FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Reconhecimento de Elementos Gestuais com Kinect

Miguel Medeiros Correia

PREPARAÇÃO DA DISSERTAÇÃO

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Orientador: Eurico Manuel Elias de Morais Carrapatoso (PhD)

Co-orientador: António Abel Vieira de Castro (PhD)

11 de Fevereiro de 2013

Conteúdo

Lista de Figuras

Abreviaturas e Símbolos

ASL	American Sign Language
DMF	Deformable Model Fitting
DPM	Deformable Part-based Model
fps	Frames per Second
GPU	Graphics Processing Unit
HOG	Histogram of Oriented Gradients
HSV	Hue, Saturation and Value
KDE	Kernel Distribution Estimation
LGP	Língua Gestual Portuguesa
LP	Língua Portuguesa
MoG	Mixture of Gaussians
PSO	Particle Swarm Optimization
RGB-D	Red, Green, Blue and Depth data
SDK	Software Development Kit
SLR	Sign Language Recognition
SOV	Sujeito-Objeto-Verbo

Capítulo 1

Introdução

A linguagem é uma capacidade mental complexa que se desenvolve em criança de forma inconsciente e informal. Surge sem uma perceção da sua lógica subjacente. Por isto Pinker [1], define-a como um instinto, diz-nos para olhar para a linguagem não como uma prova da singularidade humana mas como uma adaptação biológica para comunicar informação. Por sua vez, a língua é a materialização dessa capacidade usando um determinado conjunto de regras, um código, seja este realizado através da fala, gesto, imagem ou escrita.

A língua usada para comunicação depende do grupo de indivíduos que a usam. Podemos categorizar o tipo de comunicação em dois grupos, o oral e o não oral. No primeiro insere-se a língua falada, como a Língua Portuguesa, enquanto no segundo temos a escrita, o gesto e a imagem. Podemos, ainda, dividir em termos de *emissor-recetor*, categorizando conforme o método de emissão e o de receção. Para aqueles que ouvem, os ouvintes, a língua estabelece-se em termos *orais-auditivos* enquanto que para os não ouvintes, ou surdos, geralmente estabelece-se em termos *gestuais-visuais*, onde gestual define-se como o conjunto de elementos linguísticos manuais, corporais e faciais necessários para a articulação de um sinal.

Foi Darwin em 1871 na sua obra *A Descendência do Homem* [2] o primeiro a articular a linguagem como uma espécie de instinto, dizendo primeiro que a língua é uma arte visto que tem que ser aprendida. No entanto, uma criança apresenta uma tendência instintiva para comunicar, como pode ser observado pelo seu balbucio. A língua é aprendida pelo contacto, por ouvir aqueles que nos rodeiam e por imitação. Esta forma de aprendizagem é muito difícil de um surdo usar uma vez que não tem acesso ao *feedback* auditivo. Torna-se então, para este, extremamente difícil produzir o som da palavra. É por isso que usam a Língua Gestual, uma modalidade de comunicação baseada no *gestual-visual*. É esta a sua língua materna, é esta a língua que usam diariamente para comunicar e na qual os seus pensamentos são formulados. A Língua Gestual que usam, em Portugal denominada Língua Gestual Portuguesa (LGP), tem uma estruturação, ou uma sintaxe, completamente diferente da empregue na Língua Portuguesa. Esta diferença sintática dificulta a compreensão total quando a comunicação é apresentada de uma forma escrita.

Na Língua Gestual, enquanto o emissor constrói uma oração a partir dos elementos manuais, corporais e faciais, o recetor usa o sistema percetual visual, ao invés do sistema percetual auditivo,

para entender o que é comunicado. Assim, a informação linguística é construída tendo em conta as capacidades percetuais do sistema visual humano. Desta forma, as relações espaciais na Língua Gestual são muito complexas. Não é natural, para aqueles que usam a Língua Gestual, a produção ou compreensão de língua escrita uma vez que a sua língua natural possui uma estrutura paralela, com a utilização de gestos complexos que envolvem simultaneamente diversas partes do corpo do sinalizador. Desta forma, a aprendizagem da Língua Portuguesa pelos surdos é um processo de aquisição de uma segunda língua, o que acaba por dificultar a compreensão de texto escrito na estrutura da Língua Portuguesa.

1.1 Caracterização do tema

O desenvolvimento de tecnologias de comunicação como o rádio e a televisão sempre teve como alvo a pessoa ouvinte. Cada programa é feito na língua do país de origem e traduzido para a língua do público alvo, seja por dobragem ou por legendagem. No entanto, as duas soluções não cobrem as necessidades da pessoa surda.

No caso da legendagem, a estrutura não é a da língua principal deste, dificultando a compreensão completa da informação. A solução mais viável praticada é a utilização de intérpretes, pessoas que conseguem traduzir o que é dito para a sua forma gestual. Esta solução nem sempre é possível, seja pela dificuldade da interpretação em si, seja pelo custo de produção, ou até pelo atraso intrínseco gerado pela forma de comunicação.

A Língua Gestual é uma língua complexa. Tem a sua própria sintaxe e é composta por diversos elementos organizados no espaço. A complexidade do gesto realizado com as mãos, a posição e trajetória que estas fazem no espaço tridimensional, a expressão facial durante a execução do movimento e até o posicionamento do corpo são, todos em conjunto, o que trazem significado a um gesto.

A natureza multi-modal do gesto na língua, assim como a deteção dos ligeiros e precisos movimentos das mãos, são dos principais desafios no reconhecimento da Língua Gestual. Estes problemas levaram a que a maioria das soluções tecnológicas fossem complexas e baseadas nas mais variadas tecnologias, criando uma confusão de sistemas que mal colaboram entre si ou são demasiado ineficientes.

Enquanto que o reconhecimento automático de fala avançou já ao ponto de estar comercialmente disponível, o reconhecimento de Língua Gestual é uma área investigação muito recente. Existe uma necessidade de desenvolver métodos e mecanismos capazes de traduzir, de forma viável e acessível, a Língua Gestual. O indivíduo surdo não dispõe facilmente de ferramentas que lhe permitem traduzir corretamente a sua língua. O desenvolvimento desta área tem o benefício de trazer a esta comunidade as capacidades de interpretação entre línguas que temos hoje em dia com ferramentas de reconhecimento automático de fala. Poderá também disponibilizar uma forma de interpretação da Língua Gestual mais acessível trazendo a interpretação a mais conteúdos multimédia, através da introdução da adequada tradução para a Língua Gestual da mesma forma que acontece com a legendagem.

Este trabalho foca-se em preencher a lacuna na falta de ferramentas eficientes e acessíveis de interpretação da Língua Gestual.

1.2 Objetivos

É necessário desenvolver formas viáveis de tradução para Língua Gestual, transmitir informação em tempo real e de uma forma de simples compreensão, intuitiva e natural.

Novos desenvolvimentos na área de visão por computador trazem-nos tecnologias que prometem resolver problemas desta categoria. Uma dessas tecnologias é o sensor *Kinect* da Microsoft, um dispositivo de interação baseado em sensores de movimentos criado inicialmente como interface de controlo da plataforma de jogos Xbox 360 e agora melhorado para ser utilizado com um computador Windows.

O Kinect concede-nos uma forma fácil e rápida de rastrear o movimento de um sujeito. O movimento do tronco, da cabeça, dos membros e agora, com a última versão, obtém-se, também um fácil rastreio facial. Um dos desafios é o rastreio de pequenos movimentos efetuados pelas mãos e a distinção entre estes.

O nosso principal objetivo é o de conceber formas de deteção dos elementos gestuais – movimento corporal, expressão facial e elementos manuais – que compõem a Língua Gestual Portuguesa, de uma forma simples e eficiente reduzindo a complexidade dos sistemas de aquisição utilizados no momento, tentando assim aproximar a tecnologia ao cumprimento da necessidade. Com esta capacidade tentaremos sintetizar os elementos gestuais utilizados na língua para os reproduzir automaticamente.

Iremos estudar a possibilidade de usar o sensor *Kinect* da Microsoft para detetar os elementos gestuais que compõem a Língua Gestual Portuguesa. Começaremos por analisar como recolher as características relevantes para a identificação dos elementos gestuais a partir de imagens e segmentos de vídeo. Passaremos, então, a estudar a Língua em si a fim de melhor adaptar a deteção das características e criar mecanismos de os identificar. Finalmente, vamos desenvolver mecanismos de deteção, rastreio e análise dos elementos gestuais, utilizando o Kinect.

Em segundo plano, existe a possibilidade de usar a informação captada pelo Kinect para sintetizar automaticamente a língua. Pensamos ser possível criar um sistema de tradução bidirecional de Língua Gestual Portuguesa em tempo real.

1.3 Motivação

A área de sistemas multimédia e a possibilidade da sua aplicação nas mais variadas áreas, desde a interação homem-computador, à análise de imagem e áudio sempre interessou pessoalmente ao autor. O gosto pela inovação e desenvolvimento tecnológico, aliado a um desejo pessoal de criar novas soluções práticas e viáveis a problemas ou necessidades comuns são um dos grandes motivadores. Problemas complexos podem sempre ser divididos em pequenos problemas exequíveis.

Os avanços na tecnologia de reconhecimento de fala e a atual capacidade de sintetizar e analisar voz em tempo real levaram a considerar o caso da Língua Gestual e a falta de mecanismos similares. Esta é uma área de investigação ainda num estágio muito inicial, novos algoritmos e novos sistemas de aquisição têm vindo a surgir. A possibilidade de contribuir numa área de investigação, trazer a uma comunidade um novo mecanismo de comunicação e o gosto pessoal do autor pela abordagem a problemas complexos são fatores que contribuíram fortemente à proposta deste projeto.

1.4 Estrutura do documento

Este documento é composto de 3 capítulos, sendo este o primeiro. Neste capítulo introduzimos o tema e caracterizamos o problema em mãos explicitando o seu contexto. Apresentamos os objetivos principais do projeto e a motivação por detrás da sua elaboração.

No capítulo 2 apresenta-se o estado da arte. Neste começamos por estudar as principais abordagens utilizadas na área de deteção e rastreio do movimento do corpo humano. Passamos então a estudar a Língua Gestual no contexto da análise de características analisando a Língua Gestual Portuguesa e identificando os maiores problemas na sua análise e reconhecimento.

Finalmente, no capítulo 3 apresenta-se a proposta de trabalho, identificando a metodologia a tomar no desenvolvimento do projeto, o plano de trabalho, os principais desafios que contamos superar e, finalmente, os resultados esperados aquando da conclusão do projeto.

Capítulo 2

Estado da arte

Para podermos detetar e identificar os elementos gestuais usados numa oração em Língua Gestual Portuguesa de uma forma automática, usando visão por computador, precisamos de efetivamente extrair a informação necessária de um sinal de vídeo. Não é difícil imaginar a amplitude do problema. Todos já assistimos, numa Língua ou noutra, a um intérprete ou um surdo a comunicar com Língua Gestual. A amplitude e complexidade dos movimentos, seja no espaço ou na própria variedade dos elementos gestuais, apresentam-se como obstáculos para a realização de um sistema eficiente e rápido.

Nas últimas décadas muita investigação tem aparecido na área de visão por computador. O surgimento de aparelhos e sensores, capazes de captar imagens e vídeo com maior e melhor resolução, a um preço acessível, assim como o aumento do poder de processamento das máquinas vieram abrir as portas para o desenvolvimento de mais e melhores técnicas na área de análise do movimento humano.

Neste capítulo vamos analisar a tecnologia que tem vindo a aparecer, focando-nos principalmente na área de deteção e análise do movimento humano. Começamos, primeiro, por estudar o primeiro passo em qualquer sistema de deteção, após a aquisição da imagem: a deteção em si. Na secção 2.1 analisamos as principais abordagens utilizadas para este problema. Analisamos como é efetuado o processo de separar as características relevantes numa imagem, separando o plano de fundo do cenário, ou *background*, que contem informação irrelevante ao contexto do problema, do primeiro plano (*foreground*). É neste último plano que se encontra a pessoa que queremos seguir e analisar. Estudamos os maiores problemas encontrados neste processo e como os ultrapassar.

De seguida, na secção 2.2 focamo-nos com mais precisão na área de deteção e análise da Língua Gestual, começando por analisar questões linguísticas na Língua Gestual Portuguesa, passando então a estudar as tendências e investigações nesta área de franca evolução nas últimas décadas.

Finalmente, na secção 2.3 estudamos o sensor Kinect, o que é, as suas potencialidades e como opera, justificando assim a escolha desta tecnologia para a realização deste projeto.

2.1 Deteção e rastreio

Quando o objetivo é o reconhecimento de pessoas ou objetos numa imagem, na área de visão por computador, o primeiro passo é o de reconhecer se na imagem em questão está ou não presente o objeto pretendido, assim como onde este se encontra. A esta tarefa é dada o nome de *deteção* [3], mais propriamente, o termo *figure-ground segmentation* é usado na terminologia inglesa e descreve o processo pelo qual o sistema visual organiza um cenário em figuras de primeiro plano (*foreground*) e fundo (*background*).

Processos de deteção são geralmente aplicados como o primeiro estágio de muitos sistemas de captura e análise sendo, portanto, um passo crucial. Existem, basicamente, duas abordagens ao problema de deteção: *deteção baseada no pixel* e *deteção baseada no objeto*. Na primeira abordagem, cada pixel de uma nova imagem é comparada com um modelo do cenário analisando se este pertence ao fundo ou ao primeiro plano. O resultado desta análise para todos os pixels da nova imagem retorna a silhueta de todas as pessoas detetadas. A deteção baseada no objeto tem como princípio movimentar uma janela deslizante por toda a imagem sendo calculada a probabilidade da existência de uma pessoa para cada posição da janela. Neste tipo de abordagem o resultado, geralmente, surge como uma caixa que engloba as pessoas detetadas na imagem. Estas duas abordagens são descritas em mais detalhe nas secções 2.1.1 e 2.1.2, respetivamente.

Em aplicações como deteção de intrusos é necessário uma série de imagens consecutivas para que seja possível qualquer processamento. Nestes casos o *rastreio* de objetos é um requisito. Na literatura, a noção de rastreio toma definições diferentes. Segundo Moeslund [4] este é composto de dois processos: deteção e correspondência temporal, em que o último é definido como o processo de associar os objetos detetados na imagem atual com aqueles detetados nas imagens prévias, retornando assim trajetórias temporais no espaço.

2.1.1 Deteção baseada no pixel

Num vasto número de sistemas de análise de movimentos são usadas câmaras estacionárias para monitorizar atividade em cenários de exterior ou de interior. Como a câmara é estacionária a deteção pode ser alcançada por simplesmente comparar o plano de fundo com cada nova imagem. A esta técnica dá-se o nome de *subtração do plano de fundo*. Uma das suas maiores vantagens é o facto de retornar uma eficiente segmentação das regiões do primeiro plano e do fundo da imagem. Subtração de fundo é muito usada para processamento posterior, como rastreio, e foi-o desde os primeiros sistemas de análise de movimento humano, como o *Pfinder* [5].

A noção de comparar cada pixel de uma imagem a um modelo do cenário onde esta se encontra é simples de se compreender. No entanto, esta abordagem depende do facto de considerar que o cenário é fixo. Tal consideração pode ser adaptada facilmente ao caso de uma imagem no interior, onde podemos controlar o ambiente e as condições de luz. Porém, o caso muda de figura quando consideramos cenários de exterior, onde as árvores mexem-se com o vento e as sombras movimentam-se com a posição do sol. Por esta razão, o desenvolvimento desta área tem vindo a focar-se em formas de modelar os pixels do plano de fundo e como atualizar estes modelos

durante o processamento. Nesta secção vamos analisar os maiores desafios na modelação de fundo, passando a discutir alguns dos métodos mais usados para a implementar subtração de fundo.

2.1.1.1 Desafios na modelação do plano de fundo

Para uma boa segmentação da imagem, é necessário ter uma boa modelação do fundo do cenário. Para o efeito é preciso fazer com que o modelo tolere alterações, seja tornando-o invariante a estas ou adaptativo. Toyama *et al.* [6] identificam uma lista de dez desafios que um modelo de plano de fundo tem de superar: *moved objects*, *time of day*, *light switch*, *waving trees*, *camouflage*, *bootstrapping*, *foreground aperture*, *sleeping person*, *waking person* e *shadows*. Por outro lado, Elgammal *et al.* [7] usam a origem da alteração para a classificar:

Alterações de Iluminação

Alterações de iluminação no cenário podem ocorrer como:

- Alterações graduais em cenários de exterior, devido ao movimento do sol relativamente ao cenário;
- Alterações repentinas como o ligar e desligar de um interruptor de luz num cenário de interior;
- Sombras projetadas por objetos no plano de fundo ou pelo movimento de objetos no primeiro plano.

Alterações de Movimento

Alterações de movimento podem ser categorizadas como:

- Deslocamento global da imagem por ligeiros desvios da câmara. Apesar de assumirmos que a câmara é estacionária pequenos deslocamentos desta podem ocorrer por fatores externos como a força do vento;
- Movimento dos elementos do plano de fundo, como o movimento das árvores com o vento.

Alterações Estruturais

Estas são alterações introduzidas ao plano de fundo da imagem pelos objetos alvo. Elgammal [3] define este tipo de alteração como algo que ocorre tipicamente quando qualquer objeto relativamente permanente é introduzido no plano de fundo do cenário. Como por exemplo, se uma pessoa se mantém estacionária no cenário por algum tempo. Toyama *et al.* [6] dividem esta categoria em *moved objects*, *sleeping person* e *waking person*.

Uma das questões centrais na modelação do plano de fundo é a decisão de que características modelar. Podem-se usar características baseadas no pixel como a intensidade ou bordas (estas podem ser identificadas com zonas na imagem, que apresentam variação local de intensidade significativa) ou características baseadas na região como o bloco da imagem. A escolha das características a modelar irá influenciar a tolerância do modelo a alterações.

Outra questão reside na escolha do modelo estatístico representativo das observações do sistema para cada pixel ou região. A escolha deste modelo irá afetar o grau de precisão da deteção. Na secção seguinte referem-se os métodos estatísticos mais utilizados no contexto de modelação do plano de fundo.

2.1.1.2 Modelação estatística do plano de fundo

Ao nível do pixel, podemos pensar no problema de subtração do plano de fundo como a necessidade de classificar se a intensidade de um determinado pixel, x_t, observada no instante t, pertence ao plano de fundo ou ao primeiro plano da imagem. No entanto, como a intensidade de um pixel do primeiro plano pode tomar qualquer valor arbitrário, podemos assumir que a sua distribuição é uniforme. Assim, reduzimos um problema de classificação de duas classes a um de apenas uma classe. Esta classificação pode ser obtida através do historial de observações que está disponível desse pixel.

Modelação paramétrica

A maioria das técnicas de subtração do plano de fundo, usa como base o modelo de plano de fundo de gaussiana única [3], segundo o qual, considerando que a distribuição de ruído de um determinado pixel tem uma distribuição gaussiana nula $N\left(0,\sigma^2\right)$, tem-se que a intensidade desse pixel é uma variável aleatória com distribuição gaussiana $N\left(\mu,\sigma^2\right)$.

A estimação dos parâmetros deste modelo reduz-se a avaliar o valor médio e a variância das observações das intensidades dos pixels, ao longo do tempo. Assim, a utilização deste modelo na prática reduz-se a subtrair uma imagem de fundo B a cada nova imagem I_t e verificar se a diferença é superior a um determinado limiar. Neste caso, a imagem de fundo B é composta pelo valor médio das imagens do plano de fundo.

Este modelo pode ser adaptado a variações lentas no cenário pela atualização iterativa da imagem de fundo. Uma solução eficiente é conhecida por *esquecimento exponencial* [3]:

$$B_t = \alpha I_t + (1-\alpha) B_{t-1} \tag{2.1}$$

onde $t \geq 1$, B_t representa a imagem de plano de fundo calculada até à imagem t e α corresponde à velocidade do esquecimento da informação do plano de fundo. Esta equação funciona como um filtro passa-baixo com ganho α que separa de uma forma eficaz o plano de fundo dos objetos em movimento. É de notar que neste caso B_t passa a representar a tendência central da imagem de fundo ao longo do tempo [8]. Este modelo é usado em sistemas como o *Pfinder* [5].

Tipicamente, em cenários de exterior, a imagem de fundo não é completamente estática, podendo variar a intensidade de determinados pixels de imagem para imagem. Nestes casos a abordagem de um única gaussiana não retorna bons resultado. Friedman *et al.* [9] apresentam um modelo onde uma mistura de três distribuições gaussianas foram usadas para

modelar o valor dos pixels em aplicações de vigilância de tráfego, usando cada uma das distribuições para representar a estrada, veículos e sombras, respetivamente. Desde o trabalho de Friedman, melhoramentos ao modelo de *mistura de gaussianas* - em inglês *Mixture of Gaussians*, *MoG* - foram apresentados, como por exemplo, o trabalho de Stauffer e Grimson [10].

Modelação não paramétrica

Em cenários de exterior é habitual haver uma vasta gama de variações muito rápidas, como ondas do mar. Este tipo de variações fazem parte do plano de fundo e a modelação deste tipo de cenário requer uma representação mais flexível da distribuição de cada pixel [11].

Uma técnica geral para estimar a função densidade de probabilidade de uma variável é a técnica de Estimação da Densidade do Núcleo - *KDE (Kernel Density Estimation)*. Os estimadores de núcleo convergem assintoticamente para qualquer função de densidade com amostras suficientes. Utilizando esta técnica pode-se evitar a necessidade de guardar o conjunto de dados completo, aplicando pesos a subconjuntos de amostras. Elgammal *et al.* [11] introduzem uma abordagem para a modulação do plano de fundo, utilizando esta técnica. Seja $x_1, x_2, \ldots, x_N$ uma amostra de intensidades de um pixel, pode-se obter uma aproximação da função de densidade de probabilidade para a intensidade do pixel, a qualquer intensidade, como:

$$Pr(x_t) = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{d} K_{\sigma_j} \left(x_{t_j} - x_{i_j} \right) \tag{2.2}$$

A função 2.2 encontra-se generalizada de forma a usar características de cor. Nesta, x_t é uma característica de cor de dimensão d num instante t e K_{σ_j} representa a função de núcleo com largura de banda σ_j na dimensão espacial de cor j [3].

Como função de núcleo podem ser usadas várias funções com propriedades diferentes, embora na literatura seja habitual o uso da função gaussiana. Neste caso, a função gaussiana é apenas usada para atribuir pesos. Ao contrário da modelação paramétrica, esta técnica é mais geral e não assume que a função densidade tenha qualquer forma específica.

Ao usar esta estimação de probabilidade, um pixel x_t é considerado como parte do primeiro plano se $Pr(x_t) < th$, onde th é um limiar associado globalmente que pode se ajustado conforme necessário.

Um dos problemas da utilização de técnicas KDE é a escolha de uma boa largura de banda do núcleo (σ). Teoricamente, quando o número de amostras tende para infinito a influência da largura de banda decresce tornando-se desprezável, mas na prática é usado um número finito de amostras. Uma largura de banda muito baixa dará lugar a uma estimação irregular, enquanto que um fator muito elevado conduzirá a uma estimação demasiado suavizada. Como são esperadas diferentes variações na intensidade de pixel de um local para o outro da imagem, é usado uma largura de banda de núcleo diferente para cada pixel. Mittal e

Paragios apresentaram uma abordagem adaptativa para a estimação da largura de banda do núcleo [12].

2.1.1.3 Supressão de sombras

Um processo de subtração do plano de fundo irá sempre detetar sombras de objetos como se fizessem parte do objeto em si. Quando os objetos são estáticos a sua sombra pode ser modelada juntamente com o plano de fundo. No entanto, a deteção de sombras de objetos que se movimentam apresenta um problema: as sombras confundem-se com o objeto a ser detetado. Pense-se no caso de análise do movimento humano, a existência da sombra do indivíduo dificulta a deteção correta do movimento dos membros.

Pode-se evitar detetar sombras ou até suprimir a sua deteção com a compreensão de como surgem. As sombras são constituídas por duas partes. Na figura 2.1 pode-se observar a representação da sombra de um objeto que se movimenta. A parte mais escura, a *umbra*, não recebe luz da fonte luminosa, e a parte mais clara, a *penumbra*, recebe alguma luz da fonte [13]. Devido às condições de luz direta e indireta, típicas de cenários de interior e exterior, é comum encontrar sombras de penumbra. Este tipo de sombra pode ser caracterizada como tendo uma intensidade menor, preservando a crominância do plano de fundo [3].

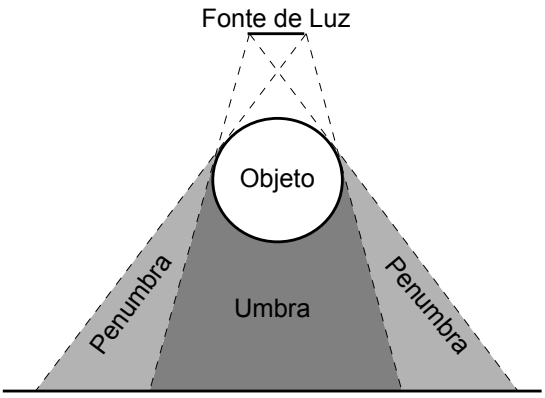

Figura 2.1: Geração de sombras sobre um objecto. Adaptado de [13].

Devido a esta propriedade de invariância à crominância, geralmente são utilizados espaços de cor também invariantes, ou menos sensíveis a alterações de intensidade de cor. Este é o caso do sistema HSV (*Hue, Saturation and Value*), onde as varíaveis H e S são invariantes a variações de intensidade de luz e a variável V, que representa a intensidade, varia. Elgammal *et al.* em [11] usam coordenadas de crominância baseadas no espaço RGB normalizado. Neste, dadas as três variáveis de R, G e B, as coordenadas de crominância são dadas por:

$$r = \frac{R}{R+G+B}, \qquad g = \frac{G}{R+G+B}, \qquad b = \frac{B}{R+G+B} \tag{2.3}$$

Uma vez que $r+g+b=1$ bastam duas variáveis para descrever o espaço de cor, temos então o espaço de crominância (r,g). Na figura 2.2 podemos comparar o resultado da deteção utilizando o espaço de cor (R,G,B) e (r,g), comprovando assim que a utilização deste espaço de cor retorna uma boa supressão da sombra da pessoa detetada.

Figura 2.2: **Esquerda** imagem original; **Centro** resultado da deteção usando o espaço (R,G,B); **Direita** resultado da deteção usando o espaço (r,g). Adotado de [3].

Embora o uso da crominância ajude a suprimir a deteção de sombras, tem o inconveniente de perder informação da intensidade da cor. Se considerarmos uma pessoa a caminhar com uma camisa branca com um plano de fundo cinzento, a pessoa não será detetada uma vez que o branco e o cinzento têm a mesma crominância. Por esta razão, é necessário utilizar sempre uma variável de intensidade da cor. No caso do espaço HSV esta é a variável V, enquanto que no espaço (r,g) é usada uma terceira variável de intensidade $s=R+G+B$, juntamente com r e g. Enquanto estas duas não variam sobre uma sombra, a variável s irá variar com a presença de sombras e zonas mais iluminadas.

A maioria das abordagens para supressão de sombras que se movimentam utilizam este raciocínio de separar as distorções provocadas pela crominância das distorções provocadas pela intensidade da cor.

2.1.2 Deteção baseada no objeto

Como foi discutido na secção 2.1.1 técnicas de subtração do plano de fundo são eficientes, no entanto, apenas em situações em que a câmara é estática. Em muitos cenários de análise de imagem a câmara é móvel: pense-se nos casos de uma câmara montada num robô. Nestes casos, a modelação do fundo não é viável, passando a ser necessário usar uma abordagem orientada ao objeto para eliminar a assunção de que o cenário terá um plano de fundo constante. Segundo Leibe [3] pode-se abordar o problema de extração de informação de um objeto usando diferentes níveis de detalhe. Do ponte de vista do rastreio este diz-nos que os objetivos principais de deteção são (a) detetar novos objetos, (b) classificar estes objetos num número de categorias de interesse, e (c) continuar a rastreá-los.

De seguida analisam-se os desafios desta abordagem juntamente com as técnicas mais usadas para atingir deteção com base nas características do objeto.

2.1.2.1 Conceito e desafios

A ideia de rastreio por deteção baseia-se em aplicar um detetor para a categoria de objetos pretendidos, a cada imagem de uma sequência de vídeo, e unir o resultado desta deteção para criar trajetórias.

Para atingir este fim é necessário, em primeiro lugar, detetar de forma fidedigna e eficiente a presença de novos objetos de interesse. Uma vez detetado um objeto e iniciado o seu rastreio é preciso que o sistema consiga distinguir se um objeto detetado numa nova imagem é aquele que se está a rastrear ou um novo, para poder associar essa deteção à trajetória do objeto ou iniciar um novo rastreio. A fim de conseguir esta deteção e associação é necessário construir um modelo de aparência, que por sua vez requer segmentação dos planos da imagem. Por fim, para limitar desvios na deteção, a segmentação tem que ser atualizada ao longo do tempo.

2.1.2.2 Abordagens à deteção de objetos

O detetor mais simples é o de *janela deslizante*, no qual uma janela de deteção de tamanho fixo é movimentado sobre toda a imagem, usando um classificador binário em cada localização da janela [3]. Para que se possam detetar objetos de diferentes tamanhos, a imagem ou a janela de deteção são redimensionadas e o processo é repetido. Usando esta técnica, a deteção de objetos reduz-se a uma simples decisão de classificação binária.

Com o uso de métodos de aprendizagem pode-se reduzir o número de decisões do classificador, o que melhora o seu tempo de execução e reduz o número de falsos positivos.

Uma das abordagens simples à deteção de objetos é a de representar as características de cada janela por um único vetor que codifica o conteúdo da janela em questão. O desafio neste caso passa pela escolha de uma representação suficientemente descritiva para capturar as características da classe do objeto com todas as suas variações e distinguir este do plano de fundo. Em 2005 foi apresentado em [14] a representação baseada em histogramas de gradientes orientados (*Histograms of Oriented Gradientes - HOG*). Esta representação divide a janela numa grelha de células, 4×4 ou 8×8, calculando para cada célula um histograma de orientação de gradientes. Em seguida, blocos de células de 2×2 são combinados para normalização do contraste. Para reduzir o efeito do ruído e do processo de quantização, a contribuição de cada pixel é pesada pela magnitude do seu gradiente. Finalmente, todos os blocos na janela de deteção são concatenados num único vetor normalizado. A representação HOG tem várias vantagens: (*a*) o uso de gradientes, ao invés da intensidade do pixel, torna o detetor tolerável a variações de iluminação como sombras; (*b*) a representação por histogramas torna o sistema mais robusto a pequenas variações das regiões da imagem; (*c*) a divisão em grelha adiciona informação localizada e dá mais detalhe à descrição do que o uso de um só histograma; (*d*) a normalização de blocos compensa variações locais de contraste [3].

Como a representação holística, como HOG, não é capaz de modelar variações locais na estrutura do objeto, como por exemplo diferentes partes do corpo, é necessário um grande número de

exemplos de aprendizagem para que o sistema possa aprender a detetar as alterações de aparência do objeto como um todo.

Uma solução mais flexível é a modelação baseada em partes deformáveis (*Deformable Part-based Model - DPM*) apresentada por Felzenszwalb *et al.* [15]. DPM representa objetos usando uma mistura de modelos de partes deformáveis a múltiplas escalas. Esta abordagem baseia-se no conceito de estruturas pictóricas, que representam objetos como uma coleção das suas partes organizadas de forma deformável. Cada parte contem propriedades da aparência local do objeto, enquanto que a configuração é caracterizada por ligações entre determinadas partes [16]. Assim, o detetor consiste num filtro global, similar ao descritor HOG e num conjunto de filtros das partes, extraídas a resoluções mais altas. O modelo define o valor da hipótese de um objeto como a soma do resultado dos filtros individuais menos o custo da deformação. A aparência das partes do objeto assim como a sua localização são aprendidas automaticamente de dados de treino. Uma vez treinado, este modelo é capaz de detetar o contorno do corpo e os seus membros.

2.1.2.3 Segmentação dos planos da imagem

Como evidenciado anteriormente, um dos desafios do rastreio através de deteção é a necessidade de associar cada nova deteção a um trajetória, descartando falsos positivos. Modelos de aparência, como modelos de cor, geralmente são usados para suportar a associação dos dados e escolher entre vários candidatos. No entanto estes modelos devem ser calculados apenas sobre a região do objeto, enquanto que o detetor de objetos retorna estes juntamente com o seu plano de fundo. Assim, para um rastreio eficiente, é necessário separar o plano de fundo do objeto detetado.

A abordagem mais simples, usada para rastreio de pedestres por Liebe *et al.* [17], passa por representar a forma do objeto por uma elipse de tamanho fixo dentro da caixa de deteção. Para detetar o mínimo possível do plano de fundo, a elipse usada por Liebe foca-se na parte superior da pessoa detetada, estendendo-se apenas ligeiramente para a zona das pernas, de forma a cobrir a maior parte possível da pessoa.

Como a elipse usada por Liebe não cobre os membros da pessoa detetada, um método de segmentação mais detalhado é preferível. Uma vez que temos a elipse, que claramente pertence à pessoa e não ao fundo, podem-se usar os pixels do seu interior para estimar a distribuição de cor do objeto, os pixels fora da elipse contêm uma estimativa da distribuição do plano de fundo. Estas duas distribuições são utilizadas como entradas de um sistema de segmentação do tipo *bottom-up* que tentará refinar o contorno do objeto. Neste tipo de abordagem, apesar da inicialização ser dada pela caixa do objeto, a segmentação não requer nenhum conhecimento à priori da forma do objeto, o que a torna aplicável a muitas categorias de objetos articulados diferentes. No entanto, a existência de sombras pode gerar uma estimação errada da distribuição do fundo, devolvendo segmentações incompletas.

Uma abordagem alternativa baseia-se em estimar a segmentação específica à classe do objeto, com base no resultado da deteção. Para que seja possível, é necessário que o detetor de objetos tenha sido treinado com exemplos de segmentação dos planos da figura, o que exige maior poder computacional do que uma simples caixa com o objeto.

2.1.2.4 Rastreio

Numa abordagem de rastreio através de deteção pura é necessária informação do detetor de objetos em cada imagem da sequência de vídeo para que se consiga seguir a trajetória do objeto. Mas, uma vez detetado um objeto a sua aparência não se irá alterar muito rapidamente. Tendo como base esta assunção podem-se usar técnicas de rastreio baseado em regiões para manter o rastreio por pequenos períodos. Esta técnica é útil em casos em que objeto pode ficar tapado, ou ocluído, permitindo o seu rastreio, assim como permite diminuir a necessidade computacional do sistema, reduzindo a quantidade de vezes que o detetor de objetos é ativado. Wu e Nevatia [18] propuseram uma abordagem a rastreio através de deteção, baseado na aparência, que usa rastreio com deslocamento da média para unir pequenas lacunas quando não existe deteção.

2.2 Língua Gestual

Muitas abordagens ao reconhecimento de Língua Gestual – na literatura inglesa é usado o termo *Sign Language Recognition (SLR)* – cometem o erro de tratar o problema como puramente de reconhecimento de gestos. A Língua Gestual é tão complexa quanto qualquer outra língua, sendo ainda que, na Língua Gestual, o significado é transmitido através de múltiplos canais em simultâneo. Cooper *et al.* [3] fazem uma simplificação do problema identificando três partes fundamentais da Língua Gestual:

1. *Características Manuais*, que englobam gestos realizados com as mãos, usando a forma da mão e movimento para transmitir um significado;

2. *Características não Manuais*, tais como expressões faciais ou a postura do corpo, que podem formar parte de um sinal ou modificar o seu significado;

3. *Ortografia Gestual*, descrevendo uma palavra, ao usar as suas letras constituintes de forma gestual, no alfabeto local. Na figura 2.3 está presente o alfabeto gestual usado na Língua Gestual Portuguesa.

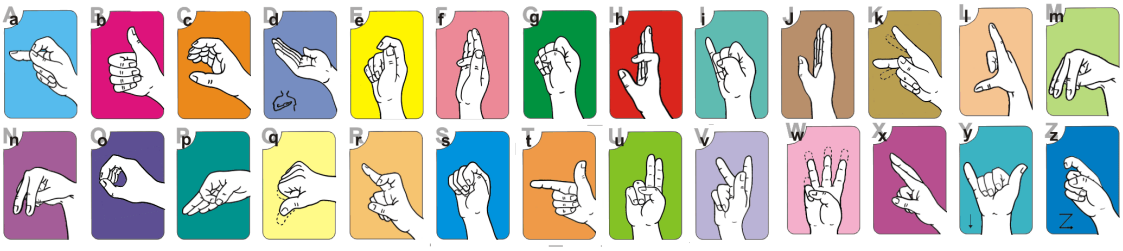

Figura 2.3: Alfabeto gestual usado na Língua Gestual Portuguesa. Adaptado de [19].

De agora em diante iremos referir-nos a *gesto*, *sinal* ou *símbolo* como um elemento gestual, composto de características manuais e não manuais, que tem um significado associado.

2.2.1 Língua Gestual Portuguesa

A Língua Gestual não é universal, sendo uma característica de cada país e cultura [20]. Assim sendo, em Portugal é usada a Língua Gestual Portuguesa. Tendo sido desenvolvida em paralelo com a Língua Portuguesa (LP), LGP não imita a sua contraparte e usa a sua própria sintaxe, tirando partido de características manuais e não manuais, num padrão simultâneo ou sequencial, arranjado no espaço tridimensional. Assim, a sua estrutura é também distinta da usada habitualmente em LP. A sua sintaxe é predominante organizada segundo *sujeito-objeto-verbo* (SOV). Por exemplo a oração "Eu vou para casa" fica, em LGP, como *(Eu) casa ir*. Outra característica, observável mo exemplo, é o fato de LGP não usar preposições (e.g.: "a", "para", "em", etc.). Ainda, no caso de o sujeito ser um pronome pessoal, e estiver implícito no contexto, poderá não ser necessário marcá-lo.

Os **verbos** em LGP são sempre realizados no infinitivo. Para marcar o tempo verbal são usados advérbios de tempo ou, na sua ausência, é usado o movimento do corpo, sendo que para a frente indica futuro e para trás indica passado.

A marcação do **género**, na LGP, surge apenas no caso de referência a seres animados, usando, normalmente, recurso aos gestos "homem" e "mulher", como marcas de masculino e feminino, respetivamente. No entanto, normalmente não é marcado o masculino, enquanto que o feminino é marcado por prefixação, isto é, o gesto *mulher* aparece antes do gesto que se pretende fletir em género. Existem ainda alguns casos em que o gesto no feminino e no masculino são diferentes (e.g.: *mãe/pai*).

A marcação de flexão em **número**, como o plural, pode ser efetuada de diferentes formas. Ana Bela Baltazar [20] descreve-as como:

Repetição, quando, para marcar o plural, o gesto é repetido;

Redobro, quando o gesto é realizado por ambas as mãos;

Incorporação, quando se usa um número para especificar quantidades reduzidas (e.g.: "quatro filhos" = *filho* + *quatro*);

Determinativo, usado para descrever quantidades não contáveis (e.g.: "muitos homens" = *homem* + *muito*).

Para realizar uma frase **interrogativa**, é usada a expressão facial que poderá ser combinada com o uso de pronomes interrogativos no final da frase. A frase **exclamativa**, por sua vez, é apoiada pela expressão facial e pela postura do tronco e da cabeça.

A diferença entre sinais é muito grande e cada individuo tem o seu próprio estilo, tal como na escrita. O individuo que pratica LGP terá uma *mão dominante*, como a mão direita para uma pessoa destra, e uma *mão não dominante*, pense-se na mão esquerda no caso anterior. O desempenho entre a mão dominante e a mão não dominante pode variar.

2.2.2 Aquisição e reconhecimento de dados

O primeiro passo num sistema de reconhecimento de Língua Gestual será sempre a aquisição dos dados. A maioria dos primeiros sistemas na área usavam luvas virtuais, como a DataGlove [21], e acelerómetros para recolher sinais específicos vindos das mãos. Nestes casos, as medidas como posição no plano (x,y,z), orientação, velocidade, etc, eram retiradas diretamente, sendo que muitas vezes os resultados dos sensores eram suficientemente bons para possibilitarem que fossem diretamente usados como características do sinal [3]. Embora este tipo de sistemas tenha a vantagem de devolver posições precisas, não permitia uma movimentação natural, restringindo a fluidez natural do movimento, alterando assim o sinal executado. Embora alguns sistemas tenham sido apresentados que reduziam este problema, os custos desta abordagem são geralmente muito elevados, levando ao uso da visão por computador.

Geralmente, no caso de visão por computador, uma sequência de vídeo é capturada usando uma combinação de câmaras. Em 1999 Segen e Kumar [22] usaram uma câmara e uma fonte de luz calibrada para calcular profundidade. Em 2004 Feris *et al.* [23] utilizam, uma série de fontes de luz externas para iluminar o cenário aplicando geometria de vários ângulos de visão para construir uma imagem de profundidade. Numa abordagem diferente, em 1998, Starner *et al.* [24] usam uma câmara frontal em conjunção com uma câmara montada na cabeça do indivíduo, apontada às mãos, para ajudar no reconhecimento de gestos. Imagens de profundidade podem ser conseguidas usando câmaras estereoscópicas, que têm a capacidade de simular a visão binocular humana usando duas ou mais lentes com sensores óticos separados. Este tipo de câmaras foi usada por Munoz-Salinas *et al.* em 2008 [25]. Recentemente o sensor Microsoft Kinect veio oferecer uma câmara de profundidade a um preço muito acessível, tornando as imagens de profundidade uma opção viável.

Uma vez adquiridos os dados, estes são descritos através das suas características. Na Língua Gestual, muitas dessas características baseiam-se nas mãos. Em particular, a forma e orientação da mão assim como a trajetória do seu movimento.

2.2.3 Características manuais

O rastreio das mãos não é uma tarefa fácil uma vez que, na Língua Gestual, os movimentos manuais são rápidos produzindo, muitas vezes, segmentos de vídeo desfocados. As mão são objetos deformáveis mudando de pose e posição no espaço. O movimento de uma mão pode ocultar o movimento da outra, assim como pode também ocultar a face do indivíduo [3].

Nos primeiros trabalhos, a tarefa de segmentação era simplificada com o uso de luvas coloridas. Zhang *et al.* [26] utilizou luvas coloridas e a geometria das mãos para detetar a sua posição e forma. As luvas usadas por Zhang *et al.* estavam codificadas de forma a que os dedos e as palmas das mãos tivessem cores diferentes. Este tipo de luvas diminui a restrição dos movimentos do indivíduo, provocada pelas luvas virtuais, mas não a elimina. Para uma abordagem mais natural, é usado um modelo da cor da pele como no trabalho de Athitsos e Sclaroff [27]. Imagawa *et al.* [28] demonstrou que usando a cor da pele obtinha-se uma boa segmentação, conseguindo segmentar

as mãos e face do indivíduo com esta técnica e aplicando, em seguida, um filtro de Kalman para o rastreio. Han *et al.* [29] demonstram que com o uso de filtros de Kalman conseguiam tornar esta abordagem robusta a oclusão. Restringindo o plano de fundo a uma cor específica, ou mantendo-o estático, consegue-se simplificar ainda mais esta tarefa. Zieren e Kraiss [30] usaram esta técnica para facilitar a segmentação do plano de fundo.

Imagens de profundidade podem ser usadas para simplificar o problema. Hong *et al.* [31] utilizam um par de câmaras estereoscópicas que, combinadas com outros sinais, permitiram construir modelos da pessoa na imagem. Por sua vez, Fujimura e Liu [32], usando a mesma tecnologia, conseguiram segmentar as mãos, embora com a assunção simplista de que as mãos seriam os objetos mais próximos da câmara.

O sensor Kinect veio oferecer aos investigadores desta área um bom meio para rastrear dados, permitindo desempenho em tempo real. Doliotis *et al.* [33] demonstram que usando este sensor, em vez do seu método anterior baseado na cor da pele, o desempenho do seu sistema aumenta entre 20% a 95%, num conjunto de dados de dez símbolos de números.

2.2.3.1 Forma da mão

Características da forma da mão são muitas vezes ignoradas, seja porque a resolução do vídeo não é suficientemente alta ou porque o poder de processamento é limitado não permitindo processamento em tempo real. Como alternativa, tende-se a aproximar a forma da mão através da extração de características geométricas como o seu centro de gravidade. O uso de luvas virtuais permite descrever a forma da mão em função dos ângulos das articulações e, de uma forma mais genérica, da abertura dos dedos, como foi demonstrado por Vogler e Metexas [34].

Com câmaras esteroscópicas, Rezaeis te al. [35] reconstroem um modelo tridimensional da mão, processando a correspondência de pontos e a estimação de movimento tridimensional, a fim de criar uma trajetória de movimento 3D completa assim como reconhecer a pose das mãos.

Oikonomidis *et al.* [36] usam o sensor Kinect para obter informação da forma da mão em tempo real, otimizando, de seguida, os parâmetros do modelo da mão usando uma variante de otimização por enxame de partículas (*Particle Swarm Optimization – PSO*) a fim de fazer corresponder a pose atual a um modelo. Embora este método consiga transmitir fielmente os parâmetros da mão, requer ainda um passo para extrair um elemento gestual conhecido.

2.2.3.2 Ortografia gestual

A ortografia gestual é uma extensão das características manuais da Língua Gestual, o seu reconhecimento requer uma boa descrição da forma da mão e, em certos Línguas, o seu movimento [3].

Com o uso de câmaras estereoscópicas para obter imagens de profundidade, Jennings [37] demonstrou um sistema de rastreio do movimento dos dedos robusto, usando contornos e cores. O sistema usa os contornos retirados de quatro câmaras, imagens esteroscópicas de duas câmaras e

cor de uma outra para detetar e rastrear os dedos. Os canais são combinados usando uma estrutura bayesiana.

Pugeault e Bowden [38] usaram o Kinect para criar um sistema de reconhecimento de ortografia gestual interativo, orientado à Língua Gestual Americana (*American Sign Language - ASL*). As mãos são segmentadas usando imagens de profundidade e de cor, sendo usados filtros de Gabor para extrair as características da pose e é usada uma técnica de aprendizagem, baseada em várias árvores de decisão, *florestas aleatórias*, para aprender a distinguir entre letras e formas. A ambiguidade entre certas formas é resolvida através de uma interface que permite ao utilizador escolher a letra correta.

2.2.4 Características não manuais

Juntamente com as características manuais, muita informação na Língua Gestual é transmitida através das características não manuais, tais como a expressão facial e pose da cabeça.

O reconhecimento da expressão facial pode ser interpretado diretamente para a Língua Gestual, ou para um sistema de interação humana mais genérico. Algumas expressões, segundo Ekman [39], são culturalmente independentes como o medo e a tristeza. A maioria da investigação na área de reconhecimento de expressões faciais, não relacionada com o reconhecimento de Língua Gestual, baseia-se nestas expressões, o que faz com que não se traduzam bem para a área em questão, sendo muitas vezes necessárias expressões exageradas para permitir o reconhecimento. Recentemente, investigadores têm trabalhado com conjuntos de dados não tão restritivos. Estas abordagens poderão provavelmente ser adaptadas à área do reconhecimento de Língua Gestual, uma vez que não têm tantas restrições e usam conjuntos de dados mais naturais [3].

Vogler e Goldstein abordam o problema de rastreio de características faciais no contexto de reconhecimento de Língua Gestual utilizando um modelo deformável da face [40]. Estes mostram que ao fazer corresponder pontos ao modelo e categorizando-os como estando dentro ou fora deste, é possível gerir oclusão pelas mãos. Eles propõem que não é necessário rastreio com oclusão completa, mas sim uma "recuperação graciosa". Este conceito sugere que quando a boca do indivíduo está escondida não é necessário saber a sua forma, podendo a informação ser retirada do que acontece antes e depois da oclusão, da mesma forma que um observador humano o faz. Porém esta teoria pode-se revelar muito difícil de comprovar.

2.3 Microsoft Kinect

Sendo originalmente apresentado como "Projeto Natal" a 1 de Junho de 2009, o sensor Kinect foi lançado a 4 de Novembro de 2010 como um acessório da consola Xbox 360 da Microsoft. Este é o fruto da parceria entre a empresa Israelita PrimeSense e a Microsoft [41].

O sensor Kinect foi criado para servir como uma forma de interação entre o utilizador e a consola Xbox 360, utilizando gestos e comandos de voz. Assim, o sensor é capaz de capturar imagens com 640×480 pixels a $30 fps$. Utilizando informação de profundidade, o sensor é ainda capaz de

produzir um modelo do esqueleto da pessoa que está a ser capturada. Com este modelo é possível definir gestos que serão reconhecidos pelo Kinect e usá-los para interagir com o computador.

Em Junho de 2011 a Microsoft lançou um *Software Development Kit* (SDK) para usar o sensor Kinect com o sistema operativo Windows 7, sendo que em Fevereiro de 2012 a versão para Windows do Kinect, *Kinect for Windows*, foi lançada.

2.3.1 Aplicações

O Kinect foi criado para revolucionar a forma como as pessoas interagem com jogos e a sua experiência, podendo interagir de uma forma natural, com o seu corpo [42]. É ainda capaz de receber comandos de voz e consegue identificar utilizadores quando estes se aproximam.

Antes do seu lançamento três demonstrações das capacidades do Kinect foram apresentadas. Estas foram *Ricochet*, *Paint Party* e *Milo and Kate* [41]. Em *Ricochet* um avatar imita todos os movimentos do utilizador e o objetivo deste jogo era acertar em bolas virtuais. *Paint Party* era uma aplicação de pintura, dando a possibilidade ao utilizador de escolher diferentes tipos de pincéis e utilizar gestos para colorir. *Milo and Kate* era a demonstração mais complexa. Criado pelos estúdios Lionhead o jogo funcionava como uma inteligência artificial emocional. O utilizador interagia de uma forma natural com um rapaz virtual de 10 anos, Milo, ou com um cão, Kate. A inteligência artificial do jogo respondia diretamente ao jogador através dos seus gestos, palavras ou ações predefinidas em situações dinâmicas. O sistema "aprendia" com o utilizador, adaptando-se às suas escolhas. Estas aplicações foram apenas usadas para demonstrar as potencialidades do sensor num ambiente de jogo.

Na altura do lançamento, quinze jogos foram apresentados que sairiam para o mercado juntamente com o Kinect, concebidos especialmente para usufruir das novas capacidades de interação oferecidas pelo sensor.

O sucesso do Kinect no mundo dos jogos despertou interesse de investigadores e praticantes de muitas e diferentes áreas como ciências de computadores, engenharia eletrotécnica e robótica. O baixo custo do sensor e as suas capacidades abriam portas para novas formas de interação com diferentes sistemas.

Na secção 2.2 referem-se alguns exemplos de sistemas de reconhecimento de Língua Gestual que usam o sensor Kinect para ultrapassar dificuldades no processamento de imagem e reconhecimento de objetos.

Muitos projetos foram desenvolvidos que usufruem deste sensor, nas mais variadas áreas, de seguida destacam-se alguns.

YScope [43]

A empresa portuguesa YDreams apresentou em 2012 o sistema YScope. Este é um sistema orientado a cirurgiões num ambiente de bloco operatório. O YScope usa o sensor Kinect para permitir que cirurgiões manipulem imagens médicas à distância, mantendo as suas mãos estéreis tanto quanto possível.

Brekel Kinect [44]

Brekel Kinect é uma aplicação que usa o sensor Kinect para permitir a captura de objetos tridimensionais e exportá-los para usar em ambientes 3D. Permite também rastreio do esqueleto para modelação e captura de movimento. É uma aplicação gratuita para uso comercial e privado.

A versão *Pro Body* é especializada na captura de movimentos de indivíduos (*motion capture* – *MoCap*). Consegue operar em tempo real sem a necessidade de pós-processamento, uma capacidade rara em sistemas MoCap. Suporta ainda a deteção de rotações das mãos, pés e cabeça detetando até 2 indivíduos num cenário.

O software tem ainda a versão *Pro Face* especializada em deteção tridimensional da face, sua posição e rotação.

3Gear Systems [45]

Uma vez que o Kinect é direcionado para trabalhar em capturas de corpo completo, a 3Gear Systems desenvolveu um SDK capaz de detetar gestos complexos produzidos pelas mãos do utilizador. A sua tecnologia traz ao Kinect a possibilidade de reconstruir uma representação precisa dos movimentos dos dedos do utilizador.

De momento o sistema usa dois sensores Kinect para evitar oclusão, sendo estes montados um pouco acima do monitor do computador. Esta montagem, juntamente com a tecnologia, permite construir interfaces interativas que respondem a pequenos gestos confortáveis, ao invés de largos gestos como o abanar dos braços.

SigmaNIL [46]

Similarmente ao 3Gear Systems, SigmaNIL é uma *framework* para visão por computador direcionada a interfaces naturais. É capaz de rastrear com precisão os movimentos dos dedos, a posição da mão, reconhecimento de gestos e rastreio do esqueleto da mão.

O sistema SigmaNIL usa apenas um sensor, suportando qualquer sensor de profundidade, como o Kinect. É capaz de interagir com as bibliotecas de base OpenNI e KinectSDK sendo, ainda, desenhado de forma modular para que se possa adicionar funcionalidades.

O interesse de muitos praticantes de várias áreas levou ao desenvolvimento de comunidades na Internet para discutir projetos que usam a tecnologia do sensor Kinect. Um dos melhores exemplos é a comunidade KinectHacks.net [47]. Apenas uma mês após o lançamento do Kinect esta comunidade já contava com 9 páginas com pequenas descrições de projetos [42]. Este número tem vindo a crescer, tendo, na altura da escrita deste documento, 70 páginas.

O interesse geral e o potencial dos projetos que têm vindo a surgir não foram descuidados pela Microsoft. Em 2012 surgiu o programa *Microsoft Accelerator for Kinect* [48] que é um projeto de apoio a empresas no seu início que usam o Kinect. Tendo recebido centenas de candidaturas de todo o mundo, o programa selecionou onze empresas que receberam apoio da Microsoft para desenvolver os seus produtos. As empresas focam-se em áreas muito variadas desde a interação homem-computador [49] à terapia física e cognitiva [50].

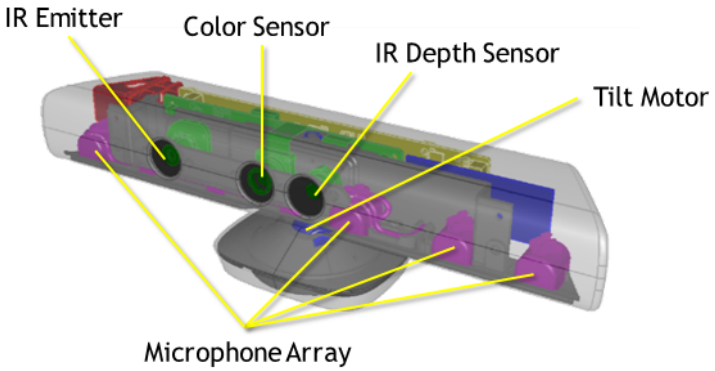

Figura 2.4: Componentes do sensor Kinect. Adotado de [51].

2.3.2 Sensor Kinect

O software interno do Kinect foi desenvolvido pela Rare, uma subsidiária da Microsoft Game Studios. Por sua vez, a tecnologia do sensor de profundidade, assim como o seu núcleo de processamento, foram desenvolvidos pela companhia PrimeSense [41]. O aparelho é constituído por um sensor de profundidade, uma câmara RGB, um acelerómetro, um motor e uma série de 4 microfones. Na figura 2.4 apresenta-se uma imagem da posição dos constituintes no aparelho.

2.3.2.1 Sensor de profundidade

O sensor de profundidade em si consiste num emissor de infravermelhos e uma câmara de infravermelhos, capaz de detetá-los. O emissor de infravermelhos cria um padrão estruturado de luz infravermelha e a câmara lê a reflexão desses raios. Por sua vez, a câmara interpreta a deformação da projeção e converte essa informação em valores de profundidade, medindo a distância entre o objeto e o sensor. Esta medição baseia-se em triangulação tendo em conta o emissor, a câmara e as posições dos pixels no cenário[41]. Existe uma distância de $7.5cm$ entre o emissor e a câmara de infravermelhos, razão pela qual é necessário calibrá-los para que a medição dos valores de profundidade seja correta.

A câmara de infravermelhos opera a $30fps$, criando imagens de 1200×960 pixels que são decimadas para 640×480 pixels com 11bits, o que resulta numa sensibilidade de $2^{11} = 2048$ níveis. O valor da profundidade é codificado numa escala de cinzentos.Quanto mais escuro for o pixel, mais próximo do sensor está esse ponto no espaço. Pixels pretos indicam que não existe informação de profundidade para esse ponto. Isto ocorre no caso dos pontos estarem muito longe, impossibilitando uma boa medição da sua profundidade, no caso de estarem numa sombra onde não haja pontos do emissor de infravermelhos, no caso de o objeto refletir mal a luz infravermelha (como no caso de espelhos ou cabelo) ou, finalmente, no caso de os pontos estarem muito próximos do sensor, uma vez que o campo de visão do Kinect é limitado devido ao emissor de infravermelhos e à câmara [42].

Uma das alterações introduzidas no sensor na versão para Windows foi a criação do *near mode*, que permite que o Kinect reconheça pessoas e objetos de uma forma mais precisa entre os $40cm$ e os $4m$ [52], sendo que o seu ângulo de visão é de 57^{o} na horizontal e 43^{o} na vertical [51].

2.3.2.2 Câmara RGB, motor, acelerómetro e microfones

A câmara RGB consegue captar imagens de 640×480 pixels, com 8 bits por canal, a $30fps$. O Kinect pode ainda operar num modo de alta resolução, captando imagens de 1280×1024 pixels a $10fps$ [41].

O motor e o acelerómetro trabalham em conjunto. O motor proporciona um mecanismo para inclinar o aparelho com uma amplitude de $\pm 27^{o}$. Por sua vez, o acelerómetro, configurado a uma amplitude de $2G$, onde G representa a aceleração provocada pela força da gravidade, é usado para determinar a orientação do Kinect [51].

O sistema de microfones é composto por uma série de quatro destes aparelhos. Com os microfones o Kinect é capaz de gravar áudio, determinar a localização da fonte sonora e a direção da onda de áudio [51].

2.3.3 Imagens de profundidade – RGB-D

Pixels numa imagem de profundidade indicam medidas calibradas de profundidade e não uma medida da intensidade da cor. As câmaras com capacidade de analisar a profundidade do cenário apresentam vantagens sobre as câmaras normais. Entre estas vantagens destacam-se a capacidade de operar a baixos níveis de luz, o facto de serem invariáveis à cor e textura assim como o facto de serem capazes de resolver ambiguidades na silhueta de um indivíduo [53].

A forma mais natural de captura de dados através do Kinect é através de imagens RGB-D. Este tipo de imagens resulta da combinação dos canais de cor vermelho (R), verde (G) e azul (B) com informação de profundidade (D).

A natureza distinta da origem da informação deste tipo de imagens – sendo que as cores têm natureza visual e a profundidade uma natureza geométrica – permite usar esta informação para realizar facilmente tarefas que seriam complexas apenas com imagens RGB, como segmentação de objetos em tempo real e reconhecimento de pose [41].

As imagens de profundidade trouxeram melhoramentos nas áreas de processamento de imagem e visão por computador. Até ao surgimento de sensores a preço acessível como o Kinect, os investigadores estavam limitados no acesso a imagens RGB-D. Da necessidade de alguns investigadores surgiram conjuntos de dados, acessíveis na Internet, que contêm imagens de profundidade, entre os quais se destacam o NYU Depth Dataset [54], o RGB-D Object Dataset [55] e o Cornell-RGBD-Dataset [56].

O uso destes conjuntos de dados permite o uso de imagens de profundidade em investigação mesmo que não se tenha acesso a um sensor Kinect.

2.3.4 Rastreio do esqueleto

Uma das grandes contribuições do Kinect foi a sua capacidade inovadora de reconhecimento e rastreio do movimento humano em tempo real adaptado a diferentes pessoas, de diferentes tamanhos e formas, sem necessidade de calibração. Esta capacidade baseia-se nos avanços feitos por Shotton *et al.* [53] em rastreio do esqueleto [42].

Rastreio do esqueleto é um processo que representa o corpo humano por um número de articulações representativas de partes do corpo, como a cabeça, o pescoço, os ombros e os braços. Cada articulação é representada pela sua localização no espaço 3D.

Shotton *et al.* desenvolveram um método de prever as posições no espaço tridimensional das articulações do corpo humano, de uma forma rápida e precisa, a partir de uma única imagem de profundidade, sem usar informação temporal: foi desenvolvido um passo intermédio que trata a segmentação das partes do corpo humano como uma tarefa de classificação baseada no pixel. A avaliação em separado de cada pixel evita a necessidade de uma pesquisa combinatória sobre todas as articulações do corpo [53]. Com este processo, o sistema consegue identificar as diferentes articulações do corpo do indivíduo detetado, criando um esqueleto com as propostas das posições das suas posições no espaço tridimensional.

Na figura 2.5 apresenta-se uma visão geral deste procedimento. No primeiro passo realiza-se a classificação das partes do corpo numa abordagem baseada no pixel, atribuindo a cada pixel uma cor. Cada cor corresponde à probabilidade desse ponto pertencer a uma determinada articulação. É criada uma hipótese da posição de cada articulação no espaço tridimensional ao encontrar o centróide global de cada parte, através de um deslocamento de média. O último passo do processo é o mapeamento das hipóteses das articulações para as do esqueleto, considerando a continuidade temporal e o conhecimento de dados já aprendidos pelo sistema [42].

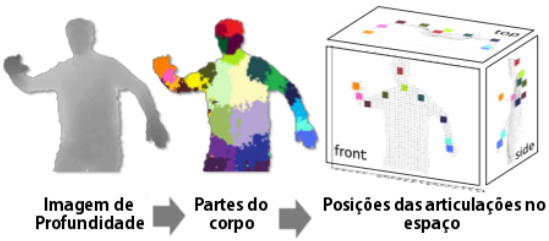

Figura 2.5: Processo de estimação da posição das articulações do corpo humano desenvolvido por Shotton *et al.* Adotado de [53].

Para treinar o sistema foram geradas imagens de profundidade realistas de humanos com diferentes formas e tamanhos em diferentes poses. Para melhorar a velocidade de processamento, o classificador pode ainda correr em paralelo na unidade de processamento gráfico (GPU). O algoritmo desenvolvido por Shotton *et al.* consegue correr a $200 fps$ na GPU da Xbox 360. Mais ainda, a abordagem discriminativa aprendida consegue superar casos de auto-oclusão e poses cortadas pela imagem.

2.3.5 Rastreio da posição da cabeça e da expressão facial

A deteção da expressão facial e da posição da cabeça têm sido uma área de investigação ativa na área de visão por computador. A sua realização tem impacto em diversas áreas como interação homem-computador e reconhecimento facial. A maioria das abordagens tende a focar-se em imagens bidimensionais. A falta de características faciais distintas neste tipo de imagem obriga a explorar técnicas baseadas na aparência e em modelos da forma.

Investigação mais recente foca-se em adaptar modelos deformáveis a digitalizações tridimensionais da face humana [42]. Para o efeito usam-se digitalizadores capazes de captar imagens 3D de alta qualidade usando sistemas de lasers de luz estruturada, produzindo bons resultados, embora a um custo elevado ou sendo necessário muito tempo para produzir uma digitalização.

O sensor Kinect tem a capacidade de juntar vídeo 2D e imagens de profundidade a $30fps$, a um custo baixo. Porém, a informação de profundidade do Kinect não é extremamente precisa, contendo muito ruído.

Cai *et al.* [57] desenvolveram um algoritmo de ajuste de um modelo deformável (*deformable model fitting – DMF*) através de máxima verosimilhança, capaz de lidar com a entrada ruidosa das imagens de profundidade, para ser usado com o Kinect. É usado uma modelo linear deformável da cabeça humana com combinação linear de uma face neutra, uma série de formas básicas com coeficientes estáticos ao longo do tempo, que representam uma pessoa em particular, e uma série de formas básicas com coeficientes, estes dinâmicos ao longo do tempo, que representam as expressões faciais [42].

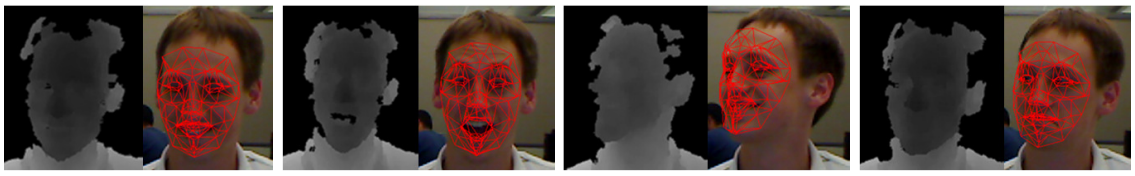

Figura 2.6: À esquerda a imagem de profundidade obtida pelo Kinect e à direita o resultado correspondente do algoritmo desenvolvido por Cai *et al.* Adotado de [57].

Uma vez que o Kinect retira a informação de profundidade através de triangulação, os valores obtidos para esta medida não têm todos a mesma precisão. O erro da medição de profundidade aumenta com o quadrado da distância. Assim, para formular a distância entre o modelo da face e o mapa de profundidade, cada ponto do mapa tem a sua própria matriz de covariância adequada para modelar a sua incerteza. Além disso, as características faciais obtidas através do vídeo bidimensional são rastreadas ao longo das tramas e facilmente integradas na estrutura DMF. Na figura 2.6 podemos observar o resultado do rastreio deste algoritmo.

2.4 Sumário

O estudo exposto neste capítulo visou abordar os temas relevantes para uma melhor compreensão de como funciona um sistema de deteção e análise, focando-se maioritariamente na deteção e análise de movimento humano.

Começámos por abordar o problema da deteção. Vimos que existem, na literatura, duas grandes abordagens a este problema: a baseada no pixel e a baseada no objeto. Estudámos como cada uma destas técnicas é realizada, analisando os seus maiores desafios e como estes são ultrapassados.

Passámos de seguida a centrar-nos mais na área de interesse deste projeto: o reconhecimento de Língua Gestual. Começámos por analisar a língua e os elementos que constituem um gesto. Olhámos de seguida para questões linguísticas da Língua Gestual Portuguesa a fim de melhor entender as diferenças entre esta e a Língua Portuguesa. Finalmente, analisámos abordagens a deteção e reconhecimento de cada uma das categorias constituintes de um gesto para aumentar o nosso entendimento das dificuldades na conceção de um sistema de reconhecimento de Língua Gestual.

Por fim, analisámos o sensor Kinect. Este sensor desenvolvido pela Microsoft trouxe ao mercado uma ferramenta a preço acessível capaz de captar imagens, vídeo e imagens de profundidade, tendo ainda a capacidade de efetuar deteção de movimento em tempo real. Começámos por estudar áreas em que este sensor é usado na atualidade, passando a analisar o sensor em si e os seus componentes. Finalmente, estudámos como o Kinect capta imagens de profundidade e como realiza rastreio e deteção.

Capítulo 3

Proposta de trabalho

Neste capítulo apresentam-se a metodologia e o plano de trabalhos para o desenvolvimento do projeto a ser realizado na unidade curricular Dissertação. Identificam-se os principais desafios que se espera superar durante o desenrolar do projeto assim como os resultados esperados no final da sua implementação.

3.1 Metodologia

O objetivo mais ambicioso deste trabalho é o de conceber um sistema capaz de interpretar Língua Gestual Portuguesa, utilizando o Kinect. Tendo a noção da complexidade do problema em mãos, pretende-se desenvolver o projeto de uma forma incremental. Desta forma dividiu-se o projeto em diferentes fases, definidas como:

1. Deteção e análise de elementos manuais;
2. Deteção e análise da pose da cabeça e expressão facial;
3. Rastreio da posição do corpo;
4. Combinação das fases 1, 2 e 3 para análise de gestos complexos;

Cada uma das fases 1, 2 e 3 deverá ser implementada de forma isolada e suportada por testes de validação. A fase 4 será abordada após verificadas todas as fases anteriores.

3.2 Principais desafios

Como em qualquer projeto, esperam-se algumas dificuldades aquando da sua implementação. Os principais desafios foram já brevemente identificados no capitulo 2, no entanto identificamo-los mais claramente nesta secção.

A deteção dos pequenos e precisos movimentos dos dedos não é uma capacidade nativa do sensor Kinect. Esta limitação terá que ser superada. Pretende-se desenvolver um mecanismo capaz

de estimar as articulações dos dedos das mãos, a fim de melhorar o rastreio dos movimentos, criando um esqueleto, numa abordagem similar à usada pelo Kinect para deteção das partes do corpo. Com esta abordagem espera-se obter uma boa estimativa do movimento dos dedos, resolver o problema de uma mão esconder a outra, assim como a oclusão de dedos de uma mão.

O SDK do Kinect permite, como se viu em 2.3.5, detetar a pose da cabeça e expressões faciais. No entanto, as expressões faciais usadas na Língua Gestual Portuguesa tomam variadas formas, algumas das quais testes com o sensor mostraram não serem identificadas nativamente. É exemplo disso a expressão "bochechas cheias". Para superar este problema presume-se conseguir refinar a malha de deteção da expressão facial, tornando-a mais sensível na zona das bochechas do indivíduo. Com este processo espera-se tornar o sistema mais sensível a expressões mais comedidas, evitando a necessidade de exagerar a expressão.

Unir as características manuais e não manuais num só classificador para identificar gestos complexos pode-se revelar uma tarefa complicada, seja pela necessidade de recursos com elevada capacidade de processamento ou pela falta de qualidade na deteção dos três elementos gestuais principais utilizando apenas um Kinect. Neste último caso, se o problema se apresentar, pretende-se adaptar a solução usando mais do que um sensor para melhor qualidade na deteção de cada característica. Espera-se diminuir a complexidade do sistema, reduzindo a necessidade de recursos, com o desenvolvimento de algoritmos refinados e eficientes.

3.3 Plano de trabalho

Para proceder à realização do projeto em mãos definiu-se um conjunto de etapas, com o objetivo de melhor planear as tarefas a desenvolver para um bom funcionamento do trabalho. Cumpridos todos os objetivos planeados, pretende-se ter um sistema capaz de reconhecer elementos gestuais que constituem a Língua Portuguesa.

Para que a realização deste planeamento fosse fidedigna foi necessário primeiro estudar o estado da arte, analisar soluções e identificar os principais problemas no âmbito do tema. Seguiu-se para a identificação de requisitos e divisão do sistema em várias partes, de forma a otimizar a abordagem ao problema.

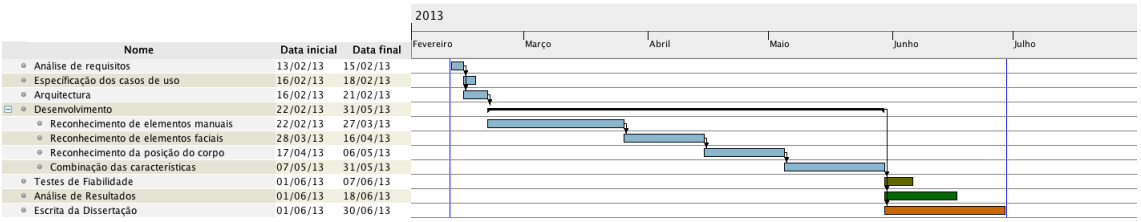

Figura 3.1: Diagrama de Gantt relativo à planificação do projeto.

Uma vez concluído este estudo, foi feito um planeamento a ser seguido aquando do desenvolvimento do projeto. Este planeamento, que pode ser consultado na figura 3.1, na forma de um diagrama de Gantt, que identifica as principais tarefas a executar assim como a sua duração.

Referências

[1] S. Pinker. *The Language Instinct: How the Mind Creates Language*. P. S. Series. HarperCollins, 2007.

[2] C.R. Darwin. *The descent of man, and selection in relation to sex*. The descent of man, and selection in relation to sex. 1871.

[3] Thomas B. Moeslund, Adrian Hilton, Volker Krüger, e Leonid Sigal, editores. *Visual Analysis of Humans - Looking at People*. Springer, 2011.

[4] T.B. Moeslund, A. Hilton, e V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst. (USA)*, 104(2-3):90 – 126, 2006/11/.

[5] Christopher R. Wren, Ali J. Azarbayejani, Trevor J. Darrell, e Alexander P. Pentland. Pfinder: real-time tracking of the human body. volume 2615, páginas 89 – 98, Philadelphia, PA, USA, 1996.

[6] K. Toyama, J. Krumm, B. Brumitt, e B. Meyers. Wallflower: principles and practice of background maintenance. volume vol.1, páginas 255 – 61, Los Alamitos, CA, USA, 1999//.

[7] Ahmed Elgammal, Ramani Duraiswami, David Harwood, e Larry S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151 – 1162, 2002.

[8] Xiang Gao, T.E. Boult, F. Coetzee, e V. Ramesh. Error analysis of background adaption. volume vol.1, páginas 503 – 10, Los Alamitos, CA, USA, 2000.

[9] N. Friedman e S. Russell. Image segmentation in video sequences: A probabilistic approach. Em *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, páginas 175–181. Morgan Kaufmann Publishers Inc., 1997.

[10] W.E.L. Grimson, C. Stauffer, R. Romano, e L. Lee. Using adaptive tracking to classify and monitor activities in a site. páginas 22 – 9, Los Alamitos, CA, USA, 1998//.

[11] A. Elgammal, D. Harwood, e L. Davis. Non-parametric model for background subtraction. *Computer Vision—ECCV 2000*, páginas 751–767, 2000.

[12] A. Mittal e N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. volume Vol.2, páginas 302 – 9, Los Alamitos, CA, USA, 2004//.

[13] J. Stander, R. Mech, e J. Ostermann. Detection of moving cast shadows for object segmentation. *IEEE Trans. Multimed. (USA)*, 1(1):65 – 76, 1999/03/.

[14] N. Dalal e B. Triggs. Histograms of oriented gradients for human detection. Em *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, páginas 886 –893 vol. 1, june 2005. doi:10.1109/CVPR.2005.177.

[15] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, e D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627 –1645, sept. 2010. doi:10.1109/TPAMI.2009.167.

[16] P.F. Felzenszwalb e D.P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vis. (Netherlands)*, 61(1):55 – 79, 2005/01/.

[17] A. Ess, B. Leibe, K. Schindler, e L. van Gool. A mobile vision system for robust multi-person tracking. Em *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, páginas 1 –8, june 2008. doi:10.1109/CVPR.2008.4587581.

[18] Bo Wu e Ram Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. volume 1, páginas 951 – 958, New York, NY, United states, 2006.

[19] Associação Portuguesa de Surdos. Alfabeto manual, 2011. Último acesso em 2012/02/06. URL: http://www.apsurdos.org.pt/.

[20] A.B. Baltazar. *Dicionário de língua gestual portuguesa*. Porto Ed., 2010.

[21] Mohammed Waleed Kadous e Computer Science Engineering. Machine recognition of auslan signs using powergloves: Towards large-lexicon recognition of sign language. Em *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, páginas 165–174, 1996.

[22] J. Segen e S. Kumar. Shadow gestures: 3D hand pose estimation using a single camera. volume Vol. 1, páginas 479 – 85, Los Alamitos, CA, USA, 1999.

[23] R. Feris, M. Turk, R. Raskar, K. Tan, e G. Ohashi. Exploiting depth discontinuities for vision-based fingerspelling recognition. Em *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, páginas 155–155. IEEE, 2004.

[24] Thad Starner, Joshua Weaver, e Alex Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371 – 1375, 1998.

[25] R. Munoz-Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, e A. Carmona-Poyato. Depth silhouettes for gesture recognition. *Pattern Recognit. Lett. (Netherlands)*, 29(3):319 – 29, 2008/02/01.

[26] L.G. Zhang, Y. Chen, G. Fang, X. Chen, e W. Gao. A vision-based sign language recognition system using tied-mixture density HMM. Em *International Conference on Multimodal Interfaces: Proceedings of the 6 th international conference on Multimodal interfaces*, volume 13, páginas 198–204, 2004.

[27] V. Athitsos e S. Sclaroff. Estimating 3D hand pose from a cluttered image. volume vol.2, páginas 432 – 9, Los Alamitos, CA, USA, 2003.

[28] K. Imagawa, Shan Lu, e S. Igi. Color-based hands tracking system for sign language recognition. páginas 462 – 7, Los Alamitos, CA, USA, 1998.

[29] J. Han, G. Awad, e A. Sutherland. Automatic skin segmentation and tracking in sign language recognition. *IET Comput. Vis. (UK)*, 3(1):24 – 35, 2009/03/.

[30] J. Zieren e K.F. Kraiss. Non-intrusive sign language recognition for human-computer interaction. Em *Proc. IFAC/IFIP/IFORS/IEA symposium on analysis, design and evaluation of human machine systems*, 2004.

[31] Seok-Ju Hong, Nurul Arif Setiawan, e Chil-Woo Lee. Real-time vision based gesture recognition for human-robot interaction. volume 4692 LNAI, páginas 493 – 500, Vietri sul Mare, Italy, 2007.

[32] Kikuo Fujimura e Xia Liu. Sign recognition using depth image streams. volume 2006, páginas 381 – 386, Southampton, United kingdom, 2006.

[33] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard, e V. Athitsos. Comparing gesture recognition accuracy using color and depth information. Em *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments*, página 20. ACM, 2011.

[34] C. Vogler e D. Metaxas. Handshapes and movements: Multiple-channel american sign language recognition. *Gesture-Based Communication in Human-Computer Interaction*, páginas 431–432, 2004.

[35] A. Rezaei, M. Vafadoost, S. Rezaei, e A. Daliri. 3D pose estimation via elliptical Fourier descriptors for deformable hand representations. páginas 1871 – 5, Piscataway, NJ, USA, 2008.

[36] I. Oikonomidis, N. Kyriazis, e A. Argyros. Efficient model-based 3D tracking of hand articulations using Kinect. *BMVC, Aug*, 2, 2011.

[37] C. Jennings. Robust finger tracking with multiple cameras. páginas 152 – 60, Los Alamitos, CA, USA, 1999.

[38] N. Pugeault e R. Bowden. Spelling it out: Real-time asl fingerspelling recognition. páginas 1114 – 19, Piscataway, NJ, USA, 2011.

[39] P. Ekman. Basic emotions, T. Dalgleish, MJ Power, Editors. *Handbook of cognition and emotion*, páginas 45–60, 1999.

[40] Christian Vogler e Siome Goldenstein. Facial movement analysis in ASL. *Universal Access in the Information Society*, 6(4):363 – 374, 2008.

[41] L. Cruz, D. Lucio, e L. Velho. Kinect and rgbd images: challenges and applications. páginas 36 – 49, Los Alamitos, CA, USA, 2012.

[42] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multimed. (USA)*, 19(2):4 – 10, 2012/02/.

[43] YDreams. YScope : YDreams and Hospital Santa Maria da Feira Launch Revolutionary App in the Health and Technology Sectors, Julho 2012. Último acesso em 2012/02/04. URL: http://www.ydreams.com/index.php#/en/aboutus/media/whatsup/2012/YSCOPEYDREAMSHOSPITALSANTAMARIA/.

[44] Jasper Brekelmans. Brekel Kinect, 2012. Último acesso em 2012/02/06. URL: http://www.brekel.com/.

[45] Robert Wang, Chris Twigg, e Kenrick Kin. 3Gear Systems, 2012. Último acesso em 2012/02/06. URL: http://threegear.com/.

[46] Sigma ArGe. SigmaNIL, 2012. Último acesso em 2012/02/06. URL: http://www.sigmanil.com/.

[47] KinectHacks.net. KinectHacks.net, 2011. Último acesso em 2012/02/06. URL: http://www.kinecthacks.com/.

[48] Microsft BizSpark. The Microsoft Acelebrator for Kinect, 2012. Último acesso em 2012/02/06. URL: http://www.microsoft.com/bizspark/kinectaccelerator/.

[49] inc. Ubi Interactive. rbi - Turns every surface into a 3D multitouch screen, 2012. Último acesso em 2012/02/06. URL: http://www.ubi-interactive.com//.

[50] Jintronix Inc. JINTRONIX, 2012. Último acesso em 2012/02/06. URL: http://www.jintronix.com/.

[51] Microsoft. MSDN - Kinect for Windows Sensor, 2012. Último acesso em 2012/02/05. URL: http://msdn.microsoft.com/en-us/library/hh855355.aspx.

[52] Kinect for Windows Team. Near mode: What it is (and isn't), Janeiro 2012. Último acesso em 2012/02/05. URL: http://blogs.msdn.com/b/kinectforwindows/archive/2012/01/20/near-mode-what-it-is-and-isn-t.aspx.

[53] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, e A. Blake. Real-time human pose recognition in parts from single depth images. páginas 1297 – 304, Piscataway, NJ, USA, 2011.

[54] Pushmeet Kohli Nathan Silberman, Derek Hoiem e Rob Fergus. NYU Depth Dataset V2, 2012. Último acesso em 2012/02/06. URL: http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html.

[55] Xiaofeng Ren Kevin Lai, Liefeng Bo e Dieter Fox. RGB-D Object Dataset, Novembro 2012. Último acesso em 2012/02/06. URL: http://www.cs.washington.edu/rgbd-dataset/.

[56] Ashutosh Saxena. Cornell RGBD Dataset, 2009. Último acesso em 2012/02/06. URL: http://pr.cs.cornell.edu/sceneunderstanding/data/data.php.

[57] Qin Cai, David Gallup, Cha Zhang, e Zhengyou Zhang. 3D deformable face tracking with a commodity depth camera. volume 6313 LNCS, páginas 229 – 242, Heraklion, Crete, Greece, 2010.