Detecção de face para obtenção de dados de posição

Gabriel F. Klimczak, Thiago Koman
Engenharia da Computação
Universidade Positivo, Curitiba, Brazil
{gabriel.klimczak99, thiagokomans}@gmail.com

I. INTRODUÇÃO

O uso de redes neurais convolucionais está sendo uma relevante alternativa para a solução dos mais diversos tipos de problemas, como o reconhecimento facial, reconhecimento de caracteres, dados de posição de pessoas e objetos, entre outros. O artigo em questão tem o objetivo de propor uma nova metodologia de detecção da posição de pessoas baseada no reconhecimento da face humana. Este sistema faz o uso de redes neurais convolucionais integradas com um sistema de detecção de movimento e algoritmos para o cálculo do seu posicionamento dentro de uma área de três dimensões.

O objetivo do sistema é recolher dados para o auxílio da disposição dos produtos dentro de uma área comercial, retornando para o usuário do sistema os pontos de maior movimentação, possibilitando dessa forma, uma melhor experiência para o cliente final.

Os sistemas atuais de detecção de movimento que possuem o objetivo de coletar dados de posição, necessitam que as câmeras utilizadas no processo de obtenção de imagens estejam posicionadas no teto do ambiente, sendo focadas para baixo, de modo que as pessoas que passam pela área sejam visualizadas de cima para baixo. Se a câmera estiver fora do posicionamento horizontal ideal, os dados coletados são imprecisos e incertos. Outro problema encontrado nessas soluções é que não possuem verificação de agente, ou seja, não é possível saber se o objeto que está sendo contabilizado é de fato um ser humano, eles apenas detectam e armazenam o movimento. O sistema de detecção vertical de dados de posição proposto tem os mesmos objetivos, mas não requer um posicionamento tão preciso das câmeras, garantindo que o objeto contabilizado é uma pessoa, pois se baseia na detecção facial.

O sistema de detecção vertical de dados de posição detecta a posição central da face de uma pessoa e retorna esse valor para uma função externa com as posições relativas aos eixos x e y. Os valores de x e y são obtidos através da posição no plano em duas dimensões e o valor z com base no tamanho da face detectada. Dessa forma, ambientes que possuem o pé-direito relativamente baixo podem chegar a utilizar apenas uma câmera em toda a área de monitoramento, ao invés de uma câmera para cada partição da área total.

Visando a comodidade do usuário, facilidade de manuseio e redução de custos, o sistema é embarcado em um microcontrolador com as tecnologias wi-fi, câmera e giroscópio integradas, possibilitando a obtenção de dados de posição da câmera (para cálculos relacionados a distância e ângulo do objeto detectado), imagens (para detecção da posição x e y do objeto) e transmissão remota de dados.

Além das vantagens já citadas nesse artigo, o sistema de detecção vertical de dados de posição entrega os dados processados em tempo real, e tem a opção de armazenamento das faces detectadas, podendo ser utilizado também em sistemas de vigilância.

II. CONCEITOS GERAIS

Os materiais e métodos serão utilizados para a criação de um sistema de reconhecimento facial para adquirir o posicionamento de um objeto - pessoa.

A. Visao Computacional

A visão computacional é uma ciência com o objetivo de tentar processar e analisar as imagens gravadas por câmeras de uma ampla variedade de maneiras para entender seu conteúdo ou para extrair informações geométricas. O termo visão computacional significa algo como visão baseada em computador.

As tarefas típicas sobre o uso da visão computacional são o reconhecimento e a medição da estrutura geométrica dos objetos, bem como os movimentos. É feito o uso de algoritmos de processamento de imagem, como segmentação, e métodos de reconhecimento de padrões, por exemplo, para classificar objetos. São utilizados métodos estatísticos ou probabilísticos, métodos de processamento de imagens, geometria projetiva, inteligência artificial e computação gráfica. As ferramentas vêm principalmente da matemática, em particular da geometria, álgebra linear, estatística, pesquisa operacional e análise funcional. Além disso, há uma estreita relação com campos relacionados, como fotogrametria, sensoriamento remoto e cartografia.

As áreas de aplicação são, por exemplo a navegação autônoma de robôs (sistemas de assistência ao motorista), a indústria cinematográfica para a criação de mundos virtuais (realidade virtual), a indústria de jogos para imersão e interação em espaços virtuais (realidade aumentada), a detecção e rastreamento de objetos (por exemplo pedestres) ou para registrar imagens médicas de tomografia computadorizada e detectar tecido doente, etc.

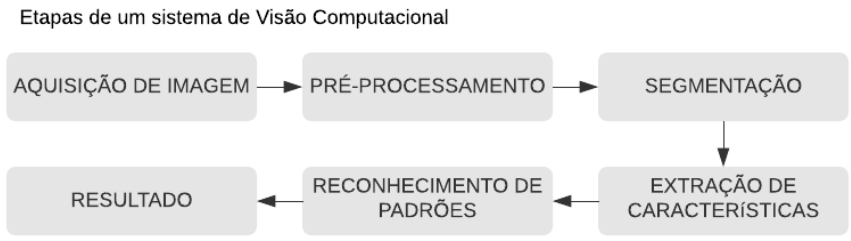

Figura 1. Diagrama das etapas da visão computacional. Fonte: o autor

B. Redes Neurais

As redes neurais artificiais baseiam-se principalmente na rede de muitos neurônios McCulloch-Pitts ou em pequenas modificações dos mesmos. A topologia de uma rede (atribuição de conexões a nós) deve ser bem pensada dependendo de sua tarefa. Após a construção de uma rede, segue-se a fase de formação, na qual a rede “aprende”. Em teoria, uma rede pode aprender com os seguintes métodos: Desenvolvimento de novas conexões; Excluir conexões existentes; Alterar o peso (os pesos wij do neurônio j para o neurônio i); Ajustando os valores de limiar dos neurônios, se estes tiverem valores de limiar; Adicionando ou excluindo neurônios; Modificação das funções de ativação, propagação ou saída.

Além disso, o comportamento de aprendizagem muda quando a função de ativação dos neurônios ou a taxa de aprendizagem da rede muda. Em termos práticos, uma rede “aprende” principalmente modificando os pesos dos neurônios. Uma adaptação do valor do limiar também pode ser feita por um neurônio. Como resultado, as RNAs são capazes de aprender funções não lineares complexas usando um algoritmo de "aprendizado" que tenta determinar todos os parâmetros da função a partir dos valores de entrada existentes e de saída desejados usando uma abordagem iterativa ou recursiva.

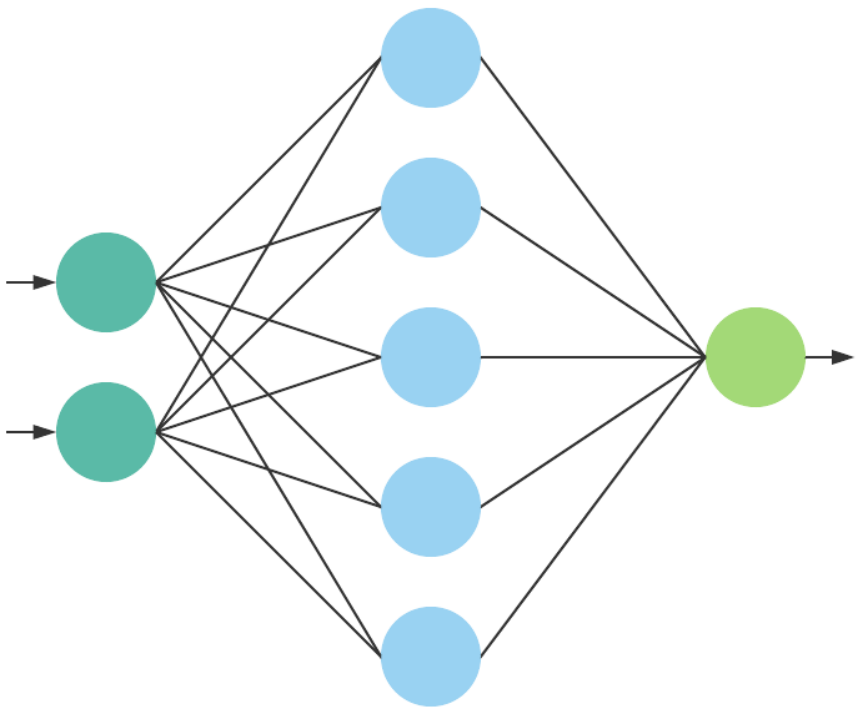

Figura 2. Uma rede neural artificial é um grupo interconectado de nós. Cada nó circular representa um neurônio artificial e cada linha representa uma conexão da saída de um neurônio artificial para a entrada de outro. Fonte: o autor

Rede neural convolucional

Uma rede neural convolucional (CNN ou ConvNet) é uma rede neural artificial. Basicamente, a estrutura de uma rede neural convolucional clássica consiste em uma ou mais camadas convolucionais, seguidas por uma camada de pooling. Em princípio, essa unidade pode ser repetida qualquer número de vezes; se for repetida o suficiente, fala-se de redes neurais convolucionais profundas, que se enquadram no campo do aprendizado profundo.

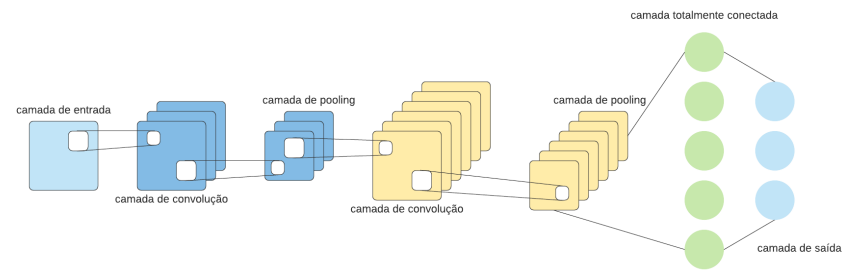

Figura 3. Arquitetura geral da CNN. Fonte: o autor

Camada Convolucional

A entrada é geralmente uma matriz bidimensional ou tridimensional (por exemplo, os pixels de uma imagem em tons de cinza ou em cores). Consequentemente, os neurônios estão dispostos na camada convolucional.

A atividade de cada neurônio é calculada usando uma convolução discreta. Uma matriz de convolução comparativamente pequena (kernel do filtro) é movida sobre a entrada passo a passo. A entrada de um neurônio na camada convolucional é calculada como o produto interno do kernel do filtro com a seção de imagem atualmente subjacente. Correspondentemente, neurônios vizinhos na camada convolucional reagem a áreas sobrepostas (frequências semelhantes em sinais de áudio ou arredores locais em imagens).

Deve-se enfatizar que um neurônio nesta camada reage apenas a estímulos em um ambiente local da camada anterior. Isso segue o modelo biológico do campo receptivo. Além disso, os pesos para todos os neurônios de uma camada convolucional são idênticos (pesos compartilhados). Como resultado, por exemplo, cada neurônio na primeira camada convolucional codifica a intensidade na qual uma borda está presente em uma área local específica da entrada. A detecção de bordas como a primeira etapa no reconhecimento de imagem tem um alto nível de plausibilidade biológica.

Análogo ao córtex visual, tanto o tamanho dos campos receptivos quanto a complexidade das características reconhecidas (por exemplo, partes de um rosto) aumentam nas camadas convolucionais inferiores.

Camada de pooling

Na próxima etapa, o agrupamento de informações supérfluas é descartado. Para reconhecimento de objeto em imagens, por exemplo, a posição exata de uma borda na imagem é de interesse insignificante - a localização aproximada de um recurso é suficiente. Apesar da redução de dados, o desempenho da rede geralmente não é reduzido pelo pool. Pelo contrário, oferece algumas vantagens significativas: Requisitos de espaço reduzidos e maior velocidade de cálculo; A possibilidade resultante de criar redes mais profundas que podem resolver tarefas mais complexas; Crescimento automático do tamanho dos campos receptivos em camadas convolucionais mais profundas.

C. Software e Hardware

O software de detecção e as redes neurais são estruturados em python devido a documentação e suporte para processamento de imagens que a linguagem oferece. Para a obtenção e processamento de imagens, a biblioteca openCV do python tem demonstrado notoriedade, pois é robusta e de fácil entendimento e programação, e por conta disso foi escolhida.

Para a obtenção de imagens, é considerado o uso de apenas uma câmera com resolução mínima de 780p x 420p, pois através da resolução mínima, é possível detectar faces a até 10 metros de distância.

O embarcado é equipado com câmera que possui a resolução mínima, podendo ser modificada conforme necessidade. Também possui um giroscópio capaz de detectar o ângulo e posicionamento da câmera, sendo possível dessa forma, realizar cálculos sem a necessidade de uma parametrização complexa.

Por ser um sistema embarcado, não existe a necessidade de um servidor externo para processamento, armazenamento ou disponibilização dos dados coletados, pois a raspberry tem a capacidade de processar, armazenar e divulgar os resultados em uma rota local, de modo que apenas quem tem acesso à rede de conexão do sistema embarcado consegue acesso aos dados.

Figura 4. Tamanho de uma raspberry pi 3 em comparação com a mão humana. Fonte: TechTudo.

D. Funcionamento do Software

Todo o processo de captura de imagens, processamento e disponibilização dos dados ocorre ao mesmo tempo, por conta disso é necessário definir restrições de qualidade de dados, pois seria impossível processar dados a 60 FPS (Frames por segundo) com resolução de 780p x 420p em uma raspberry pi 3. Por conta disso as imagens são processadas a uma média de 8 FPS, podendo oscilar conforme a disponibilidade do microcontrolador.

Após o processo de captura, as imagens são processadas através de algoritmos, utilizando a biblioteca OpenCV. Dados de

posição de face são extraídos das imagens através de filtros e rede neural, e são armazenados em um banco de dados local.

Os dados coletados e armazenados são disponibilizados em um servidor web local de modo que pode ser acessado através do ip do microcontrolador acompanhado da porta padrão 5065, sendo possível acessá-los através do seguinte formato: (http://<ip>:<porta_padrão>). Em casos onde a porta padrão já está sendo utilizada, a porta poderá ser alterada.

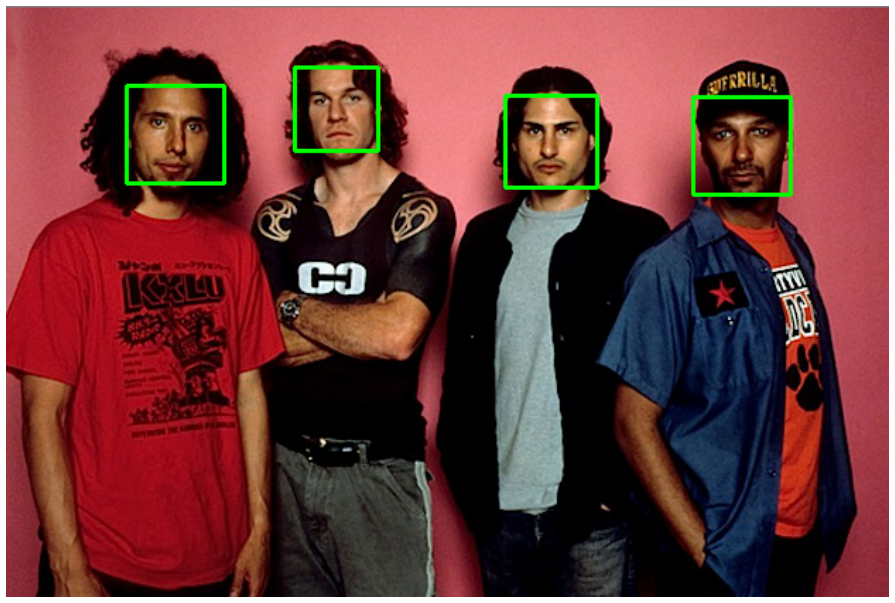

Figura 5. Imagens da banda Rage Against The Machine processada em um algoritmo de detecção de face. Fonte: Updated Code.

III. MATERIAL E MÉTODOS

Os materiais utilizados no projeto são compostos por uma raspberry pi 3, equipada com câmera de resolução 720x480. Com relação aos métodos, é utilizada uma rede neural, treinada para detectar a face de seres humanos, juntamente com outros recursos de visão computacional.

A detecção da face é um recurso primordial do sistema, pois é através da posição horizontal, posição vertical e tamanho da área da face que serão calculadas as posições do indivíduo detectado. Após a fase de experimentos, o sistema proporcionará um modelo para plotagem 2D das posições das faces.

Também será utilizado o método de conversão do modelo RGB, padrão da imagem original, para escala de cinza, possibilitando a utilização de uma rede neural já treinada, disponível na própria documentação do OpenCV.

IV. EXPERIMENTOS

Para obter um modelo de coleta de dados robusto foram realizados os seguintes testes.

A. Encontrar a proporção do tamanho da face, em pixels, pela distância da face, em metros.

Distância (m)	Tamanho da Face (px)
0,5	300
1	150
1,5	100
2	72
2,5	60
3	50

Figura 6. Tabela com os resultados obtidos no teste de Distância X Tamanho da Face. Fonte: Os autores.

B. Encontrar a função que representa a curva obtida pelos pontos encontrados no primeiro teste.

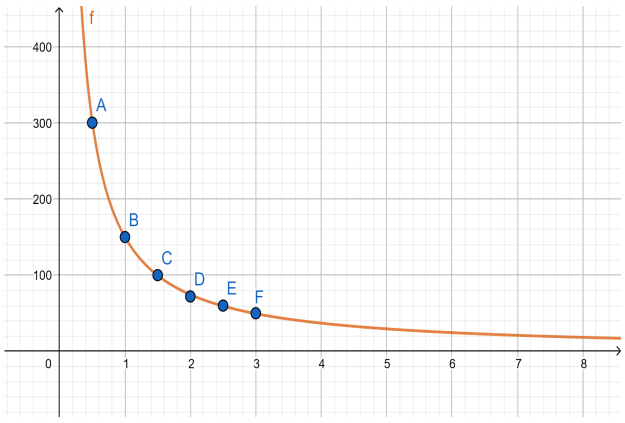

Figura 7. Função obtida através dos pontos do teste A. Fonte: Os autores.

C. Encontrar a proporção da tela com base na distância da face encontrada.

Distância (m)	Largura da tela (m)	Altura de tela (m)
0,5	0,41	0,32
0,75	0,65	0,46
1	0,9	0,63
1,5	1,42	0,88
2	1,9	1,2

Figura 8. Tabela com os resultados obtidos no teste de Distância X Tamanho da imagem. Fonte: Os autores.

D. Encontrar as funções que representam a reta obtida pelos pontos encontrados no teste anterior.

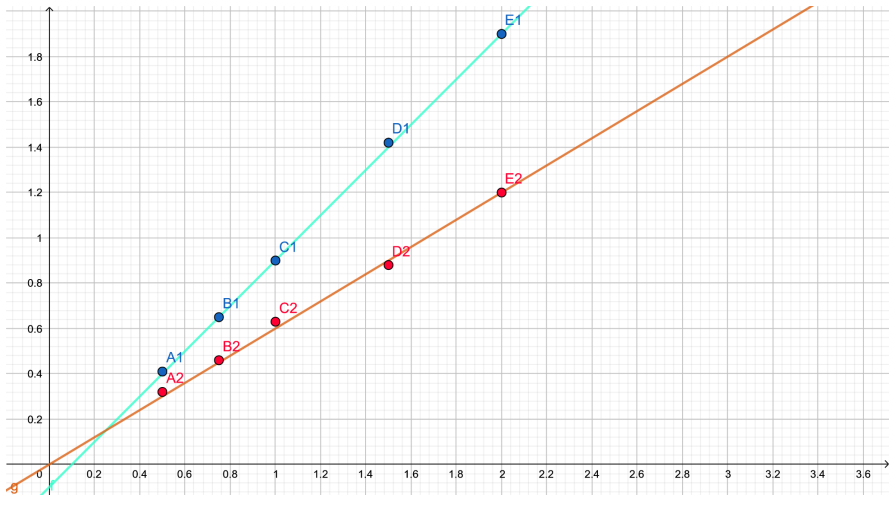

Figura 9. Função obtida através dos pontos do teste C. A reta em vermelho representa a abertura da câmera no eixo vertical e a linha em azul, representa a abertura da câmera na horizontal. Fonte: Os autores.

V. RESULTADOS

Os resultados obtidos são compatíveis com os objetivos traçados. Os valores coletados através do processamento de imagens possui uma margem de erro de 10% com relação a confiabilidade do posicionamento e 8% com relação a detecção da face humana. Os dados coletados são dispostos na tela, como resultado do processamento do vídeo, que está sendo capturado no momento do processamento. Através da imagem abaixo, é possível entender um pouco melhor o resultado obtido.

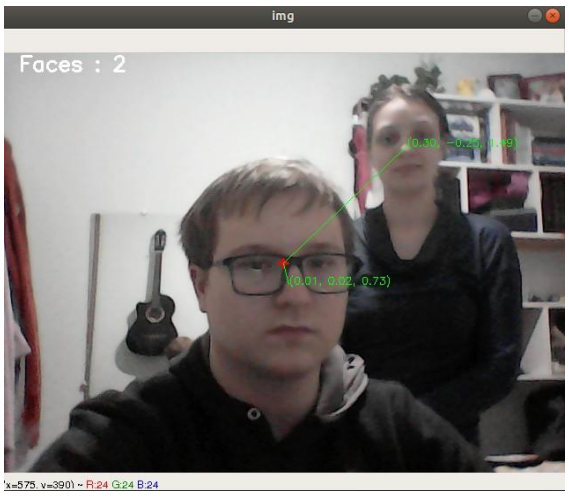

Figura 10. Imagem com os dados coletados pelo programa. Fonte: Os autores.