Bilayer Segmentation of Live Video in Uncontrolled Environments for Background Substitution: An Overview and Main Challenges

S. R. R. Sanches, R. Nakamura, V. F. Silva and R. Tori

***Abstract*— Bilayer segmentation of live video in uncontrolled environments is an essential task for home applications in which the original background of the scene must be replaced, as in videochats or traditional videoconference. The main challenge in such conditions is overcome all difficulties in problem-situations (e.g., illumination change, distract events such as element moving in the background and camera shake) that may occur while the video is being captured. This paper presents a survey of segmentation methods for background substitution applications, describes the main concepts and identifies events that may cause errors. Our analysis shows that although robust methods rely on specific devices (multiple cameras or sensors to generate depth maps) which aid the process. In order to achieve the same results using conventional devices (monocular video cameras), most current research relies on energy minimization frameworks, in which temporal and spacial information are probabilistically combined with those of color and contrast.**

***Keywords*— Background Substitution, Bilayer Segmentation, Computer Vision, Image Processing, Live Video.**

I. INTRODUÇÃO

A SEGMENTAÇÃO (extração de elementos de interesse em imagens) com o objetivo de extrair uma ou mais pessoas em primeiro plano é um problema que tem sido alvo de pesquisas desde o início do século passado [1]. Produções de cinema e televisão que, até o final da década de 1970, apoiavam-se em tecnologia óptica (analógica) [2], tradicionalmente utilizam métodos que permitem isolar elementos de uma imagem – na maioria das vezes são pessoas em primeiro plano –, com o objetivo de gerar cenas a partir da combinação desses elementos com novos planos de fundo [3], [4], [5].

Os métodos mais tradicionais de segmentação partem do princípio de que a captura do vídeo se realiza em ambientes controlados, com fundos de cor única – normalmente azul ou verde – e iluminação devidamente direcionada para que a tonalidade do fundo se mantenha constante [5], [6], [7]. De forma simplificada, tais métodos procuram isolar o elemento de interesse por meio da eliminação da cor do fundo, que é conhecida do sistema.

A partir da década de 1980, novos métodos começaram a surgir, baseados nos recursos da tecnologia digital [7] e, mais recentemente, métodos capazes de extrair elementos de interesse, não apenas em tempo real senão também a partir de imagens com planos de fundo arbitrários, passaram a ser desenvolvidos [8], [9], [10], [11], [12], [13], [14].

Essa possibilidade impulsionou pesquisas em outras áreas de aplicação a utilizarem imagens segmentadas, principalmente as voltadas para aplicações em que os elementos de interesse são pessoas posicionadas em primeiro plano na cena. Exemplos a serem citados são os sistemas de videoconferências ou *videochats* [9], [10], [11], [14] e de Realidade Aumentada [15].

Muitas são as variações que podem ocorrer, no ambiente ou no comportamento do elemento de interesse durante o processo de captura do vídeo. Tais variações podem interferir diretamente na qualidade da segmentação realizada nessas condições; inclusive, dependendo do método utilizado, a simples presença de objetos com superfícies reflexivas pode tornar-se um problema.

Idealmente, os métodos de segmentação que atuam em ambientes não controlados (também chamados de ambientes naturais) deveriam tratar de todas essas situações. No entanto, o custo computacional, além da dificuldade implícita do tratamento de algumas delas, faz que boa parte desses métodos tratem apenas os problemas de maior ocorrência na aplicação. Problemas de menor ocorrência, porém, também devem ser identificados e tratados pelo algoritmo, para que se obtenham soluções robustas.

Inserido neste contexto, o presente trabalho tem como objetivos: (i) apresentar os principais conceitos envolvidos no desenvolvimento de métodos de segmentação, capazes de extrair o elemento de interesse, em tempo real, a partir de uma sequência de imagens obtidas em ambiente não controlado; (ii) expor o estado-da-arte, na forma de uma classificação dos métodos mais praticados; e (iii) levantar os problemas que devam ser tratados, tanto os de maior quanto os de menor ocorrência.

Para isso, foram analisados os trabalhos encontrados na literatura abordem o desenvolvimento de métodos cuja finalidade seja a divisão de uma imagem de entrada em duas camadas (*bilayer*): elemento de interesse e plano de fundo, para posterior substituição do plano de fundo original.

S. R. R. Sanches, Escola Politécnica da Universidade de São Paulo (POLI-USP), São Paulo, Brasil, silviorrs@usp.br

R. Nakamura, Escola Politécnica da Universidade de São Paulo (POLI-USP), São Paulo, Brasil, ricardo.nakamura@poli.usp.br

V. F. Silva , Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP), São Paulo, Brasil, valdinei.freire@usp.br

R. Tori, Escola Politécnica da Universidade de São Paulo (POLI-USP), São Paulo, SP, Brasil, tori@acm.org

Métodos utilizados em aplicações que segmentam múltiplos elementos de interesse, como compressão de vídeo [16], ou nas que não têm como objetivo a segmentação para substituição do fundo da cena – identificação de pessoas para sistemas de segurança [17], reconhecimento de gestos [18] e rastreamento [19] –, não foram analisados neste trabalho.

Nesses casos, embora muitas abordagens sejam aplicáveis, a precisão na separação do elemento de interesse do seu fundo original não é um requisito tão rígido quanto nos métodos utilizados em aplicações de substituição de fundo. Aplicações em que o fundo deve ser substituído exigem métodos de segmentação que proporcionem qualidade na combinação com o novo fundo, resolvendo, inclusive, o problema do *matting* [20], que será discutido na seção III, mas com soluções que sejam eficientes para atender os requisitos de execução em tempo real [10], [21].

As demais seções deste artigo estão organizadas da forma como segue. Na seção II, apresentam-se as aplicações principais e em potencial, que se apoiam em processos de segmentação de imagens (e vídeos) digitais. Os conceitos envolvidos, enfatizando a importância da representação de transparência de pixels e da geração de uma camada de primeiro plano, são detalhados na seção III. Na seção IV, discute-se a segmentação tratada como um problema de minimização de energia (otimização), além da utilização de grafos como estrutura de auxílio ao processo. A quase totalidade dos métodos atuais adotam essas abordagens.

Uma classificação dos métodos que representam o estado-da-arte em segmentação de vídeo, realizada em tempo real e em ambientes não controlados, é apresentada na seção V. Para a classificação, consideraram-se as abordagens mais praticadas e o tipo de equipamento necessário.

Na seção VI, faz-se um levantamento dos principais problemas encontrados quando se aplica cada abordagem analisada. Finalmente, na seção VII, expõem-se as conclusões.

II. Principais Aplicações

A evolução dos métodos de segmentação de imagens com o objetivo de extrair pessoas ou objetos em primeiro plano, que, mais tarde foram estendidos para vídeos, deve-se principalmente à necessidade de a indústria cinematográfica combinar, de forma convincente, imagens filmadas em locais diferentes em uma única faixa de filme [22].

Embora existissem abordagens alternativas [2], as composições criadas desde as primeiras décadas do século passado utilizavam métodos que, de certa forma, isolavam os elementos de interesse – pessoas, na maioria dos casos –, para posterior combinação com um novo fundo [1], [3], [4].

Mesmo com a substituição dos equipamentos ópticos por computadores e *software* especializados, a maioria dos conceitos desenvolvidos para aquela tecnologia permaneceram aplicáveis à esfera digital. No entanto, o uso de computadores, potencializado pelo surgimento da computação gráfica, representou um passo significativo na evolução [22] e na criação de novas técnicas, e, consequentemente, de novas aplicações.

Apesar de haver um volume considerável de pesquisas voltadas à indústria cinematográfica [20] – ou aplicações similares que realizam o processo *offline* –, a possibilidade de extrair o elemento de interesse em tempo real fez surgir novas linhas direcionadas a outras áreas, em que essa característica constitui um requisito.

Segmentar uma imagem (em duas camadas) em tempo real para substituir o fundo original da cena é uma tarefa comum em programas de televisão, exibidos ao vivo, como ocorre nos informativos de previsão do tempo apresentados em telejornais. O fundo original da imagem, que possui um apresentador em primeiro plano, é substituído pelo mapa de determinada região do país (Fig. 1).

Figura 1. Substituição de fundo utilizada em telejornais para informar a previsão do tempo [10]. O fundo original da cena é substituído por um mapa de determinada região do país. Métodos que agem em ambientes não controlados podem ser utilizados nesse tipo de aplicação.

Nesse caso, o ambiente normalmente é formado por cor única e a segmentação ocorre utilizando-se de métodos que se baseiam na eliminação da cor do fundo [7]. Métodos que atuam em ambientes não controlados, no entanto, também podem ser utilizados nesse tipo de aplicação [10], [23], possibilitando, inclusive, que a captura do vídeo se realize em ambientes externos [24].

Do mesmo modo, tornam-se aplicações em potencial os sistemas de Realidade Aumentada [15] e os jogos imersivos [25], em que, a representação humana no ambiente virtual (avatar) se constrói com base na imagem do usuário.

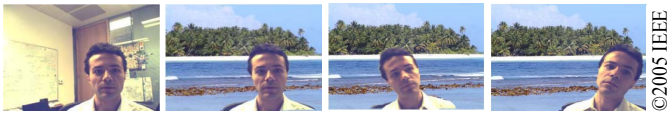
Nesses sistemas, em muitos casos, não se realiza uma simples substituição do fundo da cena. A imagem segmentada pode ser utilizada em modelo bidimensional, aplicada em um modelo geométrico ou volumétrico [26], ou podem se combinar várias camadas de imagem, que contenham o mesmo elemento de interesse em diferentes pontos de vista, para sintetizar um modelo 3D do avatar [27]. A precisão na separação do elemento de interesse do fundo original, nesses casos, também é fator preponderante.

Além disso, podem-se encontrar pesquisas voltadas a sistemas de videoconferência que realizam a segmentação da imagem dos participantes, com o objetivo de preservar o local da captura do vídeo [28], ou de produzir uma nova imagem com efeitos 3D [29].

Os *videochats*, que surgem com a popularização das conexões de rede de alta velocidade, também são um grupo de aplicações em potencial. Ao contrário das videoconferências tradicionais, em que os sinais de áudio e vídeo são, em alguns sistemas, transmitidos via satélite, os *videochats* são executados quase sempre em computadores pessoais

(inclusive *laptops*), utilizando a Internet como meio de comunicação.

Essas aplicações podem realizar a segmentação e a substituição de fundo como forma de redução de banda ou de obtenção de privacidade (Fig. 2), [9], [10], [11], [13], [14], [28], [30], [31]. Mesmo os telefones móveis 3G podem adicionar esse recurso aos seus serviços [32].



(a) (b) (c) (d)

Figura 2. Substituição de fundo como forma de obtenção de privacidade em videoconferências [28]. O plano de fundo original (a), que é arbitrário, é substituído por uma nova imagem, antes de os quadros de vídeo serem enviados pela rede aos demais participantes (b), (c) e (d).

Esquemas de compressão de vídeo [16], que segmentam múltiplos elementos de interesse em tempo real para identificar objetos (e pessoas), sistemas que necessitam do rastreamento [19] de pessoas, como os de segurança, em que indivíduos devem ser detectados e isolados para diminuir a área de atuação de algoritmos de análise de comportamento humano [17], e sistemas de reconhecimento de gestos [18], em que as mãos do usuário precisam ser isolada do fundo, são exemplos de aplicações que utilizam métodos de segmentação que atuam em ambientes não controlados.

Nesses casos, no entanto, o objetivo principal não é a segmentação para substituição do fundo da cena [17]. Embora muitas abordagens sejam aplicáveis, a precisão na separação do elemento de interesse do seu fundo original não é um requisito tão rígido quanto nos métodos utilizados em aplicações de substituição de fundo.

III. Segmentação Binária, Transparência de Pixels e Representação do Elemento de Interesse

Segundo [33], uma imagem (ou quadro de vídeo) pode ser definida como uma função bidimensional $z(x,y)$, onde x e y são coordenadas no plano espacial, e a amplitude de z, em cada par de coordenadas (x,y), é sua intensidade naquele ponto. Em imagens digitais, os valores de (x,y) e da amplitude de z são finitos.

O processo de segmentação consiste na subdivisão dessa imagem em estruturas com conteúdo semântico relevante para uma determinada aplicação [34]. Em outras palavras, o que determina o nível dessa subdivisão consiste no problema a ser resolvido, pois o processo apenas se finaliza quando o elemento de interesse para a aplicação em questão estiver isolado [33].

Uma prática comum em processos de segmentação é tratar o elemento de interesse, que foi extraído do seu contexto original, como uma camada de imagem. Para tornar a representação de uma camada de primeiro plano possível, faz-se necessária a utilização de formatos de pixel que permitam controlar sua transparência, como mostrado na Fig. 3.

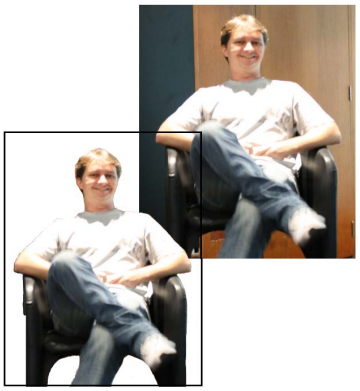


Figura 3. Representação de uma camada de primeiro plano. O elemento de interesse é extraído do plano de fundo original tornando transparentes os pixels que pertencem ao fundo e mantendo opacos os que pertencem ao elemento de interesse.

A tarefa de estimar níveis de transparência, conhecida na indústria cinematográfica como "problema do *matting*", foi definida matematicamente em [35], por meio da introdução do canal alfa, uma solução para controlar a interpolação linear das cores de duas camadas de imagens. Efeitos como suavização de bordas, além da preservação da transparência de objetos translúcidos, podem ser obtidos com esse tipo de recurso.

Segundo [35], a imagem I_z é modelada como uma combinação de uma camada de primeiro plano F_z e uma de fundo B_z, utilizando-se o canal alfa α_z como na equação

$$I = \alpha_z F_z + (1-\alpha_z) B_z \qquad (1)$$

onde α_z pode ser qualquer valor entre [0,1]. Se α_z =1 ou 0, o pixel pertence à camada de primeiro plano e à camada de fundo, respectivamente. Aos pixels cujas tonalidades são influenciadas pelas duas camadas – o que ocorre com frequência em objetos transparentes ou nas bordas de objetos opacos – valores intermediários de alfa devem ser estimados para que a separação do elemento de interesse seja mais precisa [20].

Na equação 1, restringindo-se o valor de alfa a assumir apenas os valores 0 ou 1, transforma-se o problema do *matting* em outro problema clássico: a segmentação binária, objeto de estudo deste trabalho, em que cada pixel pertence totalmente à camada de primeiro plano ou a camada de fundo [20].

Segundo [20], a maioria das pesquisas que buscam soluções para o problema do *matting* não trata o problema da segmentação binária. Algoritmos de *matting* são frequentemente custosos do ponto de vista computacional, uma vez que são normalmente voltados para composição de imagens estáticas ou vídeos pré-gravados. Por esse motivo, muitos métodos não têm compromisso com seu tempo de execução, pois podem ser aplicados *offline*.

Os métodos de segmentação desenvolvidos para essas aplicações (*offline*), normalmente, utilizam, além da imagem original, uma máscara da mesma imagem, chamada *trimap*,

que pode ser produzida manualmente pelo usuário, ou estimada por qualquer método de segmentação binária, que não é necessariamente parte do método principal, responsável pelo *matting* [20].

Um *trimap* é composto por três regiões: primeiro plano, plano de fundo e regiões desconhecidas, em que o pixel não pertence nem totalmente ao fundo, nem totalmente ao elemento de interesse [20]. Apenas nessas regiões ambíguas (ou desconhecidas) atuam os algoritmos que estimam valores intermediários de alfa.

Por outro lado, aplicações executadas em tempo real, baseadas ou não em *trimaps*, exigem que todo o processo seja automático e que a solução para o problema da segmentação em duas camadas seja de rápida execução e inclua estimativas de transparência de pixels na geração da camada de primeiro plano.

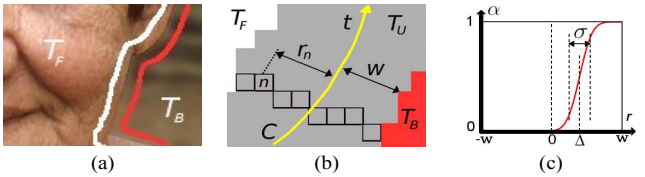
Uma técnica muita utilizada para estimar transparência de pixels em métodos de segmentação para aplicações de tempo real é conhecida como *border matting* [21]. A suavização nas bordas do elemento de interesse, que permite produzir cenas compostas com qualidade aceitável para aplicações de substituição de fundo, pode ser obtida por meio da técnica.

Resumidamente, o algoritmo de *border matting* toma como base uma polilinha C, mostrada em amarelo na Fig. 4(b), que contorna o elemento de interesse. O conjunto de pixels que pertencem a C pode ser obtido automaticamente a partir da segmentação binária, que produz uma borda rígida.

Um *trimap* $\{T_B, T_U, T_F\}$ é calculado (Fig. 4(a)), onde T_B e T_F são os conjuntos de pixels que pertencem ao plano de fundo e ao elemento de interesse, respectivamente. T_U é o conjunto de pixels em uma faixa de tamanho $\pm w$ pixels, de ambos os lados de C. O objetivo é calcular um mapa de transparência α_n, n$\in$ T_U, utilizando um modelo baseado no proposto em [38], que define a forma como α varia dentro de T_U. Os parâmetros do contorno C, $t=1,\ldots,T$ têm periodicidade T, a medida que a curva C é fechada [21].

Um índice $t(n)$ é atribuído para cada pixel $n \in T_U$, como mostrado na Fig. 4(b). Os valores de α são obtidos por meio de uma função, $g : \alpha_n = g(r_n; \Delta_{t(n)}, \sigma_{t(n)})$, onde r_n indica a distância do pixel n até C (Fig. 4(c)).

Os parâmetros Δ, σ determinam, respectivamente, o centro e a largura da transição de 0 até 1 no conjunto de valores possíveis de α. Todos os pixels com o mesmo índice t têm os mesmos valores de parâmetros Δ_t e σ_t. Os parâmetros $\Delta_1,\sigma_1,\ldots,\Delta_t,\sigma_t$ são estimados por meio de funções de minimização de energia, detalhadas em [21].



(a) (b) (c)

Figura 4. Border matting. (a) Imagem original sobreposta pelo trimap. (b) Notação para a parametrização do contorno C, obtido da segmentação binária, e do mapa de distâncias. Para cada pixel em T_U são atribuídos valores do parâmetro t, do contorno, e da distância r_n de C. (c) Função g, que define a atribuição dos valores de α [21].

IV. Segmentação como um problema de minimização de energia

Entre as características comuns identificadas em abordagens recentes de segmentação de vídeos em duas camadas, importa ressaltar o fato de muitas soluções tratarem a segmentação como um problema de minimização de energia.

Greig et al. [37] foram os primeiros a descobrirem que algoritmos de fluxo máximo/mínimo para otimização combinatorial podem ser utilizados também para minimizar funções de energia em visão computacional [38].

Em um processo de atribuição de rótulos a pixels de uma imagem, a partir de um conjunto de pixels P e de um conjunto de rótulos L, o objetivo é encontrar um rótulo f (i.e., realizar um mapeamento de P em L), que minimize determinada função de energia [39].

Para uma divisão da imagem em duas camadas, o conjunto L possui dois rótulos: elemento de interesse e plano de fundo (segmentação binária). Níveis de transparência de pixels são determinados, normalmente, em um passo posterior a segmentação binária, por meio de técnicas como *border matting* [21], discutida na seção III.

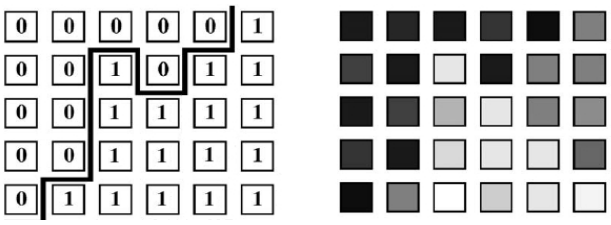
A função de energia utilizada no primeiro trabalho de [37], e por muitos métodos de segmentação atuais, a serem descritos na seção V, pode ser representada da forma

$$E(f) = \sum_{p \in P} D_p(f_p) + \sum_{p,q \in N} V_{p,q}(f_p, f_q) \qquad (2)$$

onde $N \subset PXP$ é o conjunto dos pixels que possuem relação de vizinhança. O termo $D_p(f_p)$ é uma função derivada dos dados observados, que mede o custo para atribuição do rótulo f_p ao pixel p. O termo $V_{p,q}(f_p, f_q)$ é responsável pela medição do custo, para atribuir os rótulos f_p , f_q aos pixels adjacentes p,q, e é utilizado para manutenção das descontinuidades na imagem [39]. Aos pixels vizinhos que possuem contraste alto são atribuídos custos menores, pois existe maior probabilidade de pertencerem a conjuntos diferentes. Na Fig. 5 mostra-se um exemplo de imagem rotulada.

Uma abordagem, aplicável em tempo real, bastante utilizada para minimizar energia, consiste em transformar a segmentação binária em um problema de corte em grafos [39]. A ideia básica é construir um grafo específico para determinada função de energia ser minimizada, de modo que o corte mínimo no grafo minimize também a energia do sistema. Grande parte dos trabalhos utiliza um arcabouço geral, proposto em [40], para essa finalidade.

Segundo [40], dado um grafo direcionado $G=(V,\varepsilon)$ com arestas de pesos não negativos e dois vértices terminais s (*source*) e t (*sink*), um corte s-t $C=(S,T)$ é um particionamento dos vértices em V em dois conjuntos disjuntos S e T, de modo que $s \in S$ e $t \in T$. O custo total do corte é a soma dos custos de todas as arestas que partem de S e chegam em T [39].



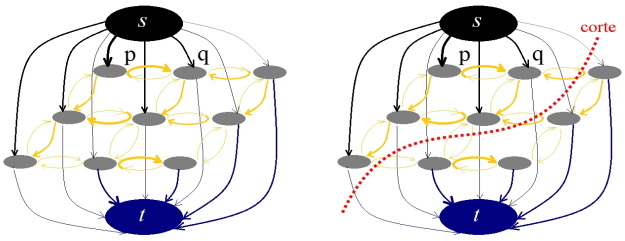
(a) (b)

Figura 5. Um exemplo de imagem rotulada, adaptada de [38]. A imagem (a) representa um conjunto de pixels P com intensidades observadas I_p, para cada $p \in P$. Em (b) é mostrada a atribuição de um rótulo $f_p \in \{0,1\}$ para cada pixel $p \in P$. As linhas mais espessas, mostradas em (b), representam rótulos de descontinuidades entre pixels vizinhos.

$$c(S,T) = \sum_{u \in S, v \in T, (u,v) \in \varepsilon} c(u,v) \qquad (3)$$

O problema do corte mínimo é encontrar um corte C com o menor custo, o que é equivalente ao cálculo do fluxo máximo de s até t. Existem muitos algoritmos que resolvem esse problema em tempo polinomial, como o do fluxo máximo otimizado, proposto em [38].

Importa observar que um corte $C=(S,T)$ é um processo de atribuição de rótulos f que mapeia o conjunto de vértices $V-\{s,t\}$ em $\{0,1\}$, onde $f(v)=0$ implica $v \in S$ e $f(v)=1$ implica $v \in T$. Isso significa que um corte é um particionamento binário de um grafo visto como uma atribuição de rótulos com dois valores possíveis [39]. Um exemplo de um grafo utilizado como estrutura auxiliar para minimização de energia aplicada à segmentação de vídeos é mostrado na Fig. 6.



(a) (b)

Figura 6. Exemplo de um grafo, adaptado de [38]. Os custos das arestas são representados por sua espessura. Um grafo de corte similar foi utilizado pela primeira vez na visão computacional em [37], para restauração de imagens binárias. O grafo G é mostrado em (a) e o corte em G pode ser visualizado em (b).

O conjunto de vértices V é formado pelos pixels $P \cup \{s-t\}$, e o custo entre $\{s,t\}$ para elementos em P é justamente a função D_p, ao passo que o custo entre elementos de P é justamente $V_{p,q}$.

V. Classificação das Abordagens

Em aplicações em que a extração do elemento de interesse se realiza em ambientes controlados, os erros de classificação de pixels podem ser evitados por meio da intervenção do usuário. O direcionamento manual de luzes e a distribuição dos elementos da cena, por exemplo, podem ser adequados, para que a cor do fundo se mantenha constante, impedindo a ocorrência de sombras, reflexos ou ruídos sobre o fundo.

Em ambientes não controlados, por sua vez, o plano de fundo é arbitrário e qualquer situação que atrapalhe a segmentação deve ser tratada pelo algoritmo, evitando intervenções do usuário para modificar o ambiente. Nesses casos, como não existe o conhecimento prévio da cor do fundo, outras informações, que podem ser obtidas da sequência de imagens, passam a ser fundamentais para que um elemento de interesse seja isolado. Algumas abordagens utilizam, ainda, equipamentos específicos, ou mais de um dispositivo para obter novas informações que auxiliem a segmentação.

Grande parte dos métodos automáticos computacionalmente eficientes para execução em tempo real trabalha com essas informações, na forma de um conjunto de "cortes" [10]. Cor, contraste, movimento e estéreo são exemplos de cortes muito utilizados. Esses cortes combinam-se probabilisticamente e aplicam-se na imagem por meio de algum arcabouço de minimização de energia, como o mostrado na seção IV. Alguns métodos mais simplificados, no entanto, utilizam essas informações (ou apenas uma delas) de formas alternativas.

O fato de determinadas abordagens se apoiarem em dispositivos específicos, ou de exigir calibração de mais de um dispositivo, pode restringir sua aplicabilidade. Em aplicações executadas em ambientes domésticos, como *videochats*, por exemplo, imagina-se que a maioria dos participantes possuam computadores e câmeras de vídeo convencionais.

Essa observação sugere que uma classificação das abordagens considere dois grupos principais: as abordagens executadas a partir de captura realizada por câmeras monoculares (convencionais) e as que necessitam de entrada binocular ou de equipamento específico (não convencionais) para produzir informações que auxiliem a segmentação.

Abordagens apoiadas em vídeo monocular podem ser divididas em dois subgrupos, cujos métodos são classificados de acordo com a técnica adotada. São eles: subtração de fundo e movimentação do elemento de interesse. Ainda que muitos métodos utilizem mais de uma técnica, uma delas, normalmente, tem maior importância que as demais no processo – i.e., sua utilização isolada resulta na rotulação correta da maioria dos pixels da imagem. A classificação sugerida neste trabalho toma como base a técnica principal utilizada no método.

As abordagens apoiadas em equipamentos específicos utilizam esse tipo de recurso para geração de mapas de profundidade da cena. Desse modo, a distância de cada pixel em relação à câmera constitui-se na principal informação a ser utilizada no processo de segmentação. Métodos que pertencem a esse grupo podem ser divididos também em dois subgrupos: os métodos baseados em estéreo e os baseados em sensores.

Nas subseções seguintes, será apresentada uma visão geral dos trabalhos que representam o estado-da-arte em

segmentação em tempo real de sequência de imagens em ambientes não controlados, considerando a classificação descrita. A Fig. 7 exibe um diagrama que sintetiza tal classificação.

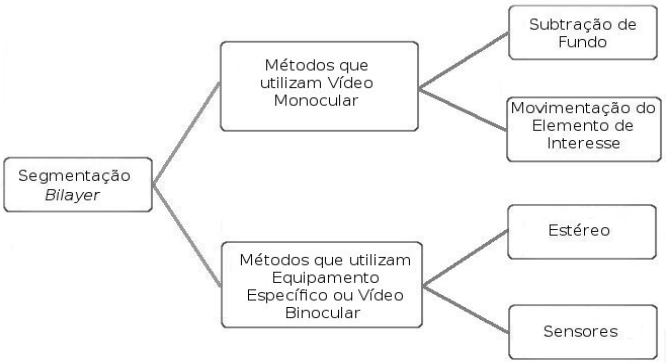


Figura 7. Classificação dos métodos de segmentação de sequência de imagens em duas camadas que atuam em ambientes não controlados. A utilização ou não de equipamento específico é o principal critério para o agrupamento dos métodos, seguido da técnica principal adotada como base.

A. Métodos que utilizam Vídeo Monocular

Grande parte dos métodos analisados são capazes de realizar a segmentação utilizando uma única câmera de vídeo convencional (vídeo monocular). Tais métodos podem ser divididos em dois subgrupos, classificados de acordo com sua abordagem principal.

1) Subtração de Fundo: A abordagem da subtração de fundo [43] consiste, basicamente, na comparação do quadro de vídeo no tempo atual (Fig. 8(b)) com um modelo do fundo (Fig. 8(a)). Como mostrado na Fig. 8(c), a camada de primeiro plano é gerada com base nos pixels não coincidentes dessas duas imagens. Esses pixels pertencerão ao elemento de interesse.

Métodos mais simplificados calculam a diferença do quadro atual e do anterior com base em um *threshold* [41], [42] ou calculam um modelo do plano de fundo por meio da média ou da mediana de alguns quadros anteriores [43]. Outros utilizam ainda informações do quadro atual, considerando também uma taxa de aprendizado [41]. Esses métodos, denominados básicos [41], apoiam-se na história recente dos pixels e não estabelecem quaisquer correlações espaciais entre pixels vizinhos.

(a) (b) (c)

Figura 8. Método da Subtração de Fundo [48]. Em (a) exibe-se um modelo do fundo e em (b) um quadro de vídeo no tempo atual. Os pixels não coincidentes nas duas imagens fazem parte do elemento de interesse (c). Cores semelhantes no fundo e no elemento de interesse podem provocar erros de classificação, como mostrado em (c).

Algoritmos mais sofisticados, por sua vez, utilizam, por exemplo, misturas de modelos de cores gaussianos [44], estimadores de densidade de *kernel* [45], estimadores de *Mean-Shift* [46] ou decomposição da imagem em autoespaços (*Eigenbackground*) [47]. Desse modo, obtêm-se métodos capazes de lidar com planos de fundo que apresentam maiores variações [42].

Na utilização de métodos baseados em subtração de fundo, a maior dificuldade encontrada não se relaciona à diferenciação dos pixels em si, mas na construção automática de um modelo do fundo e na manutenção deste modelo, que é comparado quadro a quadro com a imagem atual [49]. Essa abordagem, apesar de ter sua aplicação voltada tradicionalmente aos sistemas de segurança [17], também é utilizada em métodos que funcionam como base para aplicações de substituição de fundo [29], [30], [50].

2) Movimentação do Elemento de Interesse: Muitos métodos que agem em ambientes não controlados têm como característica comum a busca por informações que permitem mapear a movimentação do elemento de interesse.

No entanto, uma das técnicas mais aplicadas para identificação de elementos em movimento em uma sequência de imagens – o cálculo do fluxo óptico (*optical flow*) [51] – é normalmente evitada devido ao seu custo computacional [10] e a impossibilidade de representar o elemento de interesse como um modelo rígido, dado que este, em grande parte das aplicações, é uma pessoa em primeiro plano [51].

Métodos de segmentação atuais identificam pixels em movimento utilizando informações de cor, aliada à observação da coerência temporal da sequência de imagens [10], [13], [31]. Processos de aprendizado *offline*, baseados em "*ground-truths*", também são recursos utilizados por métodos desenvolvidos a partir dessa abordagem. Obtém-se, desse modo, as probabilidades de cada pixel da imagem pertencer ao fundo ou ao elemento de interesse. Tais valores são considerados pelo modelo [10], [14], [31], no momento da classificação dos pixels.

Alguns métodos assumem que o plano de fundo seja estático e necessitam de inicialização na forma de um "plano de fundo limpo" [9], [29], para reduzir erros de classificação provocados por regiões de alto contraste no plano de fundo.

Em alguns trabalhos, as características do movimento são combinadas com informações a respeito da forma do elemento de interesse, para modelar correlações espaciais [11], [31]. Desse modo, pode-se classificar regiões da imagem pouco texturizadas, ou onde não houve movimentação (pixels dessas regiões não podem ser classificados com base apenas em informações de movimento, como mostrado na Fig. 9).

B. Métodos que utilizam Equipamento Específico ou Vídeo Binocular

Muitas abordagens utilizadas para extração do elemento de interesse a partir de uma sequência de imagens apóiam-se em equipamentos específicos, considerados não convencionais, ou utilizam mais de um equipamento (calibrados), com o

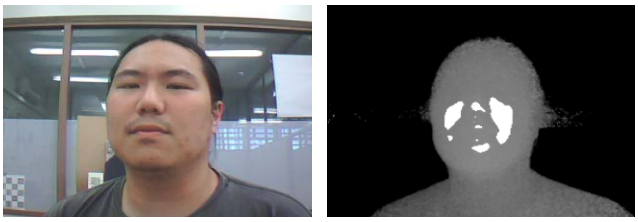
objetivo de obter novas informações que auxiliem a segmentação.

(a) (b)

Figura 9. Exemplo de um mapa de movimentação de pixels. As regiões mais claras da imagem correspondem às bordas em movimento, ao passo que as áreas mais escuras são bordas estacionárias. As regiões intermediárias representam áreas não texturizadas, que permanecem ambíguas (b) [10].

Apesar de sua utilização estar restrita a algumas aplicações atualmente, o desenvolvimento de esses métodos também se justifica pela possibilidade desses equipamentos tornarem-se convencionais no futuro.

Dois tipos de abordagens são comumente utilizadas para preencher esses mapas: estéreo e as baseadas em sensores. Ambas utilizam esses equipamentos com a finalidade de estimar mapas de profundidade da cena. Um mapa de profundidade é uma matriz, de tamanho correspondente ao da imagem, que contém a distância de cada pixel em relação à câmera. Na Fig. 10, mostra-se um mapa de profundidade da cena, que pode ser utilizado para auxiliar a segmentação.



(a) (b)

Figura 10. Mapa de Profundidade obtido por meio de um sensor TOF. Em (a) e (b), mostram-se o quadro de vídeo e o mapa de profundidade do mesmo quadro, respectivamente. Os pixels mais claros representam os mais próximos da câmera de vídeo (e do sensor), ao passo que os mais escuros são os mais distantes

1) Estéreo: Uma das formas de estimar mapas de profundidade para resolver problemas de segmentação é por meio da utilização de algoritmos de estéreo [54], [55].

A técnica do estéreo exige que dois vídeos sincronizados sejam utilizados como entrada. O principal desafio em abordagens desse tipo é a localização dos pixels correspondentes [56] nas imagens esquerda e direita, para que a profundidade de cada pixel possa ser calculada por meio de um processo de triangulação [54].

Uma estratégia adotada para encontrar correspondência nas imagens estéreo é determinar a linha epipolar, cujo processo se mostra na Fig. 11.

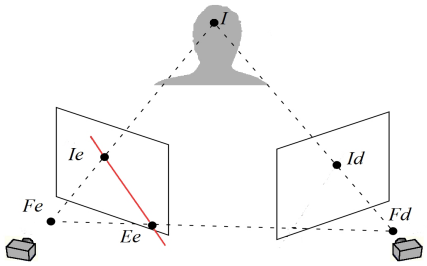


Figura 11. Determinação da linha epipolar. Um ponto *I*, pertencente ao elemento de interesse é observado por duas câmeras com seus respectivos pontos focais *Fe* e *Fd*. A projeção de *I* sobre os planos das imagens direita e esquerda são *Ie* e *Id*. A reta *IeEe* é a linha epipolar. O espaço de busca aos pontos correspondentes da imagem direita passa a ser restrito a essa reta. Como os pontos *Ie* e *Id*, e suas projeções são conhecidos, a distância do ponto *I* pode ser calculada por um processo de triangulação [54].

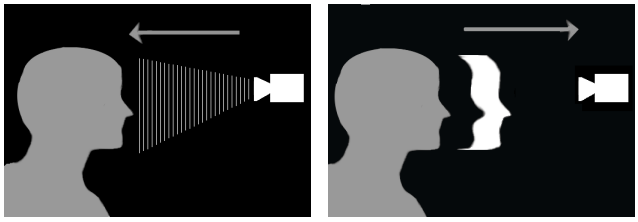
Apesar de utilizarem a distância dos pixels, obtida por meio de estéreo como informação principal, alguns trabalhos aplicam também cortes de cor e contraste [28], [57], [58] para evitar erros de classificação, principalmente nas bordas. Outros utilizam técnicas de reconhecimento de faces [59], para obter a localização do elemento de interesse e desconsiderar regiões dele distantes, tornando a segmentação mais robusta.

2) Sensores: TOF (*Time-of-Flight* – Tempo de Voo) são sensores ativos que utilizam laser para medir as distâncias entre o próprio sensor e os objetos da cena [60] (Fig. 12). Essas distâncias são utilizadas para preencher mapas de profundidades densos, utilizados em métodos de segmentação.

Basicamente, esses sensores utilizam luz pulsada [23], [24] ou luz modulada [61]. No primeiro caso, uma onda de luz constante acerta os elementos da cena, e a propagação de fótons de alta frequência mede o tempo de retorno do pulso de luz. No segundo caso, a luz emitida é modulada, e mede-se o TOF pela detecção do atraso da fase.

Equipamentos comerciais [52], que utilizam outros tipos de sensores, como os baseados em técnicas que utilizam luz estruturada [57] para a aquisição de mapas de profundidade, também têm sido utilizados para resolver problemas de segmentação em aplicações substituição de fundo.

Métodos que pertencem a esse grupo, além da informação de profundidade [23], aplicam também cortes de cor e contraste [12], ou atuam em conjunto com algoritmos de rastreamento [62], para alcançar resultados robustos. Definir um *threshold* simples, com base na distância do pixel não é suficiente, pois os valores de profundidade obtidos, na maioria das vezes, não são precisos a ponto de alcançar qualidade aceitável para aplicações de substituição de fundo [12].



(a) (b)

Figura 12. O conceito consiste em gerar uma "parede de luz", que se desloca ao longo do campo de visão da câmera. Essa parede pode ser gerada, por exemplo, como um pulso de laser de curta duração, com um campo de iluminação igual ao campo de visão da câmera (Fig. 12(a). Iluminação não-visual é utilizada para que não interfira no conteúdo visual do vídeo. Quando atinge os objetos na cena, a parede de luz é refletida de volta para a câmera, carregando uma impressão dos objetos (Fig. 12 (a)). A impressão contém todas as informações necessárias para a geração do mapa de profundidade [24].

VI. Desafios e soluções

O objetivo dos métodos de segmentação de vídeos em duas camadas que atuam em ambientes não controlados consiste na extração do elemento de interesse, sem que seja necessária a intervenção do usuário no ambiente onde a captura do vídeo se realiza. Isso significa que, além da dificuldade implícita de identificar o elemento a ser isolado em uma cena arbitrária, o algoritmo implementado no método deve tratar todas as situações desfavoráveis que podem ocorrer durante a execução de uma aplicação.

Nas subseções seguintes, expõem-se as principais situações-problema que devam ser contornadas em um processo de segmentação, seguidas das soluções empregadas pelos métodos atuais para evitar ou minimizar os erros de segmentação, provocados por tais situações.

A.Principais Problemas

Variações na iluminação, pessoas que atravessam o fundo da cena, ou a movimentação da câmera que captura o vídeo são situações comuns em ambientes não controlados. Ocorrências desse tipo são exemplos de situações desfavoráveis, que dificultam a identificação de um elemento de interesse dentro de uma sequência de imagens.

Algumas dessas situações, no entanto, podem se tornar um problema, quando se aplica determinada abordagem. Por outro lado, essa mesma situação é contornada implicitamente por métodos apoiados em outra.

Um exemplo disso é a situação em que uma pessoa atravessa o fundo da cena. Apesar de não representar um problema para métodos que utilizam mapas de profundidade – pois estes baseiam-se em informação de profundidade dos pixels –, é uma ocorrência difícil de ser contornada pelos métodos que utilizam informações de movimentação do elemento de interesse como meio de identificá-lo. Nesse caso, os pixels em movimento no fundo serão considerados como pertencentes ao elemento de interesse, caso nenhum tratamento adicional seja incorporado ao método.

Na tabela I, registram-se todas as situações-problema e suas possíveis causas, tais como identificadas nos trabalhos analisados, independentemente dos métodos que afetam.

Na tabela II, mostram-se as formas com que cada uma dessas situações-problema afeta os métodos de segmentação.

Importa ressaltar que se destacaram os problemas que podem ocorrer durante a execução da aplicação. A abordagem estéreo, por exemplo, exige um trabalhoso processo de calibração de duas (ou mais) câmeras [10]. Considera-se, neste trabalho, que tais dispositivos estejam devidamente calibrados.

Do mesmo modo, considera-se também que o problema da sincronização do sensor com a câmera de vídeo [60] esteja resolvido. A utilização de câmeras tanto binoculares quanto com sensores pré-calibradas, que podem ser encontradas no mercado, evitam problemas de calibração.

Tabela I
Problemas que podem ocorrer quando se utilizam métodos de segmentação em tempo real de sequência de imagens que atuam em ambientes não controlados e suas possíveis causas.

Problema	Possíveis Causas
Variações na iluminação	O acender ou o apagar de lâmpadas em um escritório [9], movimentação de pessoas próxima à câmera que podem provocar sombras ou acionar o ajuste automático de branco da câmera [9], [10].
Movimentação no fundo	Movimentos de cortinas, provocados por rajadas de vento [9]. Movimento de nuvens, ondas do mar, galhos e folhas de árvores [41]. Objetos ou pessoas distantes que atravessam a cena [11], [31]. Objetos que se movem até a cena e depois deixam de se movimentar ou objetos presentes na cena se afastam e revelam novas partes do fundo [9], [41].
Elemento de interesse estático	Indivíduo em primeiro plano permanece imóvel em frente a câmera [11].
Grande movimentação do elemento de interesse	O elemento de interesse se movimenta além do campo de visão da câmera [31], [51].
Oscilações da câmera	Tremulação da câmera acoplada em um computador móvel, posicionado no colo do usuário [9], [10], [11], [31].
Cores semelhantes no fundo e no elemento de interesse	Existência de objetos no plano de fundo que possuem a mesma tonalidade de parte do vestuário da pessoa em primeiro plano (Fig. 8(c)).
Regiões pouco texturizadas ou homogêneas	Imagens saturadas ou presença de elementos como paredes brancas e partes do céu [10], [28], [58].
Intensidade da luz do ambiente	Presença de superfícies reflexivas ou de vários sensores no ambiente [63].

B. Algumas Soluções Existentes

Posto que a segmentação voltada a aplicações em ambientes não controlados represente um desafio aos pesquisadores da área, para alguns dos problemas levantados há soluções eficientes. Por outro lado, várias são as situações-problema cujos erros provocados apenas se minimizam.

TABELA II
FORMAS COMO AS SITUAÇÕES-PROBLEMA PODEM AFETAR CADA ABORDAGEM EM UM PROCESSO DE SEGMENTAÇÃO REALIZADO EM TEMPO REAL, EM AMBIENTE NÃO CONTROLADO.

SITUAÇÃO/PROBLEMA	EQUIPAMENTO CONVENCIONAL		EQUIPAMENTO NÃO CONVENCIONAL	
	MOVIMENTAÇÃO	SUBTRAÇÃO DE FUNDO	ESTÉREO	SENSORES
Variações na iluminação	A mudança de cor de um pixel, provocada pela mudança de iluminação, pode ser confundida com a movimentação do elemento de interesse, produzindo erros de classificação [10].	Podem tornar as cores do quadro atual bem diferentes em relação àquelas do modelo do fundo [9].	–	–
Movimentação no fundo	O objeto, ou pessoa, que atravessa o fundo da cena se considera elemento de interesse [10], [11].	O objeto, ou pessoa, que atravessa o fundo da cena se considera elemento de interesse [9], [10], [11], [31].	–	–
Elemento de interesse estático	Impossibilita a distinção entre o que é fundo e o que é elemento de interesse, dado que a sequência de quadros é estática (utilizando apenas informação de movimentação de pixel) [11].	Pode impossibilitar a geração do modelo do fundo [41] (quando não existe inicialização na forma de um "plano de fundo limpo").	–	–
Grande movimentação do elemento de interesse	Pode causar o "problema da abertura" quando se utiliza o fluxo óptico (a posição do elemento rastreado é perdida em alguns quadros do vídeo e não é possível determinar o seu deslocamento) [51].	–	Pode causar oclusão estéreo [28], [58], [64] (o elemento de interesse não fica visível em uma das câmeras, impossibilitando a localização de pixels correspondentes nas imagens).	O elemento de interesse pode se mover além do limite de emissão de luz do sensor [60], impossibilitando o cálculo de profundidade.
Oscilações da câmera	Impede a diferenciação entre o que é movimento do elemento de interesse e o que são alterações de cores provocadas pela movimentação da câmera (uma mesma região da imagem passa a não corresponder mais à mesma região da cena)[9], [10].	Faz com que a imagem de referência não represente o plano de fundo nos quadros em que a posição da câmera é diferente da inicial [41].	Pode afetar a calibração das câmeras (caso uma delas seja movimentada). Não representa um problema quando se utiliza um equipamento binocular pré-calibrado.	–
Cores semelhantes no fundo e no elemento de interesse	Pode fazer com que as regiões das bordas do elemento de interesse sejam consideradas fundo, devido à ausência de contraste [9] (o contraste é um informação utilizada para identificação de pixels em movimento).	Pode fazer com que os pixels do fundo sejam confundidos com os do elemento de interesse, provocando erros de classificação [9].	–	–
Regiões pouco texturizadas ou homogêneas	Impossibilita a identificação de pixels em movimento, quando a movimentação desse pixel fica limitada a regiões desse tipo [10].	–	Impede a identificação dos pixels correspondentes nas duas imagens de entrada, o que é a tarefa essencial para esse tipo de abordagem [10], [28], [58].	–
Intensidade da luz do ambiente	–	–	–	Pode provocar múltiplas reflexões, causando interferências nos sinais de retorno, que são utilizados para o cálculo dos valores de profundidade dos pixels [63].

Entre as abordagens apoiadas em vídeo monocular, os métodos baseados em subtração de fundo que utilizam informações espaciais, por exemplo, vieram para solucionar muitos dos problemas que ocorriam em métodos mais simplificados, normalmente baseados em *thresholds*. Variações na iluminação, por exemplo, desde que ocorram dentro de determinados níveis, podem ser contornadas por esses métodos [29]. Quando essas variações ocorrem de forma brusca, no entanto, o problema é de difícil tratamento.

A ocorrência de grande movimentação do fundo não pode ser tratada por métodos puros de subtração de fundo. Nesse caso, outras informações obtidas da imagem são necessárias. Quando existe pequena movimentação, os erros de classificação provocados são em pequeno número e os mesmos podem ser preenchidos (quando ocorrem no

elemento de interesse), ou removidos (quando ocorrem no fundo) aplicando-se operadores morfológicos [65].

Em alguns casos, o problema do elemento de interesse estático, o que dificulta a construção automática do modelo do fundo, tem sido contornado por meio da inicialização do sistema com uma imagem "limpa" do plano de fundo – ou seja, excluindo-se o elemento de interesse [29], ou pela captação de um conjunto de imagens do fundo [30], [65].

Alguns métodos baseados em subtração de fundo tratam as oscilações na câmera por meio da utilização de um modelo de plano de fundo estendido, que contém regiões além do tamanho da janela do vídeo. Para contornar o problema das cores semelhantes no fundo e no elemento de interesse, faz-se necessário o processamento em conjunto com outras técnicas.

Com respeito aos métodos apoiados na movimentação do elemento de interesse, verifica-se que o problema de sua grande movimentação tem sido de pouca ocorrência, dado que as soluções mais recentes não se apoiam em cálculos do fluxo óptico [10], [11], [31]. A utilização de outras informações, como cor e contraste, tem sido a solução para que regiões pouco texturizadas, ou homogêneas, possam ser classificadas, quando o elemento de interesse é estacionário [10].

Oscilações na câmera e variações na iluminação são problemas que têm sido minimizados, utilizando-se informações obtidas da coerência temporal do vídeo, e com o auxílio de treinamento *offline* [9], [10], [28]. Em alguns casos, essas informações combinam-se com filtros de forma [11], para estimar a geometria do elemento de interesse e minimizar os problemas ocasionados por movimentação no fundo, além de reduzir ainda mais os provocados por oscilações na câmera e variações na iluminação [11], [31].

A utilização do modelo de movimentação em conjunto com outras técnicas, como a de subtração de fundo, pode evitar o problema das cores semelhantes no fundo e no elemento de interesse [9].

Entre as abordagens apoiadas em equipamento específico, ou em vídeo binocular, os métodos de estéreo, em que a informação de profundidade é a única, não tratam o problema da oclusão estéreo. Em alguns métodos mais sofisticados, adotam-se informações de cor, contraste e a coerência espacial entre os quadros, para evitar o problema [60].

A inclusão dessas informações evita o problema da impossibilidade de classificação nas regiões pouco texturizadas. A aplicação conjunta de algoritmos de identificação de faces também é uma solução para os problemas das regiões poucos texturizadas, ou de oscilações da câmera [59].

Em técnicas que utilizam sensores, a movimentação do elemento de interesse além do raio de ação do sensor não pode ser tratada quando a informação de profundidade é a única considerada pelo método [23], [24]. Para que se obtenha qualidade no recorte, outras informações, como cor e contraste, são necessárias [12]. O problema das múltiplas reflexões não foi abordado em nenhum dos trabalhos analisados.

VII. Conclusões

A pesquisa realizada mostra que são inúmeros os problemas que devem ser tratados pelos métodos de segmentação em tempo real de vídeos em duas camadas que atuam em ambientes não controlados. Apesar do considerável volume de pesquisas na área, ainda não existe uma solução geral. Se, por um lado, alguns métodos atuais se apresentam sofisticados, por outro, não tratam satisfatoriamente todos os problemas aqui abordados. Como consequência, as soluções existentes ainda são propensas a erros.

Além disso, ressalta-se um fator a observar: uma mesma situação pode representar um problema para um método desenvolvido por meio de determinada abordagem e ser facilmente tratada por outro originado a partir de abordagem diferente. Justifica-se a observação pelo fato de a maioria dos métodos não se utilizar uma única informação (cor ou movimento, por exemplo), mas basearem-se em várias outras que se podem obter da sequência de imagens, ou por meio de equipamento específico. No entanto, alguns trabalhos buscam soluções valendo-se de extensões de uma técnica original, em vez de utilizar outra como auxiliar.

Abordagens que se apoiam em informações adquiridas por equipamento específico (ou exigem do usuário algum tipo de calibração de equipamentos) produzem métodos que apresentam, atualmente, os resultados mais robustos. Como mostrado na tabela II, menos situações-problema precisam ser contornadas nesses casos. No entanto, existe um volume considerável de pesquisas que buscam resultados semelhantes, utilizando apenas equipamento convencional (vídeo monocular), com o objetivo de produzir métodos que possam ser utilizados em número maior de aplicações.

Para que se procedesse à classificação dos métodos analisados, conforme aqui sugerida, atentou-se para a utilização ou não desses equipamentos como principal critério de agrupamento. Podem-se classificar as abordagens mais utilizadas considerando que fazem parte de uma dessas duas principais vertentes de pesquisas: a apoiada em vídeo monocular ou em equipamento específico de auxílio, incluindo vídeo binocular.

Embora apresentem mais problemas a serem contornados, as abordagens baseadas em movimentação dos pixels revelam-se uma tendência em aplicações de substituição de fundo. Os métodos apoiados nas técnicas de subtração de fundo têm suas aplicações potenciais nos sistema de segurança e são menos utilizados no contexto da substituição do fundo.

De modo geral, na quase totalidade dos trabalhos – objeto de análise do estado-da-arte – evidenciou-se uma característica relevante nos mais recentes: o fato de a

segmentação ser tratada como um problema de minimização de energia. Essa tendência deve-se a esse tipo de abordagem produzir métodos computáveis em tempo real e com resultados mais robustos que os obtidos por outras abordagens. Em regra, combinam-se probabilisticamente as várias informações úteis à segmentação, tanto as que podem ser obtidas da imagem quanto as fornecidas por equipamento específico.

AGRADECIMENTOS

Silvio Sanches agradece à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e Valdinei Silva agradece à FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo – proc. 11/19280-8) pelos respectivos apoios financeiros.

REFERÊNCIAS

[1] F. D. Williams, "Method of taking motion pictures," U.S. Patent 1,273,435, Jul. 23, 1918.

[2] J. Foster, The Green Screen Handbook: Real-World Production Techniques. Chichester, GB: John Wiley and Sons Ltd, 2010, ch. Mattes and Compositing Defined, pp. 3–15.

[3] P. Vlahos, "Composite photography utilizing sodium vapor illumination," U.S. Patent 3,095,304, Jun. 25, 1963.

[4] ____, "Composite color photography," U.S. Patent 3,158,477, Nov. 24, 1964.

[5] ____, "Comprehensive electronic compositing system," U.S. Patent 4,100,569, Jul. 11, 1978.

[6] Y. Mishima, "Soft edge chroma-key generation based upon hexoctahedral color space," U.S. Patent 5,355,174, Out. 11, 1994, 11-10-1994.

[7] S. Gibbs, C. Arapis, C. Breiteneder, V. Lalioti, S. Mostafawy, and J. Speier, "Virtual studios: an overview," *Multimedia, IEEE*, vol. 5, no. 1, pp. 18–35, Jan-Mar 1998.

[8] J. Bergen, P. Burt, R. Hingorani, and S. Peleg, "A three-frame algorithm for estimating two-component image motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 9, pp. 886–896, 1992.

[9] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, "Background cut," *in Proceedings of European Conference Computer Vision (ECCV 2006)*, 2006, vol. 2, pp. 628–641.

[10] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, "Bilayer segmentation of live video," in *CVPR '06: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. Washington, DC, USA: IEEE Computer Society, June 2006, pp. 53–60.

[11] P. Yin, A. Criminisi, J. Winn, and I. Essa, "Bilayer segmentation of webcam videos using tree-based classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 1, pp. 30–42, 2011.

[12] L. Wang, C. Zhang, R. Yang, and C. Zhang, "Tofcut: Towards robust real-time foreground extraction using a time-of-flight camera," in *Fifth International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010.

[13] A. Parolin, G. P. Fickel, C. R. Jung, T. Malzbender, R. Samadani, "Bilayer video segmentation for videoconferencing applications", in *IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1-6.

[14] S. R. R. Sanches, V. F. Silva and R. Tori, "Bilayer segmentation augmented with future evidence", in *Computational Science and Its Applications - ICCSA 2012*, ser. Lecture Notes in Computer Science, B. Murgante, O. Gervasi, S. Misra, N. Nedjah, A. Rocha, D. Taniar, and B. Apduhan, Eds. Springer Berlin / Heidelberg, 2012, vol. 7334, pp. 699–711.

[15] S. R. R. Sanches, D. M. Tokunaga, V. F. Silva, A. C. Sementille and R. Tori, "Mutual occlusion between real and virtual elements in augmented reality based on fiducial markers", in *IEEE Workshop on Applications of Computer Vision (WACV)*, 2012, pp. 49-54.

[16] Z. Wu and C. Chen, "A new foreground extraction scheme for video streams," in *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*. New York, NY, USA: ACM, 2001, pp. 552–554.

[17] W. Nam and J. Han, "Motion-based background modeling for foreground segmentation," in *VSSN '06: Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*. New York, NY, USA: ACM, 2006, pp. 35–44.

[18] J. Bernardes, R. Nakamura and R.Tori, "Comprehensive Model and Image-Based Recognition of Hand Gestures for Interaction in 3D Environments" *The International Journal of Virtual Reality*, vol. 10, pp. 11-23, 2011.

[19] J. L. A. Samatelo and E. O. T. Salles, "A New Change Detection Algorithm for Visual Surveillance System," *IEEE Latin America Transactions*, vol. 10, no. 1, pp. 1221-1226, 2012.

[20] J. Wang and M. F. Cohen, "Image and video matting: a survey," *Found. Trends. Comput. Graph. Vis*., vol. 3, no. 2, pp. 97–175, 2007.

[21] C. Rother, V. Kolmogorov, and A. Blake, ''Grabcut: interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph*., vol. 23, no. 3, pp. 309–314, 2004.

[22] Y.-Y. Chuang, "New models and methods for matting and compositing," Ph.D. dissertation, University of Washington, 2004.

[23] G. J. Iddan and G. Yahav, "Three-dimensional imaging in the studio and elsewhere," B. D. Corner, J. H. Nurre, and R. P. Pargas, Eds., vol. 4298, no. 1. Bellingham, Washington USA: *Society of Photo-Optical Instrumentation Engineers (SPIE)*, 2001, pp. 48–55.

[24] R. Gvili, A. Kaplan, E. Ofek, and G. Yahav, "Depth keying," *SPIE Elec. Imaging*, vol. 5006, pp. 554–563, 2003.

[25] S. Wang, X. Xiong, Y. Xu, C. Wang, W. Zhang, X. Dai, and D. Zhang, "Face-tracking as an augmented input in video games: enhancing presence, role-playing and control," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ser. CHI '06. New York, NY, USA: ACM, 2006, pp. 1097–1106.

[26] T. Ogi, T. Yamada, Y. Kurita, Y. Y. Hattori, and M. Hirose, "Usage of video avatar technology for immersive communication," in *ACL 2003 Co-located Workshop: First International Workshop on Language Understanding and Agents for Real World Interaction*, 2003.

[27] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-based visual hulls," *in SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 369–374.

[28] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother, "Bi-layer segmentation of binocular stereo video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. Washington, DC, USA: IEEE Computer Society, June 2005, pp. 407–414.

[29] C. Harrison and S. E. Hudson, "Pseudo-3d video conferencing with a generic webcam," in *Proceedings of the Tenth IEEE International Symposium on Multimedia* - ISM '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 236–241.

[30] J. H. Kim, S. C. Ahn, and H.-G. Kim, "Teleconference system with a shared working space and face mouse interaction," in *PCM (2)*. Berlin Heidelberg: Springer-Verlag, 2004, pp. 665–671.

[31] P. Yin, A. Criminisi, J. Winn, and I. Essa, "Tree-based classifiers for bilayer video segmentation," in *Computer Vision and Pattern Recognition. CVPR '07*. IEEE Conference on, vol. 0. Los Alamitos, CA, USA: IEEE Computer Society, June 2007, pp. 1–8.

[32] Q. Wu, P. Boulanger, and W. F. Bischof, "Robust real-time bi-layer video segmentation using infrared video," in *Proceedings of the Canadian Conference on Computer and Robot Vision*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 87–94.

[33] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall, Inc., 2002.

[34] H. Pedrini and W. R. Schwartz, *Análise de Imagens Digitais: Princípios, Algoritmos e Aplicações*, 1st ed. Thomson Learning, 2008.

[35] T. Porter and T. Duff, "Compositing digital images," in *SIGGRAPH '84: Proceedings of the 11th annual conference on Computer*

graphics and interactive techniques. New York, NY, USA: ACM Press, 1984, pp. 253–259.

[36] E. Mortensen and W. Barrett, "Toboggan-based intelligent scissors with a four-parameter edge model," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on*., vol. 2, 1999, pp. 2 vol. (xxiii+637+663).

[37] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society*, vol. 51, no. 2, pp. 271–279, 1989.

[38] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, 2004.

[39] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 2, pp. 147 –159, feb. 2004.

[40] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," in *Computer Vision, IEEE International Conference on*, vol. 1. Los Alamitos, CA, USA: IEEE Computer Society, 2001, pp. 105–112.

[41] M. Piccardi, "Background subtraction techniques: a review," *in Systems, Man and Cybernetics*, 2004 IEEE International Conference on, vol. 4, 2004, pp. 3099 – 3104 vol.4.

[42] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," in *Proc. 13th Conf. Uncertainty in Artificial Intelligence (UAI)*, 1997, pp. 175–181.

[43] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 10, pp. 1337 – 1342, 2003.

[44] Z. Tang, Z. Miao, and Y. Wan, "Background subtraction using running gaussian average and frame difference," in *Entertainment Computing - ICEC 2007*, ser. Lecture Notes in Computer Science, L. Ma, M. Rauterberg, and R. Nakatsu, Eds. Springer Berlin / Heidelberg, 2007, vol. 4740, pp. 411–414.

[45] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proceedings of the 6th European Conference on Computer Vision-Part II*, ser. ECCV '00. London, UK: Springer-Verlag, 2000, pp. 751–767.

[46] B. Han, D. Comaniciu, and L. Davis, "Sequential kernel density approximation through mode propagation:," in *Proceedings of Asian Conference on Computer Vision – ACCV 2004*, 2004.

[47] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 831 –843, Aug. 2000.

[48] R. Nakamura, "Vídeo-avatar com detecção de colisão para realidade aumentada e jogos," Ph.D. dissertation, Escola Politécnica da Universidade de São Paulo, 2008.

[49] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," *Computer Vision, IEEE International Conference on*, vol. 1, p. 255, 1999.

[50] R. Qian and M. Sezan, "Video background replacement without a blue screen," *Image Processing. ICIP 99. Proceedings International Conference on*, vol. 4, pp. 143–146 vol.4, 1999.

[51] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, pp. 43–77, 1994.

[52] R. M. Geiss, " Visual targed tracking," U.S. Patent 2010/197 399 A1, Aug. 5, 2010.

[53] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 1, p. 195, 2003.

[54] Y. Ohta and T. Kanade, "Stereo by intra- and inter-scanline search using dynamic programming," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 7, no. 1, pp. 139–154, March 1985.

[55] I. J. Cox, S. L. Hingorani, S. B. Rao, and B. M. Maggs, "A maximum likelihood stereo algorithm," *Comput. Vis. Image Underst.*, vol. 63, no. 3, pp. 542–567, 1996.

[56] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, pp. 7–42, April 2002.

[57] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother, "Probabilistic fusion of stereo with color and contrast for bi-layer segmentation," Microsoft Research, Cambridge, Tech. Rep. MSR-TR-2005-36, March 2005. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=70156.

[58] ____, "Probabilistic fusion of stereo with color and contrast for bilayer segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 9, pp. 1480–1492, Sept. 2006.

[59] K. Law and S. Sclaroff, "Foreground object segmentation from binocular stereo video," Intelligent *Robots and Computer Vision XXIII: Algorithms, Techniques, and Active Vision*, vol. 6006, no. 1, p. 60060C, 2005.

[60] L. Bianchi, P. Dondi, R. Gatti, L. Lombardi, and P. Lombardi, "Evaluation of a foreground segmentation algorithm for 3d camera sensors," *in Image Analysis and Processing - ICIAP 2009*, ser. Lecture Notes in Computer Science, P. Foggia, C. Sansone, and M. Vento, Eds. Springer Berlin / Heidelberg, 2009, vol. 5716, pp. 797–806.

[61] S. B. Gokturk, H. Yalcin, and C. Bamji, "A time-of-flight depth sensor - system description, issues and solutions," in *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, vol. 3. Washington, DC, USA: IEEE Computer Society, 2004, pp. 35–.

[62] A. Bleiweiss and M. Werman, "Fusing time-of-flight depth and color for real-time segmentation and tracking," *in Proceedings of the DAGM Workshop on Dynamic 3D Imaging*, ser. Dyn3D '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 58–69.

[63] A. Kolb, E. Barth, and R. Koch, "Tof-sensors: New dimensions for realism and interactivity," in *Computer Vision and Pattern Recognition Workshops. CVPRW '08. IEEE Computer Society Conference on*, 2008, pp. 1 - 6.

[64] D. Geiger, B. Ladendorf, and A. Yuille, "Occlusions and binocular stereo," *International Journal of Computer Vision*, vol. 14, pp. 211–226, 1995.

[65] M. Li, "Towards real-time novel view synthesis using visual hulls," Ph.D. dissertation, Universität des Saarlandes, 2005.

Silvio Ricardo Rodrigues Sanches é bacharel (2003) e mestre (2007) em Ciência da Computação pelo Centro Universitário Euripedes de Marília. Atualmente é aluno regular do curso de Doutorado em Engenharia Elétrica (Sistemas Digitais) da Escola Politécnica da Universidade de São Paulo. Tem experiência nas áreas de Realidade Aumentada e Visão Computacional.

Ricardo Nakamura é engenheiro mecânico (1999), mestre (2002) e doutor (2008) em Engenharia Elétrica pela Universidade de São Paulo. Sua principal área de pesquisa é a de interação, particularmente em jogos digitais. Atualmente é professor doutor na Escola Politécnica da USP e integra o comitê gestor da Comissão Especial de Jogos e Entretenimento Digital da Sociedade Brasileira de Computação (SBC).

Valdinei Freire da Silva graduou-se em Engenharia da Computação pela Universidade de São Paulo (2002) e doutorou-se em co-tutela de Engenharia Elétrica pela Universidade de São Paulo e Engenharia Electrotécnica e de Computadores pela Universidade Técnica de Lisboa (2009). Realiza pesquisa na área de Inteligência Artificial, atuando principalmente nos seguintes temas: processos markovianos de decisão, aprendizado por reforço, extração de preferências e teoria da utilidade esperada. Atua desde 2010 como professor doutor na Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP).

Romero Tori é engenheiro (1982), mestre (1988), doutor (1994) e livre-docente (2003) pela Universidade de São Paulo. Atualmente é professor titular do Centro Universitário Senac de São Paulo e Professor Associado da Escola Politécnica da USP. Os termos mais frequentes na contextualização de sua produção científica, tecnológica e artístico-cultural são: Computação Gráfica, Design, Realidade Virtual, Realidade Aumentada, Multimídia, Hipermídia, Educação Virtual Interativa, Educação, Tecnologia Educacional, Educação a Distância e *Computer Games*.