Campinas, 13 a 19 de junho de 2016

Quadro a quadro

Lacuna semântica em segmentação de imagens é objeto de estudo no Instituto de Computação

CARLOS ORSI
carlos.orsi@reitoria.unicamp.br

Fotos: Antonio Scarpinetti

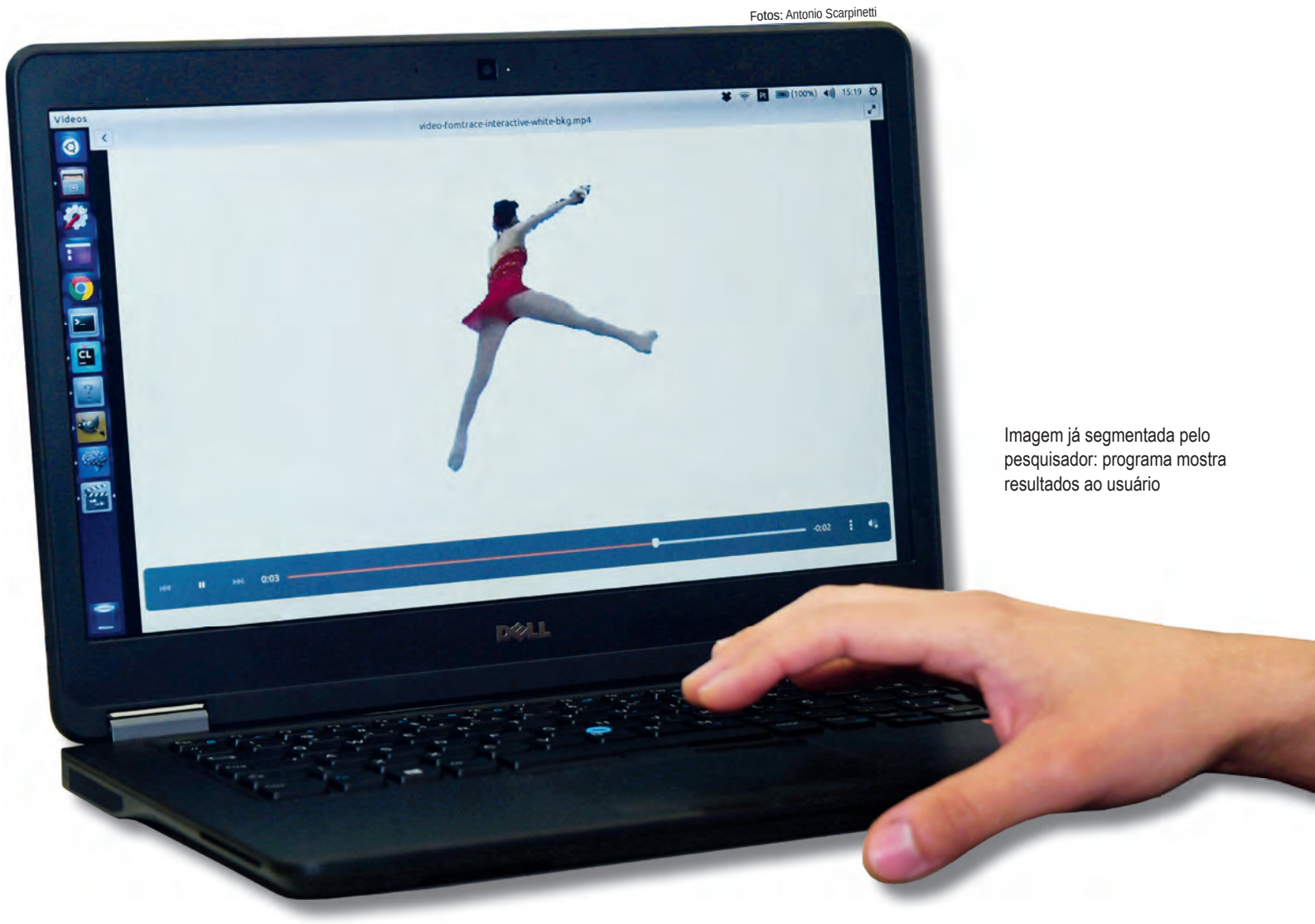

Imagem já segmentada pelo pesquisador: programa mostra resultados ao usuário

O diretor de efeitos especiais de um filme de ficção científica precisa lançar a imagem de seu protagonista, que realiza diversas manobras acrobáticas diante da parede do estúdio, sobre um fundo futurista de naves espaciais, tiros e explosões; um divulgador de ciência deseja inserir a imagem de um pássaro voando, captada na natureza, sobre um fundo animado com legendas que explicarão os movimentos do voo; um biólogo gostaria de isolar, numa animação criada a partir de dezenas de imagens de uma cultura de células em divisão, o percurso de apenas uma ou duas delas.

Cada uma dessas situações comporta soluções diferentes – no cinema, por exemplo, a mais usada é a técnica da tela verde, diante da qual o ator se movimenta para, então, ter sua imagem inserida no fundo adequado – mas todas elas são exemplos de um mesmo tipo de problema, o de segmentação da imagem, e que hoje é um dos desafios da interação entre ser humano e computador.

"É algo que parece simples, o ser humano bate o olho e diz: este é o meu objeto de interesse, o resto é fundo", explicou Thiago Spina, autor da tese de doutorado *"Interactive segmentation of objects in images and videos using graphs and fuzzy models of content knowledge"* ("Segmentação interativa de objetos em imagens e vídeos utilizando grafos e modelos nebulosos de conhecimento de conteúdo"), defendida no Instituto de Computação (IC) da Unicamp.

"Só que para a máquina preencher essa lacuna semântica de, dada a indicação do usuário, determinar com precisão quais são os pixels, os pontos que compõem a imagem, que ele realmente quer isolar não é trivial", descreve ele. "Ainda mais em vídeo, onde é preciso propagar a segmentação para os demais quadros e há movimentos em que o objeto de interesse não só muda de posição como também de forma – por exemplo, um braço pode ficar escondido atrás do corpo por alguns segundos, ou a roda de uma bicicleta mudar de ângulo".

Em sua tese, Spina apresenta um software que permite ao usuário marcar, com um pincel colorido controlado pelo mouse, como o que existe em programas comuns de edição de imagens, que parte do quadro inicial do vídeo corresponde ao objeto de interesse e que parte é o fundo que deve ser descartado. O programa então propaga a seleção para os demais quadros e apresenta o resultado ao usuário, que pode refinar suas marcações para corrigir eventuais erros.

Competição

Para fazer a segmentação e a propagação, os pixels "semente" – aqueles que foram marcados como parte do objeto e parte do fundo – "competem entre si para tentar conquistar os mais parecidos com eles", disse Spina. "E o que significa os mais parecidos? São aqueles cuja cor não é tão diferente da cor da semente. Como a cor não é tão diferente, o gradiente tende a ser baixo". Usando como exemplo um vídeo que mostra um ciclista realizando manobras acrobáticas, ao ar livre e sob um céu azul, o pesquisador descreve:

"Todos os pontos que estavam na camisa do ciclista tinham um gradiente baixo entre si, porque era todo mundo mais ou menos da mesma cor. A mesma coisa ocorreu aqui. Agora, quando o programa viu que o céu era bem diferente da camisa, aí o gradiente era alto, o peso era bem distinto e a competição entre semente de objeto e semente de fundo parou ali em cima, no contorno do ciclista. Assim, ele descobriu onde parar".

A dificuldade aumenta quando se pede ao programa que aplique os mesmos critérios baseados em pixels-semente aos demais quadros do vídeo, de modo automático. "Quando a gente quer fazer isso no vídeo inteiro, tem não só o problema de gradiente baixo, gradiente alto, como de oclusão parcial, quando uma parte do objeto de interesse passa por trás de algum objeto que é parte do fundo, ou de mudança de topologia, quando a silhueta, o contorno do objeto de interesse muda por conta de uma mudança de posição".

Mais dificuldades surgem no caso de movimentos muito rápidos, que geram um borrão onde a divisa entre objeto e fundo não é bem definida. Por isso são necessárias as correções feitas pelo usuário. Uma métrica da qualidade do sistema é o número de correções necessárias.

"A gente fez algumas medidas do número de correções. Normalmente, se você consegue fazer uma segmentação manual a cada cinco quadros, você tem um bom resultado, porque já está reduzindo o esforço do usuário em 80%", disse Spina. "Então o software que tem por aí, o meu inclusive, tenta chegar nessa meta ou superá-la".

Publicação

Tese: *"Interactive segmentation of objects in images and videos using graphs and fuzzy models of content knowledge"*
Autor: Thiago Spina
Orientador: Alexandre Falcão
Unidade: Instituto de Computação (IC)

Conhecimento humano

O orientador da tese, Alexandre Falcão, explica por que, a despeito da crescente sofisticação dos sistemas de informática – com máquinas capazes de derrotar seres humanos em jogos complexos como xadrez ou go, ou mesmo reconhecer rostos em vídeos de segurança – a intervenção humana ainda é necessária na tarefa de segmentação.

"A percepção humana da imagem é composta de duas partes: uma parte é a informação que podemos chamar de local, que você consegue extrair das cores e da textura. Você consegue perceber que o sujeito está com uma camisa vermelha. Então, todos os pontos da imagem, os pixels, que são vermelhos e pertencem à camisa têm uma textura similar, naquela adjacência. E aí você tem a camisa vermelha sobre um fundo branco, há um contraste. Esse tipo de informação é algo que a máquina consegue explorar bem", disse ele.

"Agora, quando a gente fala da outra percepção que o ser humano tem, essa é a percepção que vem da experiência, do conhecimento prévio: você sabe que o conjunto de pontos vermelhos forma uma camisa, que dentro da camisa há um homem, que o homem tem pernas, etc. Essa percepção é muito mais complicada para a máquina, a menos que ela seja treinada para chegar próximo ao que o ser humano consegue fazer". O orientador prossegue: "Então, o que acontece aí? Nesse método, ela tenta incorporar um pouco da percepção que o ser humano tem, através de um modelo: na hora que você segmenta a silhueta do indivíduo num quadro, cria-se um modelo a partir da segmentação, e o modelo é propagado para o frame seguinte".

Thiago Spina, autor do estudo: "O contorno do objeto de interesse muda por conta de uma mudança de posição"

O orientador da pesquisa, professor Alexandre Falcão: "Dificilmente você vai ver uma técnica de 'deep learning' que segmente todos os objetos de um vídeo automaticamente"

Outro componente do método, explica Falcão, está associado à percepção de como os elementos da imagem se conectam entre si. "Dois pontos na camisa estão mais fortemente conexos do que um ponto na camisa e outro no fundo. Porque, entre os dois pontos da camisa é possível achar um caminho que vai de um ao outro passando por outros que também satisfazem esse mesmo critério de cor e de textura. Já para ir de um ponto na camisa e um ponto do fundo é preciso, necessariamente, passar por uma transição, sendo que os elementos de fundo também são mais parecidos entre si do que com os elementos da camisa".

Falcão explica que tarefas como o reconhecimento de faces são mais fáceis para os computadores. "São tarefas em que você repete várias vezes a imagem daquele indivíduo para a máquina, então ela vai aprender que se trata daquele indivíduo. Mas outra coisa é determinar a extensão espacial: dizer que determinada pessoa está numa imagem e delimitar exatamente qual a extensão espacial ocupada pela pessoa são tarefas bem diferentes".

"A maioria das aplicações mais modernas de 'deep learning' (aprendizado profundo), em que grandes empresas como o Google estão investindo tanto, requer apenas o reconhecimento de padrões, a detecção de um evento, a identificação de um determinado objeto, a partir da apresentação consistente daquela imagem, forma ou textura", disse. "Mas isso tudo se baseia em ações repetitivas. Quando a forma ou a textura varia muito e a decisão requer outras informações que são difíceis de modelar, tem que ter um ser humano no processo. Acho que o grande exemplo é vídeo. Esse trabalho do Thiago é um exemplo: dificilmente você vai ver uma técnica de 'deep learning' que segmente todos os objetos de um vídeo automaticamente".

Além das aplicações nas áreas de arte e entretenimento, sistemas de segmentação de imagens também são úteis na análise de imagens médicas e em ouros campos da ciência.

Células

Spina está atualmente em contato com pesquisadores do Instituto de Tecnologia da Califórnia (Caltech) interessados em segmentar vídeos montados a partir de sequências de imagens de células vegetais em crescimento.

"Eles têm trabalhado com segmentação de células-tronco em imagens de plantas", disse. "Então, em vez de ser um objeto apenas a segmentar, são centenas de objetos, e uma vez que tenham segmentado essas centenas, eles têm que garantir que a segmentação esteja correta, isto é, precisam não só utilizar o método automático que já têm, como também fazer a parte de correção, com o auxílio do usuário, e as imagens são 3D também", descreve. O objetivo, explica Spina, é entender como as células crescem e se dividem. "A ideia é criar simulações computacionais que permitam compreender como funciona o processo de desenvolvimento dos órgãos das plantas, que se transformam em grãos e frutos para a alimentação humana, por exemplo".

"Para a análise dos vídeos das células, a propagação da segmentação de um quadro anterior para o quadro seguinte ainda é um problema em aberto", disse o autor da tese. "Por ora, o que é feito é segmentar cada quadro individualmente, com o auxílio do usuário, e depois tentar determinar a correspondência entre as células segmentadas em quadros adjacentes, para determinar se ocorreu um processo de crescimento, duplicação ou morte".