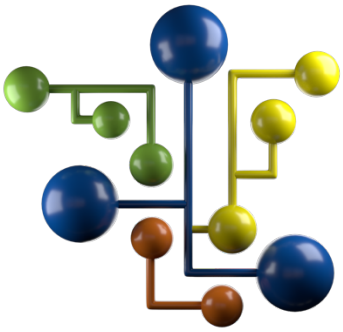

Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Rastreamento de Pessoas em Vídeos

O rastreamento de pessoas no vídeo é um problema prático importante. Se pudéssemos relatar de forma confiável a localização dos braços, pernas, tronco e cabeça em sequências de vídeo, poderíamos construir muitas interfaces de jogo melhoradas e sistemas de vigilância. Os métodos de filtragem não tiveram muito sucesso com esse problema porque as pessoas podem produzir acelerações grandes e se movem muito rápido. Isso significa que, para um vídeo de 30 Hz, a configuração do corpo no quadro i não restringe a configuração do corpo no quadro i +1.

Atualmente, os métodos mais eficazes exploram o fato de que a aparência muda muito lentamente a partir do quadro a quadro. Se pudéssemos inferir um modelo de aparência de um indivíduo de um vídeo, poderíamos usar essa informação em um modelo de estrutura pictórica para detectar essa pessoa em cada quadro do vídeo. Poderíamos então ligar essas localizações ao longo do tempo para rastrear.

Existem várias maneiras de inferir um bom modelo de aparências. Consideramos o vídeo como uma grande pilha de fotos da pessoa que desejamos rastrear. Podemos explorar essa pilha procurando por modelos de aparências que explicam muitas das imagens. Isso funcionaria através da detecção de segmentos do corpo em cada quadro usando o fato de que os segmentos têm arestas mais ou menos paralelas. Tais detectores não são particularmente confiáveis, mas os segmentos que queremos encontrar são especiais. Eles vão aparecer pelo menos uma vez na maioria dos quadros de vídeo; tais segmentos podem ser encontrados agrupando-se as respostas do detector. É melhor começar com o tronco porque é grande e porque os detectores de tronco tendem a ser confiáveis. Uma vez que temos um modelo de aparência do tronco, os segmentos superiores de perna, e assim por diante, deverão aparecer perto do tronco. Esse raciocínio gera um modelo de aparência, mas pode não ser confiável se as pessoas aparecerem contra um fundo quase fixo, onde o detector de segmento gera muitos falsos positivos. Uma alternativa é estimar a aparência para muitos dos quadros de vídeo, reestimando diversas vezes a configuração e a aparência; vemos então se um modelo de aparência explica muitos quadros. Uma alternativa bastante confiável na prática é a aplicação de um detector para uma configuração de corpo fixo para todos os quadros. Uma boa opção de configuração é a que é fácil de detectar de forma confiável, e onde há uma chance forte de a pessoa aparecer nessa configuração mesmo em uma sequência curta (caminhadas laterais é uma boa escolha). Ajustamos o detector para ter uma taxa de falso positivo baixa; assim sabemos quando ele responde que encontramos uma pessoa real; e, como localizamos o seu tronco, braços, pernas e cabeça, sabemos como esses segmentos se parecem.

Utilização da Visão

Se os sistemas de visão pudessem analisar o vídeo e compreender o que as pessoas estão fazendo, seríamos capazes de projetar melhor edifícios e lugares públicos, recolher e utilizar dados sobre o que as pessoas fazem em público; construir sistemas de vigilância mais precisos, mais seguros e menos intrusivos; construir comentaristas esportivos computadorizados; e construir interfaces homem-computador que observassem as pessoas e reagissem ao seu comportamento. Aplicações de interfaces reativas vão de jogos de computador que fazem um personagem levantar e mover-se até sistemas que economizam energia através do gerenciamento de calor e luz em um prédio ao encontrar onde estão os ocupantes e o que estão fazendo. Alguns problemas são bem

compreendidos. Se as pessoas são relativamente pequenas no quadro de vídeo, e o fundo é estável, é fácil detectar as pessoas, subtraindo uma imagem de fundo a partir do quadro atual. Se o valor absoluto da diferença é grande, essa subtração do fundo declara o pixel como sendo um pixel de imagem de frente; ao ligar blobs de imagens de frente ao longo do tempo, obtém-se uma pista.

Alguns comportamentos estruturados como balé, ginástica ou tai chi têm vocabulários específicos de ações. Quando realizados contra um fundo simples, os vídeos dessas ações são fáceis de lidar. A subtração de fundo identifica as principais regiões em movimento, e podemos construir características HOG (mantendo o controle do fluxo, em vez da orientação) para apresentar para um classificador. Podemos detectar padrões consistentes de ação com uma variante do nosso detector de pedestres, onde as características de orientação são recolhidas em histogramas ao longo do tempo e do espaço.

Algumas ações humanas complexas produzem padrões consistentes de aparência e movimento. Por exemplo, beber envolve movimentos da mão em frente do rosto. As três primeiras imagens são detecções corretas de beber, a quarta é um falso positivo (o cozinheiro está olhando para o pote de café, mas não bebendo).

Problemas mais gerais permanecem em aberto. A grande questão de pesquisa é associar as observações do corpo, os objetos próximos aos objetivos e as intenções das pessoas em movimento.

Uma fonte de dificuldade é a falta de um vocabulário simples do comportamento humano. O comportamento é como a cor, em que as pessoas tendem a pensar que conhecem uma porção de nomes de comportamento, mas não conseguem produzir longas listas de tais palavras quando necessário. Há bastante evidência de que os comportamentos combinam — você pode, por exemplo, beber um milk-shake dentro de um caixa eletrônico —, mas ainda não sabemos quais são as peças, como funciona a composição ou quantas combinações pode haver. Uma segunda fonte de dificuldade é que não sabemos que características expõe o que está acontecendo. Por exemplo, saber que alguém está perto de um caixa eletrônico pode ser o suficiente para dizer que ele está entrando no caixa eletrônico. A terceira dificuldade é que o raciocínio habitual sobre a relação entre dados de treinamento e de teste não é confiável. Por exemplo, não podemos argumentar que um detector de pedestres é seguro simplesmente porque tem bom desempenho em um grande conjunto de dados porque esse conjunto de dados pode muito bem omitir fenômenos importantes, mas raros (por exemplo, pessoas utilizando bicicletas). Não gostaríamos que nosso motorista automatizado atropelasse um pedestre que passou a fazer algo incomum.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig