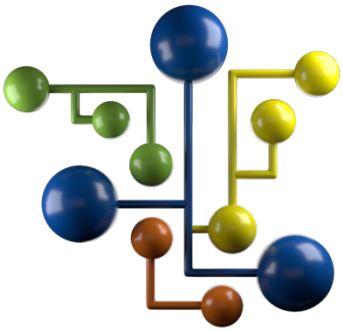


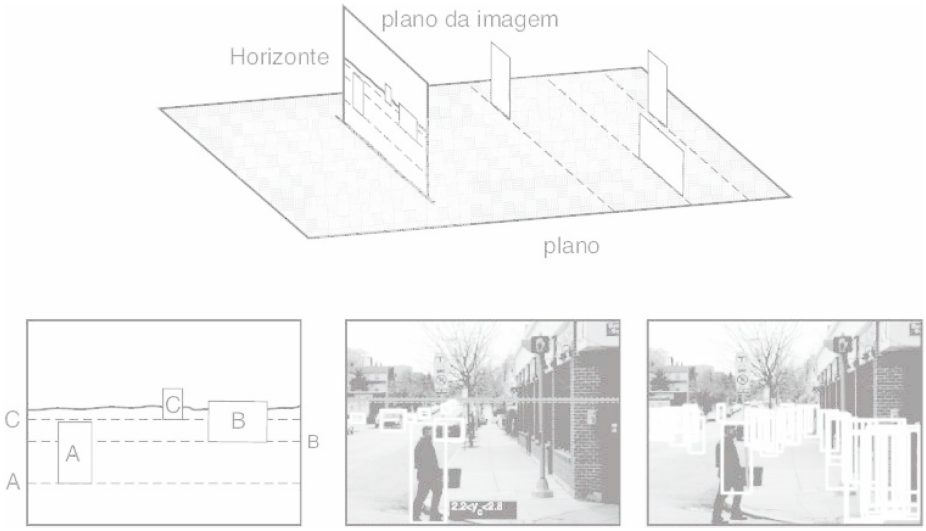
Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Objetos e Estrutura Geométrica de Cenas

A cabeça de um adulto humano típico tem cerca de 20 cm de comprimento. Isso significa que, para alguém a 13 m de distância, o ângulo subtendido para a cabeça na câmera é de 1 grau. Se virmos uma pessoa cuja cabeça parece subtender apenas meio grau, a inferência bayesiana sugere que estamos olhando para uma pessoa normal que está a 26 m de distância, em vez de alguém com a cabeça de metade do tamanho. Essa linha de raciocínio nos fornece um método para verificar os resultados de um detector de pedestres, bem como um método para estimar a distância de um objeto. Por exemplo, todos os pedestres são praticamente da mesma altura e tendem a ficar sobre o plano do chão. Se soubermos onde o horizonte está em uma imagem, podemos classificar os pedestres pela distância até a câmera. Isso funciona porque sabemos onde os pés estão, e os pedestres cujos pés estão mais próximos do horizonte na imagem estão mais afastados da câmera (figura abaixo). Os pedestres que estão mais distantes da câmera também devem ser menores na imagem. Isso significa que podemos descartar algumas respostas do detector — se um detector encontra um pedestre que é grande na imagem e cujos pés estão perto do horizonte, encontrou um pedestre enorme; isso não existe, de modo que o detector está errado. Na verdade, muitas ou a maioria das janelas de imagem não são janelas de pedestres aceitáveis e não precisam ser apresentadas ao detector.



Em uma imagem com pessoas em pé sobre um plano no chão, as pessoas cujos pés estão mais próximos do horizonte devem estar mais longe (desenho acima). Isso significa que eles devem parecer menores na imagem (desenho inferior esquerdo). Isso significa que o tamanho e a localização real dos pedestres em uma imagem dependem uns dos outros e da localização do horizonte. Para explorar isso, precisamos identificar o plano do chão, que é feito usando métodos de forma da textura. A partir dessa informação e de alguns pedestres prováveis, podemos recuperar um horizonte como mostrado na imagem central. À direita, caixas de pedestres aceitáveis dão esse contexto geométrico. Observe que os pedestres que na cena são mais altos devem ser mais baixos. Se não forem, são falsos positivos.

Existem várias estratégias para encontrar o horizonte, incluindo a busca de uma linha de horizonte irregular, com uma porção de azul acima dela, e a utilização da estimativa de orientação da superfície obtida da deformação de textura. Uma estratégia mais elegante explora o reverso de

nossas restrições geométricas. Um detector de pedestres que seja razoavelmente confiável é capaz de produzir estimativas do horizonte, se houver muitos pedestres na cena a diferentes distâncias da câmera. Isso ocorre porque a escala relativa dos pedestres é uma pista de onde está o horizonte. Por isso, podemos extrair uma estimativa do horizonte do detector; então usamos essa estimativa para podar os erros do detector de pedestre.

Se o objeto for familiar, pode-se estimar mais do que apenas a distância até ele porque o que ele parece na imagem depende fortemente de sua postura, isto é, sua posição e orientação com respeito ao telespectador. Isso tem muitas aplicações. Por exemplo, em uma tarefa de manipulação industrial, o braço do robô não pode pegar um objeto até que a postura seja conhecida. No caso de objetos rígidos, seja tridimensional ou bidimensional, esse problema tem solução simples e bem definida com base no método de alinhamento, que agora desenvolveremos.

O objeto é representado por M características ou pontos distintos m1, m2,…., mM no espaço tridimensional — talvez os vértices de um objeto poliédrico. Eles são medidos em algum sistema de coordenadas que é natural para o objeto. Os pontos são então submetidos a uma rotação R tridimensional desconhecida, seguida da translação de uma quantidade desconhecida t e projeção para dar origem aos pontos característicos de imagem p1, p2,…, pn no plano da imagem. Em geral, N ≠ M porque alguns pontos do modelo podem ser ocluídos, e o detector de atributos poderá perder algumas características (ou inventar características falsas devido ao ruído). Podemos expressar isso como

$$p_i = \Pi(\mathbf{R}m_i + \mathbf{t}) = Q(m_i)$$

para um ponto do modelo mi e o ponto de imagem correspondente p1. Aqui, R é uma matriz de rotação, t é uma translação e Π indica a projeção de perspectiva ou uma de suas aproximações, tal como projeção ortográfica em escala. O resultado líquido é uma transformação Q que trará o ponto de modelo mi em alinhamento com o ponto de imagem pi. Embora não conheçamos Q inicialmente, sabemos (para objetos rígidos) que Q deve ser o mesmo para todos os pontos do modelo.

Podemos resolver Q, dadas as coordenadas tridimensionais de três pontos do modelo e suas projeções bidimensionais. A intuição é a seguinte: podemos escrever as equações relacionando as coordenadas de pi às de mi. Nessas equações, as quantidades desconhecidas correspondem aos parâmetros de rotação da matriz R e o vetor de translação t. Tendo equações suficientes, devemos ser capazes de resolver Q. Não demonstraremos aqui, apenas afirmaremos o resultado seguinte:

- Dados três pontos não colineares m1, m2 e m3 no modelo e suas projeções ortográficas em escala p1, p2 e p3 no plano de imagem, existem exatamente duas transformações da estrutura coordenadas do modelo tridimensional para uma estrutura coordenada de imagem bidimensional.

- Essas transformações estão relacionadas por uma reflexão em torno do plano da imagem e podem ser calculadas por uma solução de forma fechada simples. Se pudéssemos identificar as características do modelo correspondentes às três características da imagem, poderíamos calcular Q, a pose do objeto.

Vamos especificar a posição e a orientação em termos matemáticos. A posição de um ponto P na cena é caracterizada por três números, as coordenadas de P (X, Y, Z) em uma estrutura coordenadas com a sua origem no orifício e o eixo Z ao longo do eixo óptico. O que temos disponível é a projeção em perspectiva (x, y) do ponto na imagem. Isso especifica o raio do orifício ao longo do qual P se encontra; o que não sabemos é a distância. O termo "orientação" poderia ser usado em dois sentidos:

1. A orientação do objeto como um todo. Pode-se especificar isso em termos de uma rotação tridimensional relacionando sua estrutura de coordenadas com a câmera.

2. A orientação da superfície do objeto em P. Isso pode ser especificado por um vetor normal, n, que é um vetor especificando a direção perpendicular à superfície. Muitas vezes, expressamos a orientação da superfície usando as variáveis obliquidade e inclinação. Obliquidade é o ângulo entre o eixo Z e n. Inclinação é o ângulo entre o eixo X e a projeção de n no plano da imagem.

Quando a câmera se move em relação a um objeto, tanto a distância do objeto como a sua orientação mudam. O que é preservado é a forma do objeto. Se o objeto for um cubo, esse fato não muda quando o objeto se move. Geômetros tentam formalizar a forma há séculos; o conceito básico é que forma é o que permanece inalterado sob algum grupo de transformações, por exemplo, combinações de rotações e translações. A dificuldade reside em achar uma representação de forma global, que seja geral o suficiente para lidar com a grande variedade de objetos no mundo real — não apenas formas simples, como cilindros, cones e esferas — e ainda podem ser facilmente recuperadas a partir de entrada visual. O problema de caracterizar a forma local de uma superfície é muito melhor compreendida. Essencialmente, pode-se fazer isso em termos de curvatura: como é que a superfície normal muda à medida que alguém se move em diferentes direções na superfície? Para um plano, não há nenhuma alteração. Para um cilindro, se alguém se move paralelamente ao eixo, não há mudança, mas na direção perpendicular a superfície normal gira a uma velocidade inversamente proporcional ao raio do cilindro, e assim por diante. Tudo isso é estudado no assunto chamado geometria diferencial.

A forma de um objeto é relevante para algumas tarefas de manipulação (por exemplo, decidir onde agarrar um objeto), mas seu papel mais significativo é o reconhecimento de objetos, em que a forma geométrica, a cor e a textura fornecem as pistas mais importantes para nos permitir identificar objetos, classificar o que está na imagem como exemplo de alguma classe já vista antes, e assim por diante.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig