R. Elschlager, "The Representation and Matching of Pictorial Structures," *IEEE Transactions on Computers,* Vols. %1 de %2C-22, pp. 67-92, 1973.

[32] M. Turk e A. Pentland, "Face recognition using eigenfaces," *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 586-591, 1991.

[33] H. A. Rowley, S. Baluja e T. Kanade, "Human Face Detection in Visual Scenes," *Advances in Neural Information Processing Systems 8,* pp. 875-881, 1995.

[34] F. Fleuret e D. Geman, "Graded Learning for Object Detection," *Proceedings of the IEEE Workshop on Statistical and Computational Theories of Vision,* 1999.

[35] P. Viola e M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision,* vol. 57, p. 137–154, 2004.

[36] B. Heisele, T. Serre, S. Mukherjee e T. Poggio, "Feature reduction and hierarchy of classifiers for fast object detection in video images," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 2, pp. 18-24, 2001.

[37] M. Aeberhard, S. Rauch, M. Bahram, G. Tanzmeister, J. Thomas, Y. Pilat, F. Homm, W. Huber e N. Kaempchen, "Experience, Results and Lessons Learned from Automated Driving on Germany's Highways," *IEEE Intelligent Transportation Systems Magazine,* vol. 7, pp. 42-57, 2015.

[38] S. Ren, L. Lu, L. Zhao e H. Duan, "Circuit board defect detection based on image processing," *2015 8th International Congress on Image and Signal Processing,* pp. 899-903, 2015.

[39] A. K. P., S. N. S. e S. K. V., "A Review of PCB Defect Detection Using Image Processing," *International Journal of Engineering and Innovative Technology,* vol. 4, pp. 188-192, 2015.

[40] N. V. Hung, L. C. Tran, N. H. Dung, T. M. Hoang e N. T. Dzung, "A traffic monitoring system for a mixed traffic flow via road estimation and analysis," *2016 IEEE Sixth International Conference on Communications and Electronics,* pp. 375-378, 2016.

[41] G. B. Keller, T. Bonhoeffer e M. Hübener, "Sensorimotor Mismatch Signals in Primary Visual Cortex of the Behaving Mouse," *Neuron,* vol. 74, pp. 809-815, 2012.

[42] P. Perona e J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 12, pp. 629-639, 1990.

[43] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision,* vol. 60, p. 91–110, 2004.

[44] H. Das, Telemanipulator and Telepresence Technologies, Boston, Massachusettes: SPIE PRESS - The International Society for Optical Engineering, 1994.

[45] D. Foyle, A. Andre, R. McCann, E. Wenzel, D. Begault e V. Battiste, "Taxiway Navigation and Situation Awareness (T-NASA) System: Problem, Design Philosophy, and Description of an Integrated Display Suite for Low-Visibility Airport Surface Operations," *SAE Transactions: Journal of Aerospace,* vol. 105, pp. 1411-1418, 1997.

[46] H. Eschena, T. Köttera, R. Rodecka, M. Harnischa e T. Schüppstuhla, "Augmented and Virtual Reality for Inspection and Maintenance Processes in the Aviation Industry," *Procedia Manufacturing,* vol. 19, pp. 156-163, 2018.

[47] Y. Zhou, H. Luo e Y. Yang, "Implementation of Augmented Reality for Segment Displacement Inspection During Tunneling Construction," *Automation in Construction,* vol. 82, pp. 112-121, 2017.

[48] J. P. Lima, R. Roberto, F. Simões, M. Almeida, L. Figueiredo, J. M. Teixeira e V. Teichrieba, "Markerless Tracking System for Augmented Reality in the Automotive Industry," *Expert Systems with Applications,* vol. 82, pp. 100-114, 2017.

[49] P. Vavra, J. Roman, P. Zonca, P. Ihnat, M. Nemec, J. Kumar, N. A. Habib e A. El-Gendi, "Recent Development of Augmented Reality in Surgery: A Review," *Journal of Healthcare Engineering,* vol. 2017, pp. 1-9, 2017.

[50] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama e K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," *CoRR,* 2016.

[51] J. Redmon, S. Divvala, R. Girshick e A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *CoRR,* 2015.

[52] R. He, J. Rojas e Y. Guan, "A 3D Object Detection and Pose Estimation Pipeline Using RGB-D Images," *CoRR,* 2017.

[53] S. Ren, K. He, R. Girshick e J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *CoRR,* 2015.

[54] G. Gkioxari, R. Girshick, P. Dollár e K. He, "Detecting and Recognizing Human-Object Interactions," *2017,* CoRR.

[55] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan e S. Belongie, "Feature Pyramid Networks for Object Detection," *CoRR,* 2016.

[56] K. He, G. Gkioxari, P. Dollár e R. Girshick, "Mask R-CNN," *CoRR,* 2017.

[57] R. Girshick, J. Donahue, T. Darrell e J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR,* 2013.

[58] K. E. A. v. d. Sande, J. R. R. Uijlings, T. Gevers e A. W. M. Smeulders, "Segmentation as selective search for object recognition," *International Conference on Computer Vision,* pp. 1879-1886, 2011.

ANEXO I - PLATAFORMA WEBTEARS

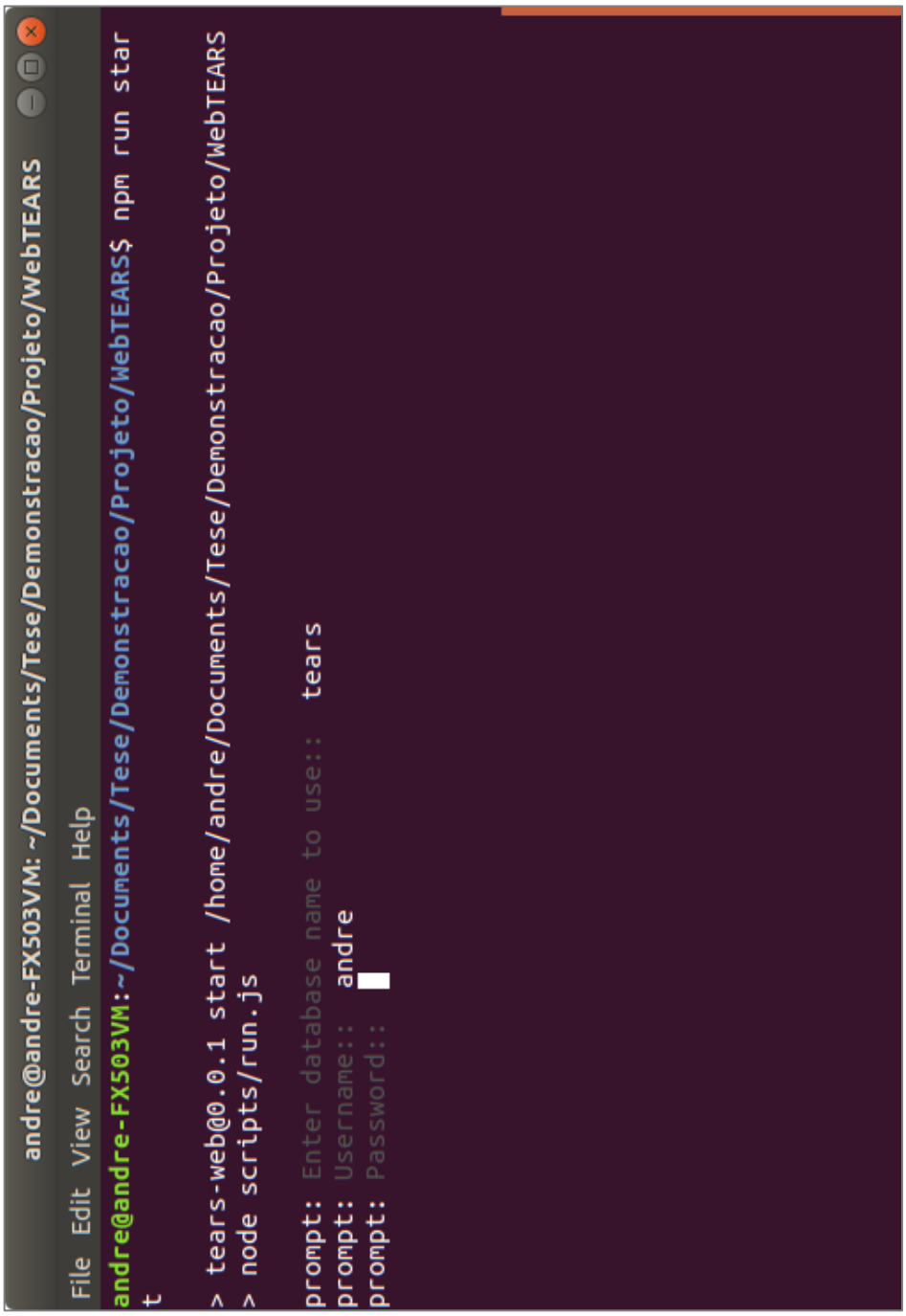

Figura 41 - Processo de arranque em terminal do servidor WebTEARS.

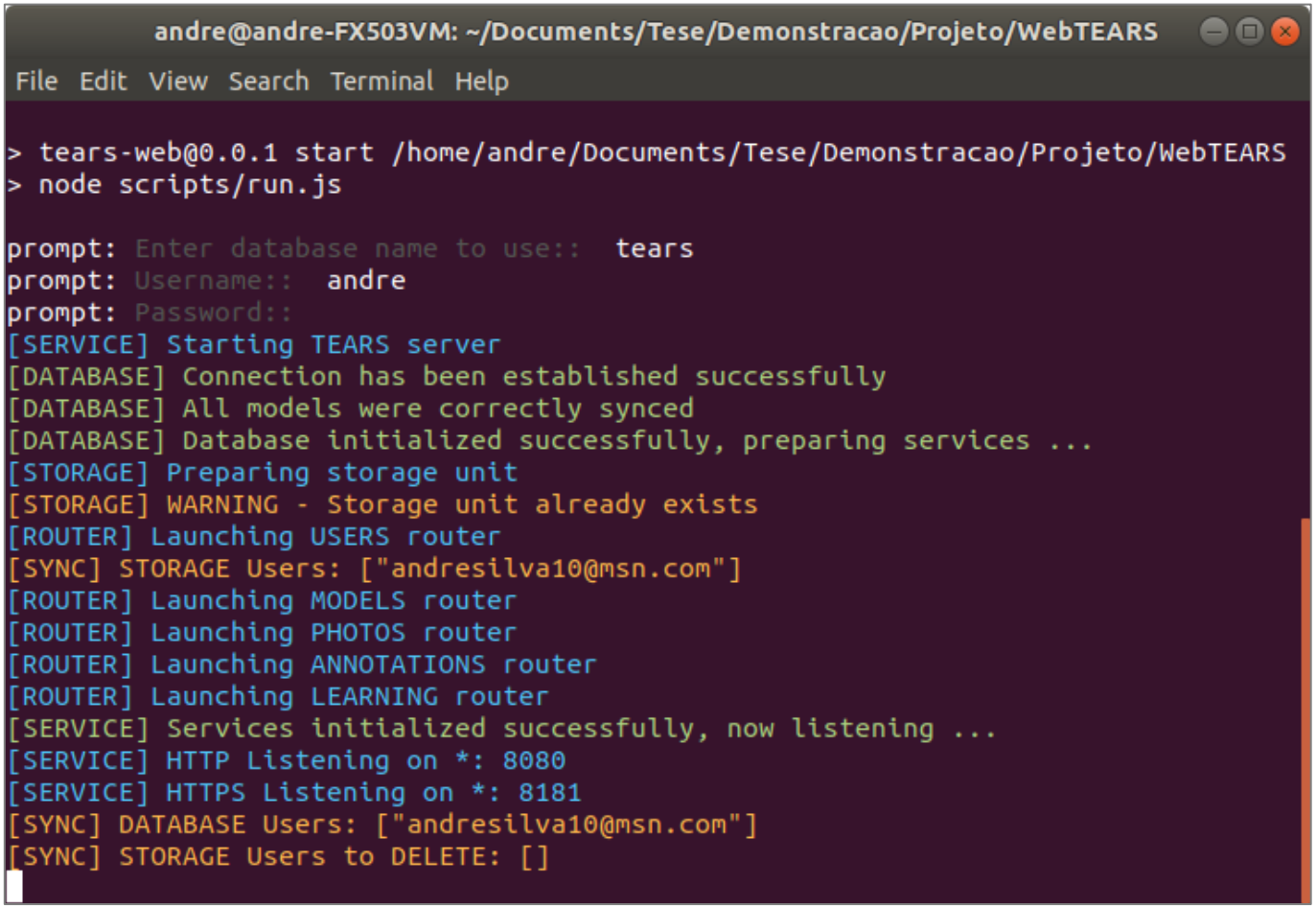

Figura 42 - Iniciação dos vários módulos e serviços presentes no servidor WebTEARS.

Figura 43 - Execução e listagem de vários pedidos a serem realizados por parte de utilizadores ao servidor WebTEARS.

ANEXO II - PLATAFORMA DESKTEARS

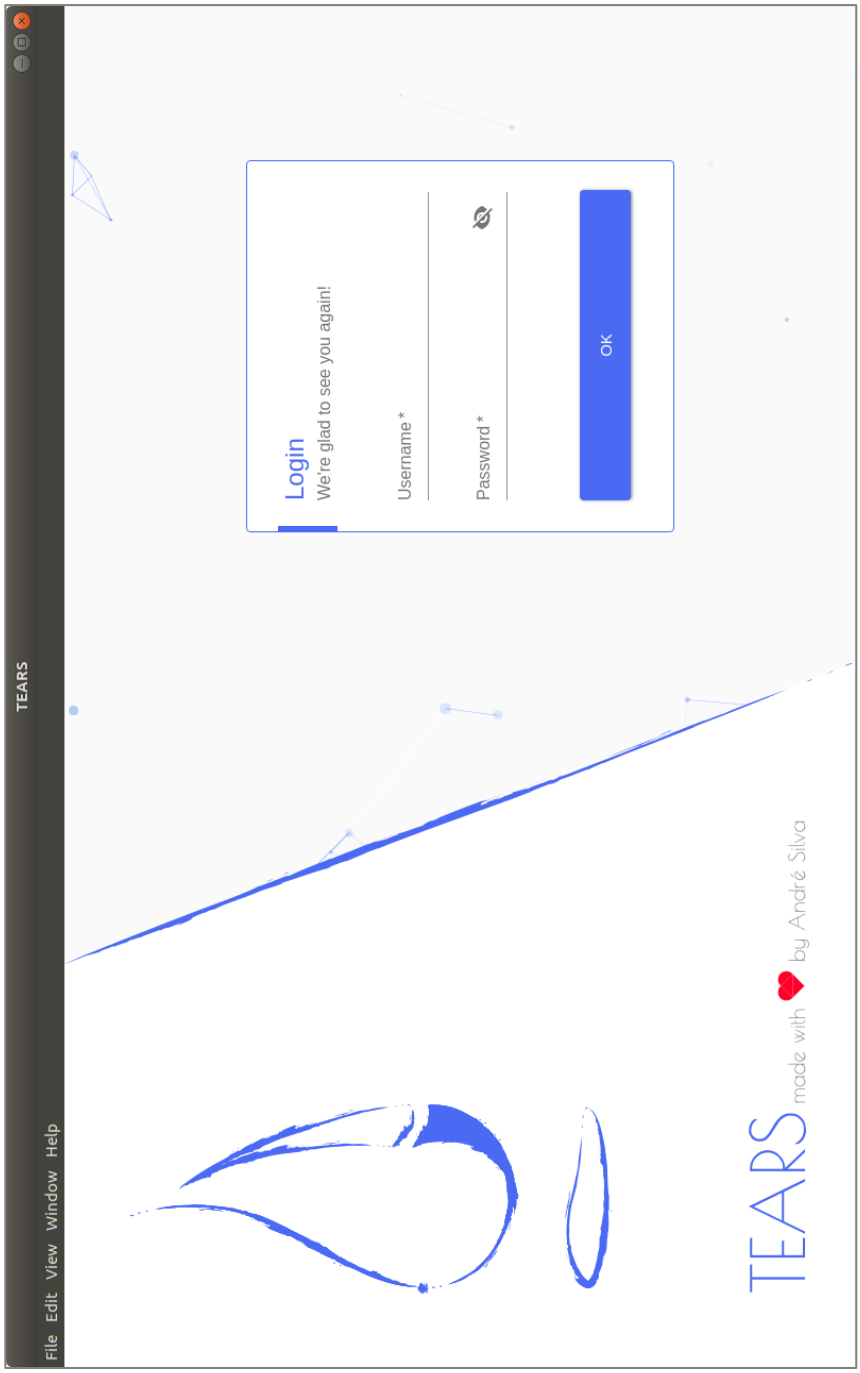

Figura 44 - Página inicial da plataforma DeskTEARS.

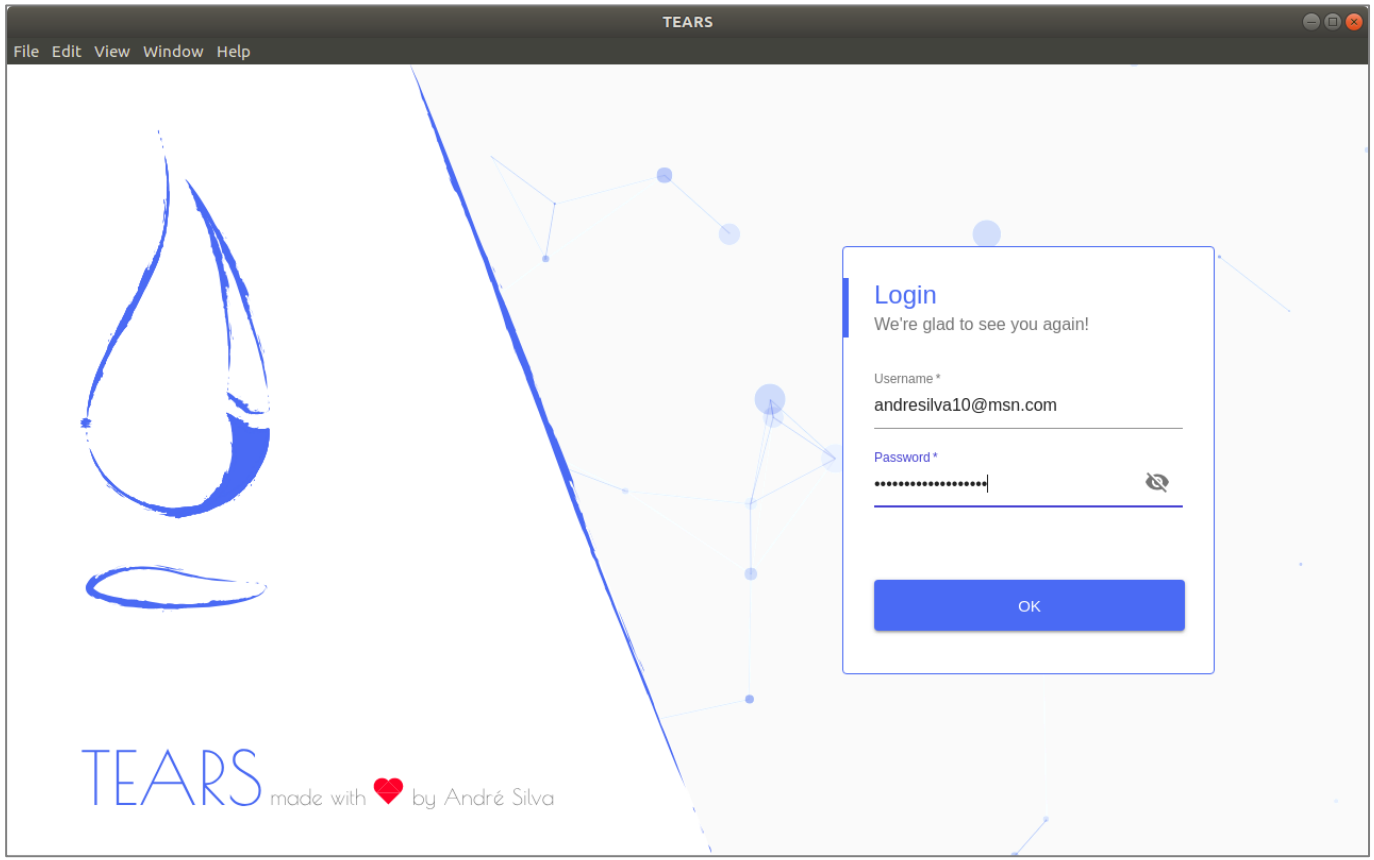

Figura 45 - Processo de autenticação do utilizador na plataforma DeskTEARS.

Figura 46 - Listagem dos objetos personalizados do utilizador na plataforma DeskTEARS.

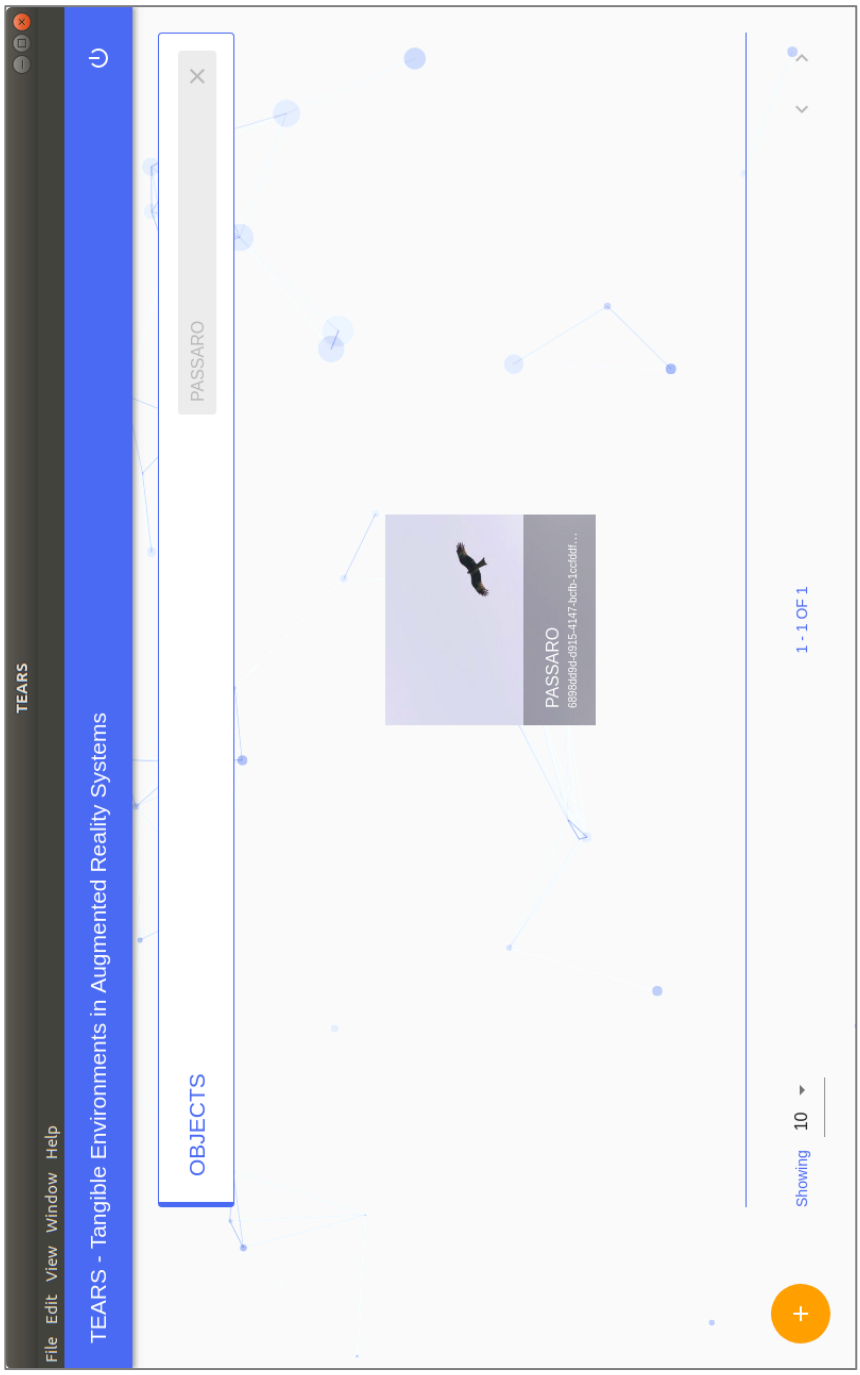

Figura 47 - Demonstração do pequeno sistema de pesquisa e filtro dos objetos existentes na plataforma DeskTEARS do utilizador.

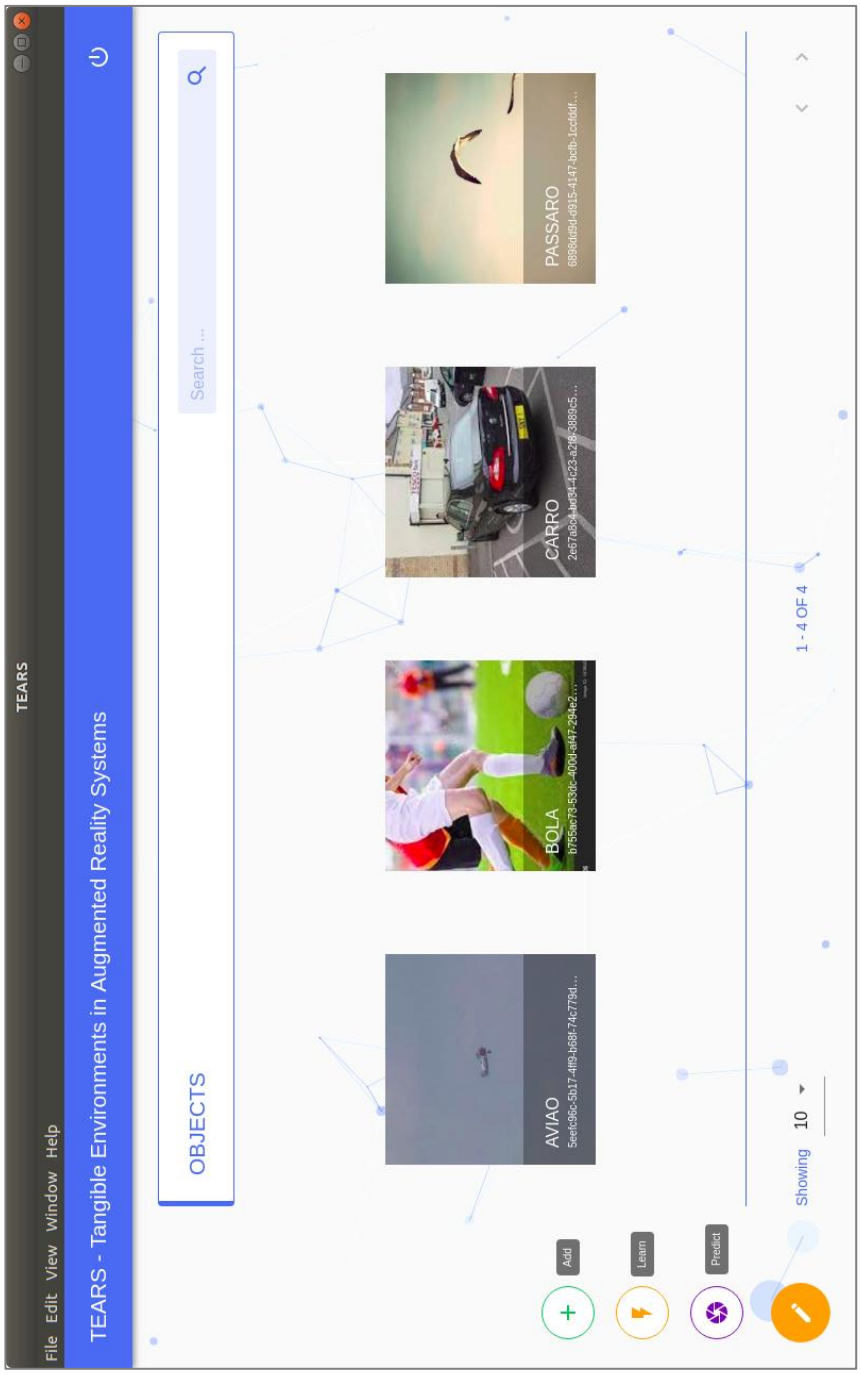

Figura 48 - Menu lateral com funcionalidades que permitem manipular a máquina de previsão do utilizador e adicionar novos objetos à plataforma DeskTEARS.

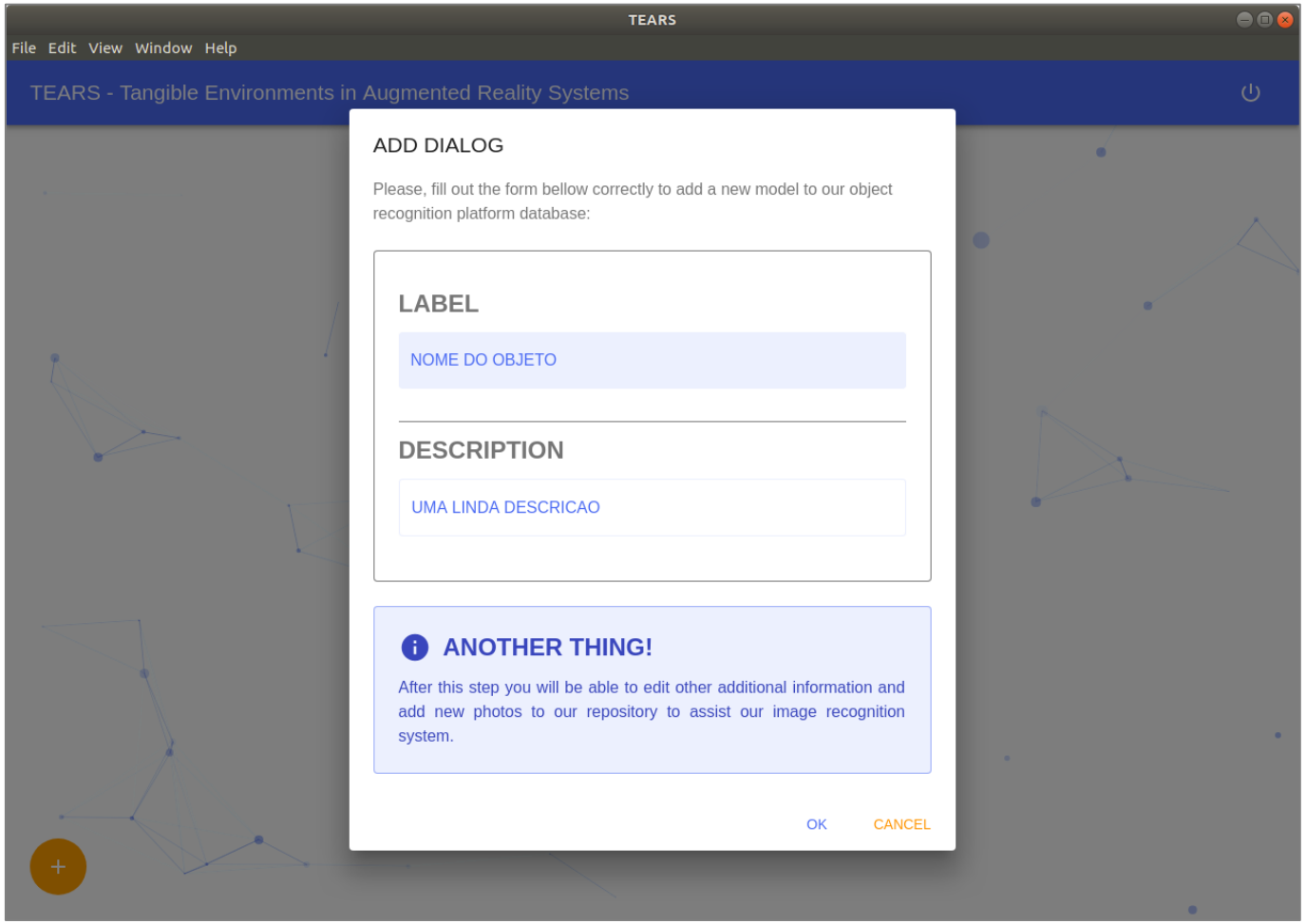

Figura 49 - Exemplo de formulário a preencher inicialmente para adicionar um novo objeto à plataforma DeskTEARS.

Figura 50 - Visualização em detalhe da informação relativa a um qualquer objeto que seja selecionado na plataforma DeskTEARS.

Figura 51 - *Popup* informativo sobre alterações que foram realizadas nos dados informativos do objeto ao o utilizador tentar fechar a janela.

Figura 52 - *Popup* informativo que surge no momento de remoção de um qualquer objeto da plataforma DeskTEARS.

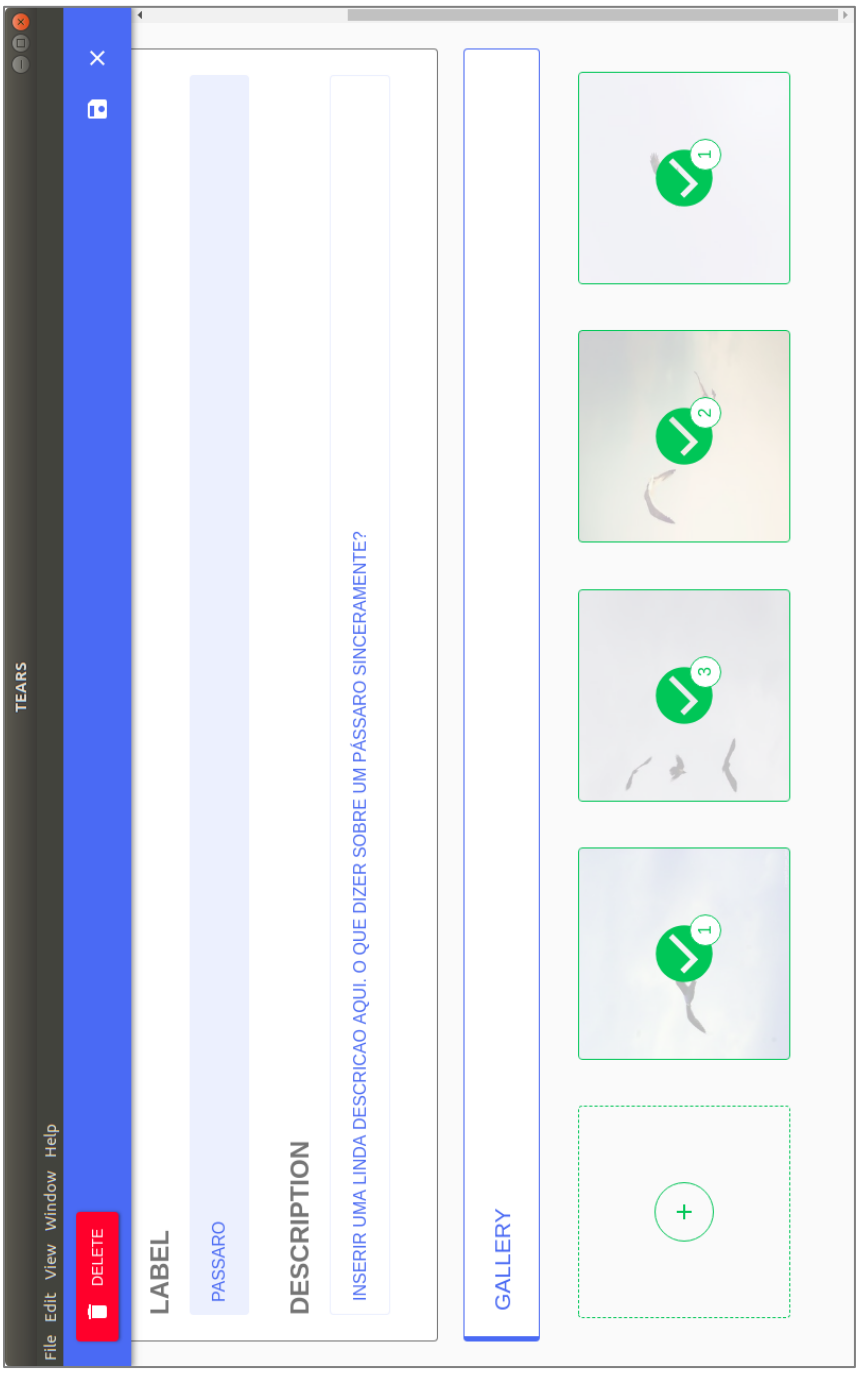

Figura 53 - Possível visualização da galeria de imagens presente no objeto selecionado.

Figura 54 - Menu adicional que surge em cada umas imagens da galeria que permite visualizar ou remover a imagem selecionada.

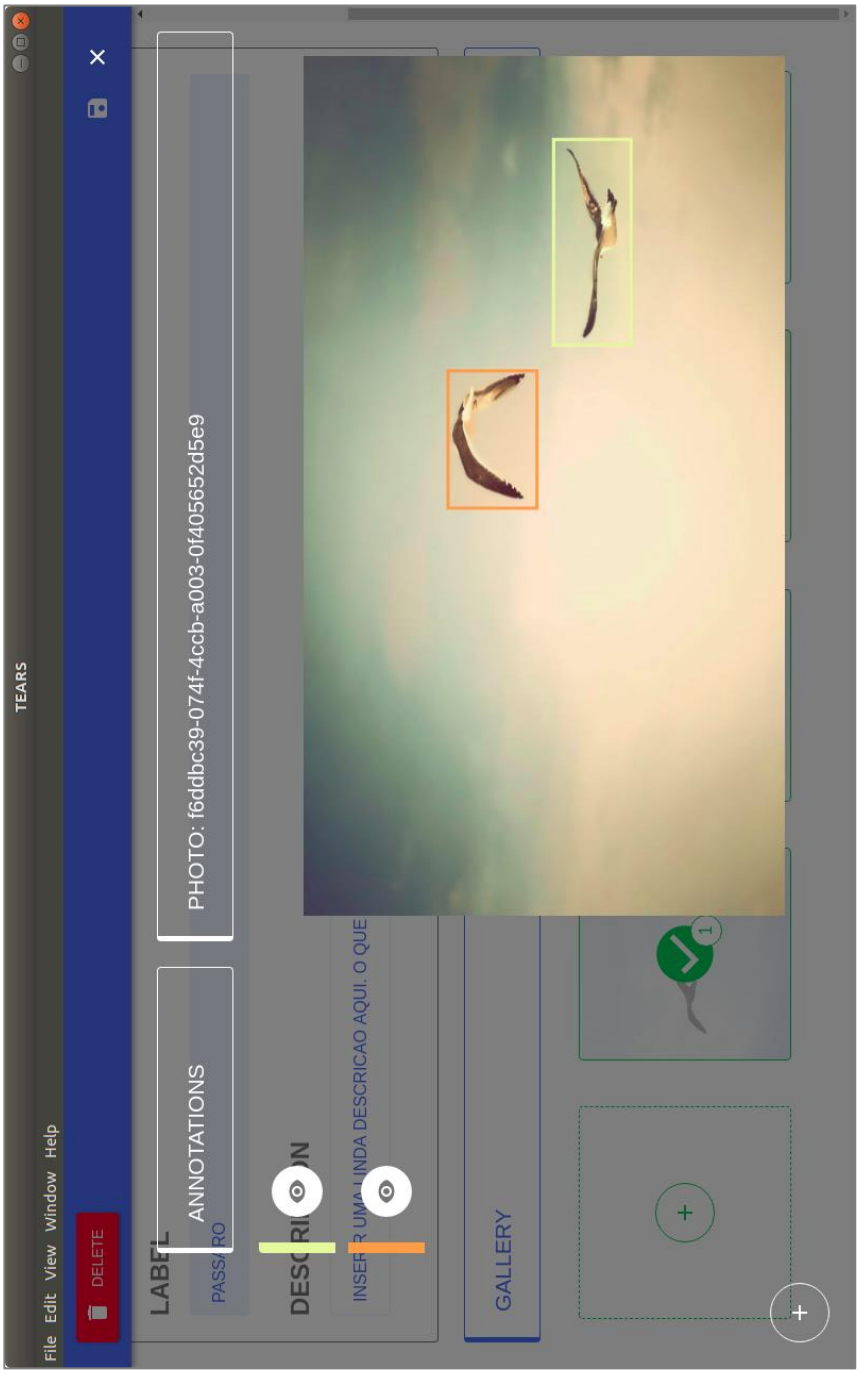

Figura 55 - Visualização da imagem selecionada onde é possível realizar a listagem das anotações realizadas aí e igualmente gerir e manipular as mesmas.

ANEXO III - PLATAFORMA UNITYTEARS

Figura 56 - Exemplo de deteção, localização e classificação de uma cadeira no ambiente que rodeia o utilizador.

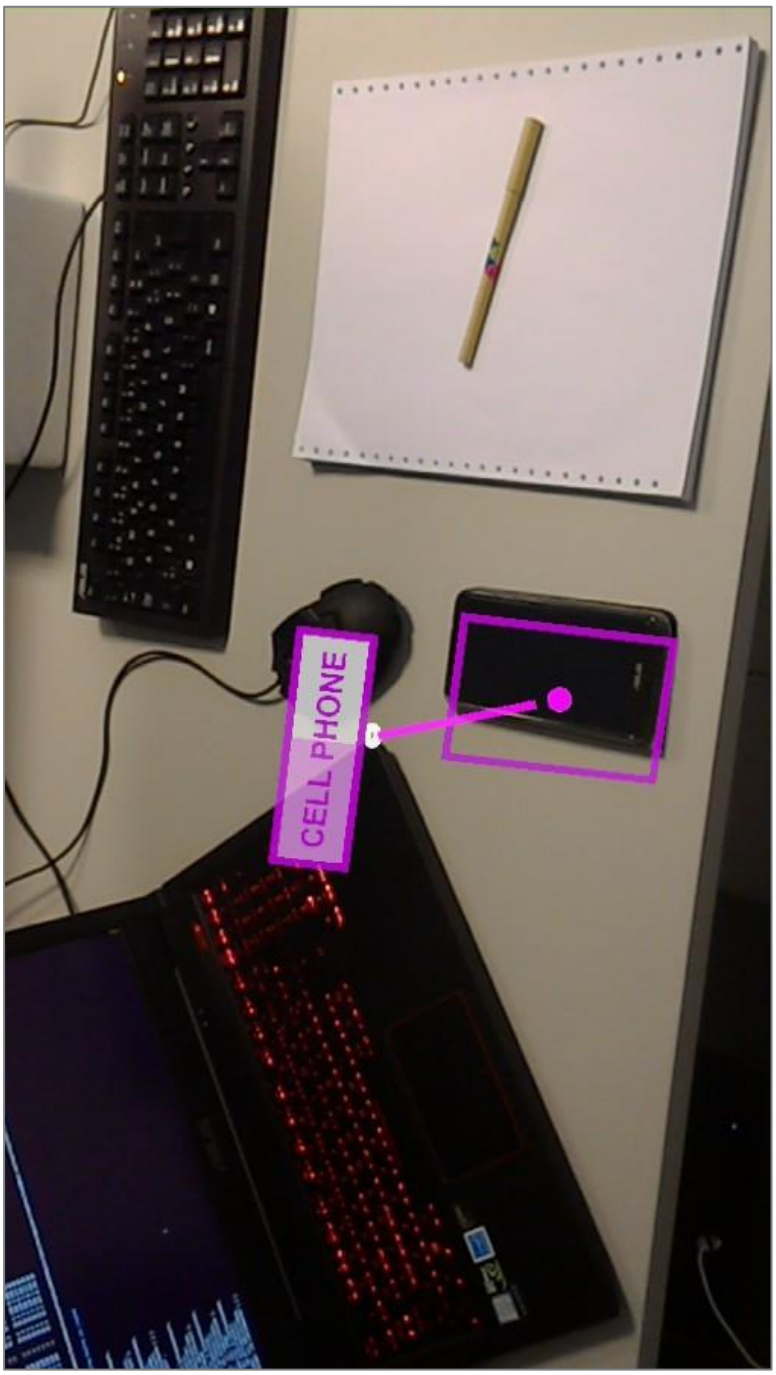

Figura 57 - Exemplo de deteção, localização e classificação de um telemóvel no ambiente que rodeia o utilizador.

Q

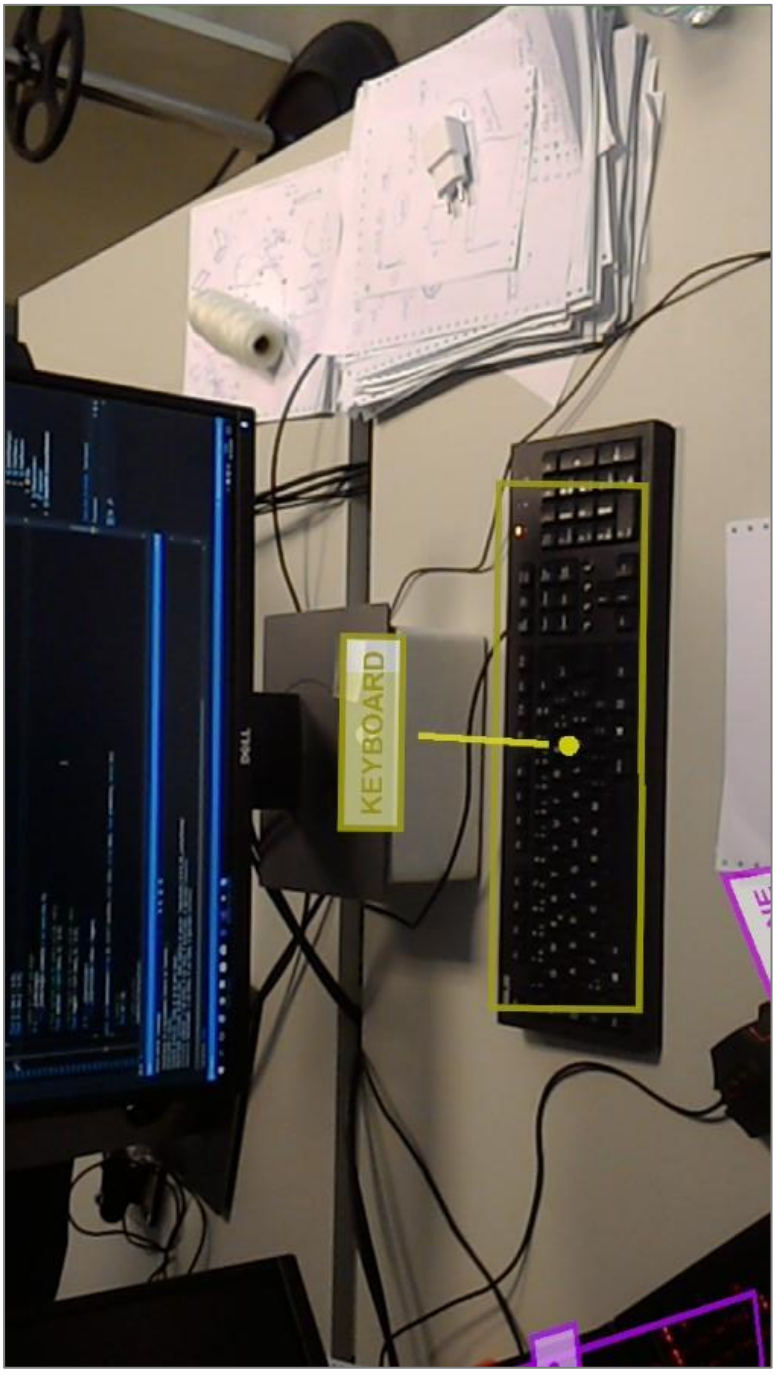

Figura 58 - Exemplo de deteção, localização e classificação de um teclado no ambiente que rodeia o utilizador.

R

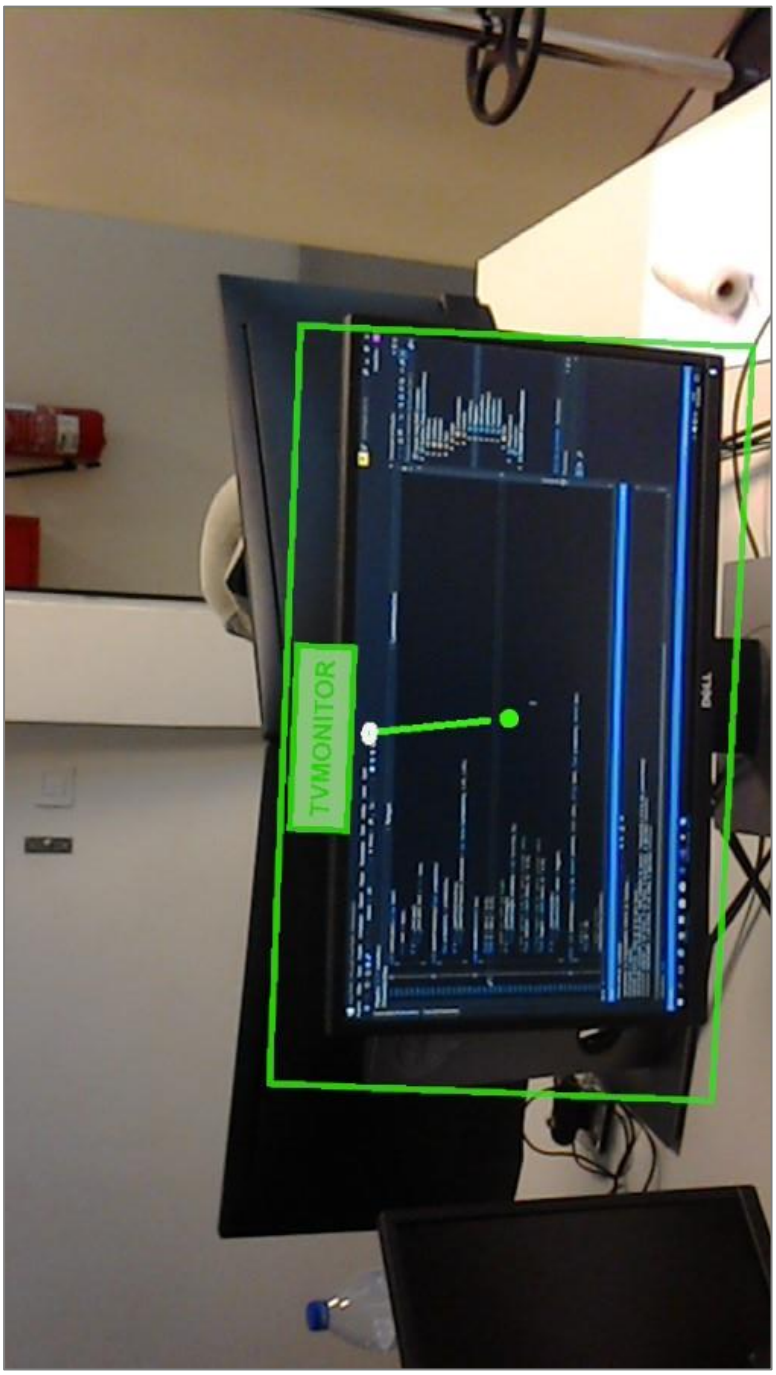

Figura 59 - Exemplo de deteção, localização e classificação de um monitor no ambiente que rodeia o utilizador.

S

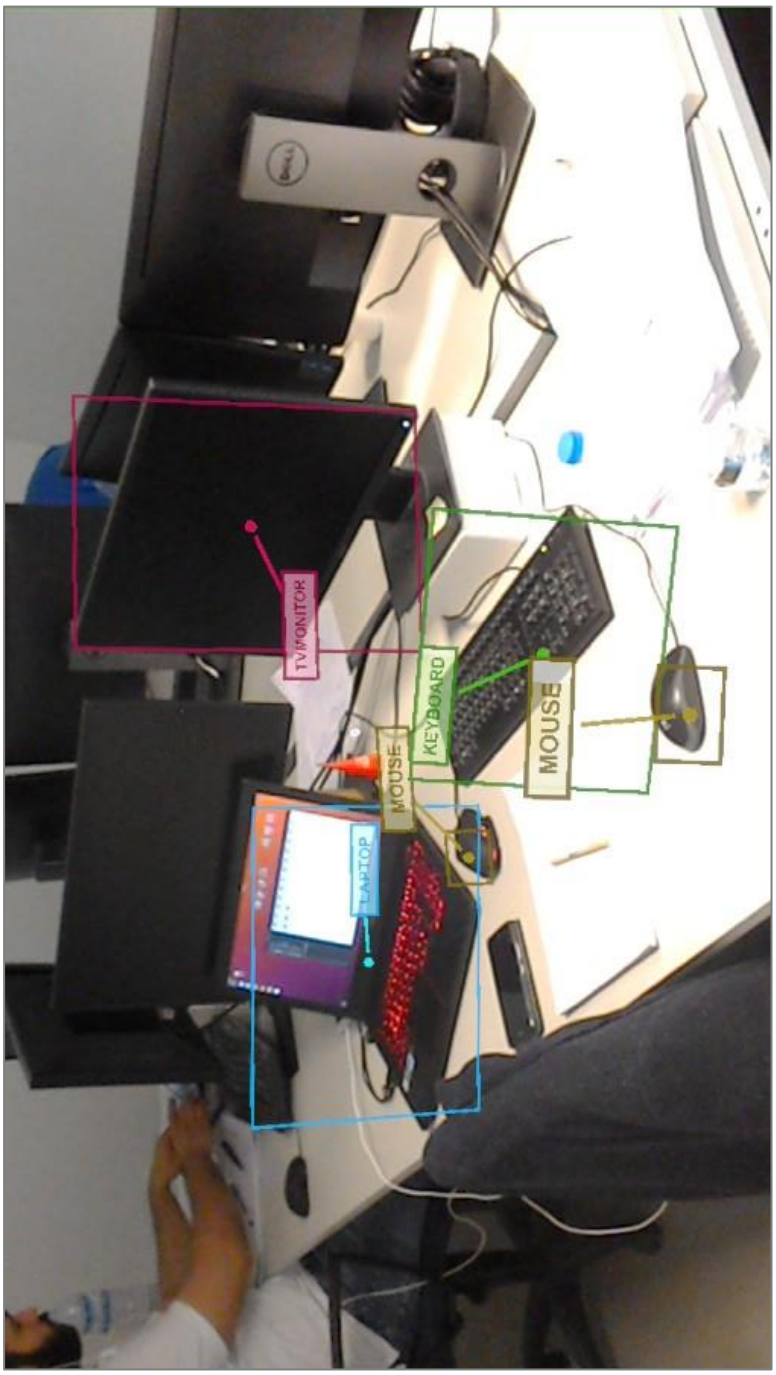

Figura 60 - Exemplo de deteção, localização e classificação de múltiplos objetos selecionados no ambiente que rodeia o utilizador.

T

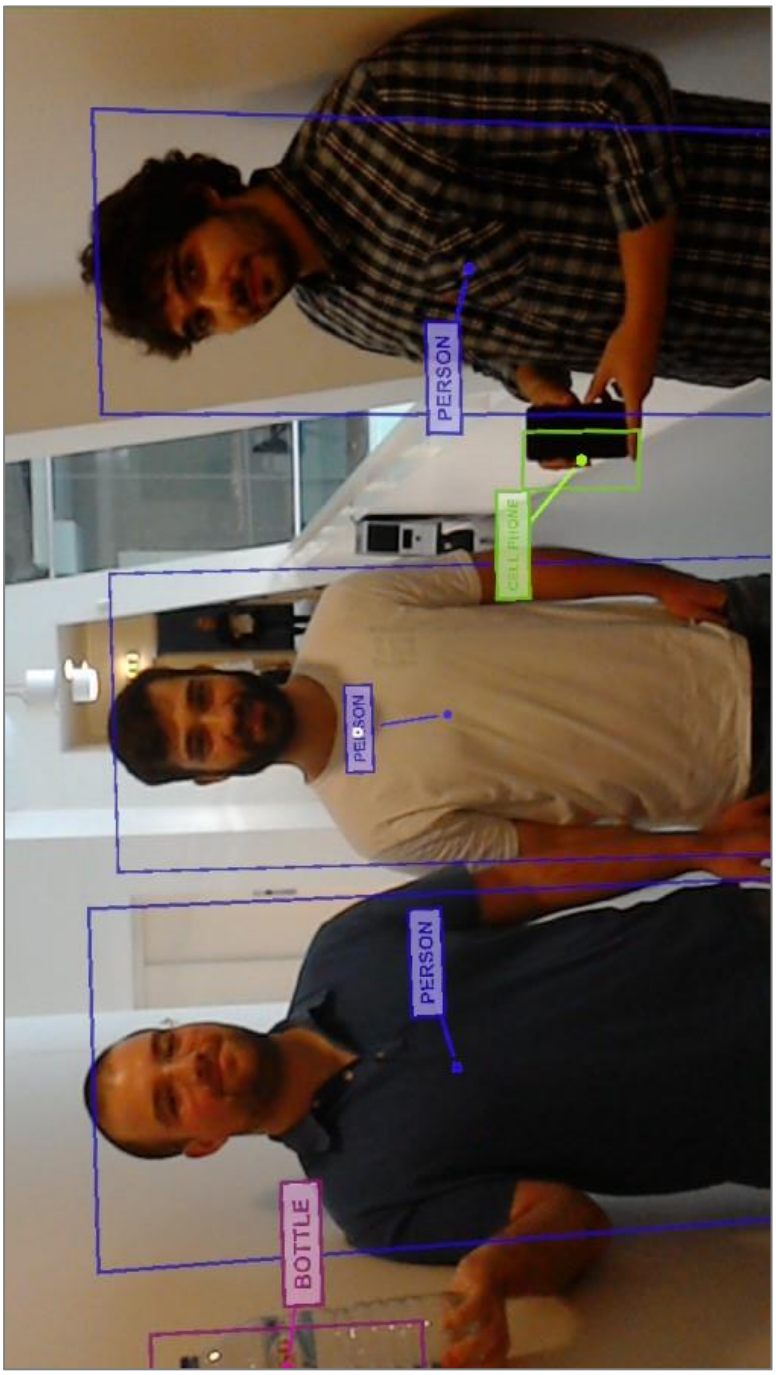

Figura 61 - Exemplo de deteção e localização de múltiplos objetos e pessoas que rodeiam o utilizador. Da esquerda para a direita, Gonçalo, João e Carlos, colegas de trabalho.

U