República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) **PI0620497-0 A2**

BRPI0620497A2

(22) Data de Depósito: 15/11/2006
(43) Data da Publicação: 16/11/2011
(RPI 2132)

(51) Int.Cl.:
G06F 17/30

(54) Título: MÉTODO PARA A CRIAÇÃO DE UMA SINOPSE DE VÍDEO, SISTEMA PARA TRANSFORMAR UMA SEQUÊNCIA DE ORIGEM DE QUADROS DE VÍDEO DE UMA PRIMEIRA CENA DINÂMICA EM UMA SEQUÊNCIA DE SINOPSE DE PELO MENOS DOIS QUADROS DE VÍDEO QUE ILUSTRAM UMA SEGUNDA CENA DINÂMICA, E, PRODUTO DE PROGRAMA DE COMPUTADOR

(30) Prioridade Unionista: 15/11/2005 US 60/736,313, 17/01/2006 US 60/759,044

(73) Titular(es): Yissum Research Development Company of the Hebrew University of Jerusalem

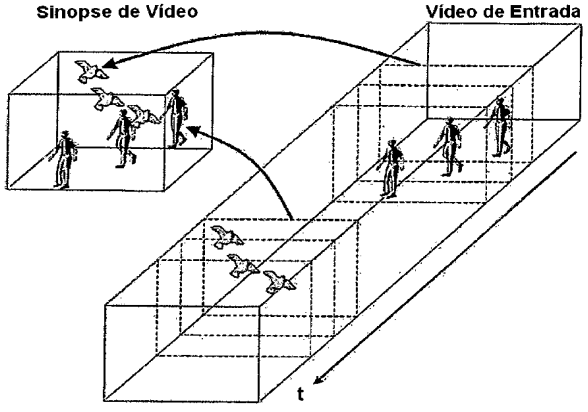
(72) Inventor(es): Alexander Rav-Acha, Shmuel Peleg

(74) Procurador(es): David do Nascimento Advogados Associados

(86) Pedido Internacional: PCT IL06001320 de 15/11/2006

(87) Publicação Internacional: WO 2007/057893de 24/05/2007

(57) Resumo: MÉTODO PARA A CRIAÇÃO DE UMA SINÓPSE DE VÍDEO, SISTEMA PARA TRANSFORMAR UMA SEQUÊNCIA DE ORIGEM DE QUADROS DE VÍDEO DE UMA PRIMEIRA CENA DINÂMICA EM UMA SEQUÊNCIA DE SINOPSE DE PELO MENOS DOIS QUADROS DE VÍDEO QUE ILUSTRAM UMA SEGUNDA CENA DINÂMICA, E, PRODUTO DE PROGRAMA DE COMPUTADOR. Trata-se de um método e um sistema implementado por computador que transformam uma primeira seqüência de quadros de vídeo de uma primeira cena dinâmica em uma segunda seqUência de pelo menos dois quadros de vídeo que apresentam uma segunda cena dinâmica. Um subconjunto de quadros de vídeo na primeira seqUência é obtido, o qual mostra o movimento de pelo menos um objeto que tem uma pluralidade de pixeis localizados nas respectivas coordenadas x,y e são selecionadas partes do subconj unto que mostram os aparecimentos não-espacialmente sobrepostos de pelo menos um objeto na primeira cena dinâmica. As partes são copiadas de pelo menos três quadros de entrada diferentes em pelo menos dois quadros sucessivos da segunda sequência sem mudar as respectivas coordenadas x,y dos pixels no objeto e de maneira tal que pelo menos um dos quadros da segunda seqUência contenha pelo menos duas partes que aparecem em quadros diferentes na primeira sequência.



Ref. 1820804590
PI 0620497-0

MÉTODO PARA A CRIAÇÃO DE UMA SINOPSE DE VÍDEO, SISTEMA PARA TRANSFORMAR UMA SEQÜÊNCIA DE ORIGEM DE QUADROS DE VÍDEO DE UMA PRIMEIRA CENA DINÂMICA EM UMA SEQÜÊNCIA DE SINOPSE DE PELO MENOS DOIS QUADROS DE VÍDEO QUE ILUSTRAM UMA SEGUNDA CENA DINÂMICA E PRODUTO DE PROGRAMA DE COMPUTADOR

CAMPO DA INVENÇÃO

A presente invenção refere-se de maneira geral à renderização baseada em imagem e vídeo, onde novas imagens e os vídeos são criados mediante a combinação de partes de múltiplas imagens originais de uma cena. Particularmente, a invenção refere-se a tal técnica para a finalidade de abstração ou sinopse de vídeo.

TÉCNICA ANTERIOR

As referências da técnica anterior consideradas como relevantes como um antecedente para a invenção são listadas abaixo e seus conteúdos são aqui incorporados a título de referência. Referências adicionais são mencionadas nos pedidos de patente norte-americanos provisórios números 60/736.313 e 60/759.044 e seus teores são aqui incorporados a título de referência. O reconhecimento das referências aqui apresentadas não deve inferido como significando que elas são de alguma maneira relevantes à Patenteabilidade da invenção aqui descrita. Cada referência é identificada por um número incluído nos colchetes e, por conseguinte a técnica anterior será mencionada por todo o relatório descritivo pelos números incluídos nos colchetes.

[1] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. *Interactive digital photomontage*. In SIGGRAPH, páginas 294-302, 2004.

[2] A. Agarwala, K. C. Zheng, C. Pal, M. Agrawala, M. Cohen, B. Curless, D. Salesin, and R. Szeliski. *Panoramic video textures*. In SIGGRAPH, páginas 821-827, 2005.

[3] J. Assa, Y. Caspi, and D. Cohen-Or. Action sinopse: *Pose*

selection and illustration. In SIGGRAPH, páginas 667-676, 2005.

[4] O. Boiman and M. Irani. *Detecting irregularities in images and in video*. In ICCV, páginas I: 462-469, Beijing, 2005.

[5] A. M. Ferman and A. M. Tekalp. *Multiscale content extraction and representation for video indexing*. Proc. of SPIE, 3229:23-31, 1997.

[6] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu. *Efficient representations of video sequences and their applications*. Signal Processing: Image Communication, 8(4):327-351, 1996.

[7] C. Kim and J. Hwang. *An integrated scheme for object-based video abstraction*. In ACM Multimedia, páginas 303-311, New York, 2000.

[8] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. *Optimization by simulated annealing*. Science, 4598(13):671-680, 1983.

[9] V. Kolmogorov and R. Zabih. *What energy functions can be minimized via graph cuts?* In ECCV, páginas 65-81, 2002.

[10] Y. Li, T. Zhang, and D. Tretter. *An overview of video abstraction techniques*. Technical Report HPL-2001-191, HP Laboratory, 2001.

[11] J. Oh, Q. Wen, J. lee, and S. Hwang. *Video abstraction*. In S. Deb, editor, Video Data Mangement and Information Retrieval, páginas 321-346. Idea Group Inc. and IRM Press, 2004.

[12] C. Pal and N. Jojic. *Interactive montages of sprites for indexing and summarizing security video*. In Vídeo Proceedings of CVPRO5, página II: 1192, 2005.

[13] A. Pope, R. Kumar, H. Sawhney, and C. Wan. Video abstraction: *Summarizing video content for retrieval and visualization*. In Signals, Systems and Computers, páginas

915-919, 1998.

[14] W02006/048875 *Method and system for spatio-temporal video warping*, pub. May 11, 2006 by S. Peleg, A. Rav-Acha and D. Lischinski. Este corresponde ao USSN 10/556,601 depositado em 02 de novembro de 2005.

[15] A. M. Smith and T. Kanade. *Video skimming and characterization through the combination of image and language understanding*. In CAIVD, páginas 61-70, 1998.

[16] A. Stefanidis, P. Partsinevelos, P. Agouris, and P. Doucette. *Summarizing video datasets in the spatiotemporal domain*. In DEXA Workshop, páginas 906-912, 2000.

[17] H. Zhong, J. Shi, and M. Visontai. *Detecting unusual activity in video*. In CVPR, páginas 819-826, 2004.

[18] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid, and W. G. Aref. *Exploring video content structure for hierarchical summarization*. Multimedia Syst., 10(2):98-115, 2004.

[19] J. Barron, D. Fleet, S. Beauchemin and T. Burkitt. *Performance of optical flow techniques*, volume 92, páginas 236-242.

[20] V. Kwatra, A. Schodl, I. Essa, G. Turk and A. Bobick. *Graphcut textures: image and video synthesis using graph cuts*. In SIGGRAPH, páginas 227-286, Julho de 2003.

[21] C. Kim and J. Hwang, *Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications*, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 2, Fevereiro de 2002, páginas 122-129.

[22] Patente U.S. N°. 6.665.003.

ANTECEDENTES DA INVENÇÃO

A sinopse de vídeo (ou abstração) é uma representação temporalmente compacta que visa a habilitação de busca e recuperação de vídeo.

Há duas abordagens principais para a sinopse de

vídeo. Em uma abordagem, um conjunto de imagens salientes (quadros chaves) é selecionado da seqüência de vídeo original. Os quadros chaves que são selecionados são aqueles que melhor representam o vídeo [7, 18]. Em uma outra abordagem, uma coleção de seqüências de vídeo curtas é selecionada [15]. A segunda abordagem é menos compacta, mas confere uma melhor impressão da dinâmica da cena. Essas abordagens (e outras) são descritas em buscas amplas na abstração de vídeo [10, 11].

Em ambas as abordagens acima, quadros inteiros são utilizados como os blocos de edificação fundamentais. Uma metodologia diferente utiliza imagens de mosaico junto com alguns meta-dados para a indexação de vídeo [6, 13, 12]. Nessa metodologia, a imagem de sinopse estática inclui objetos de tempos diferentes.

Também são conhecidas abordagens baseadas em objetos nas quais os objetos são extraídos do vídeo de entrada [7, 5, 16]. No entanto, esses métodos utilizam a detecção do objeto para identificar quadros chaves significativos e não combinam as atividades de intervalos de tempo diferentes.

No estado da técnica, também são conhecidos métodos para a criação de uma imagem panorâmica simples utilizando mini-cortes iterados [1] e para a criação de um filme panorâmico utilizando mini-cortes [2]. Em ambos os métodos, um problema com complexidade exponencial (no número de quadros de entrada) é aproximado e, portanto, eles são mais apropriados para um número pequeno de quadros. O trabalho relacionado neste campo é associado com a combinação de dois filmes utilizando mini-cortes [20].

O Pedido de Patente WO2006/048875 [14] apresenta um método e um sistema para manipular o fluxo temporal em um vídeo. Uma primeira seqüência de quadros de vídeo de uma

primeira cena dinâmica é transformada em uma segunda seqüência de quadros de vídeo que descrevem uma segunda cena dinâmica tal que, em um aspecto, para pelo menos uma característica na primeira cena dinâmica, as respectivas partes da primeira seqüência de quadros de vídeo são amostradas a uma taxa diferente do que as partes circundantes da primeira seqüência de quadros de vídeo; e as partes amostradas são copiadas em um quadro correspondente da segunda seqüência. Isso permite que a sincronia temporal das características em uma cena dinâmica seja mudada.

DESCRIÇÃO RESUMIDA DA INVENÇÃO

De acordo com um primeiro aspecto da invenção, é apresentado um método implementado em computador para a criação de uma sinopse de vídeo a partir da transformação de uma fonte de seqüências de quadros de vídeo de uma primeira cena dinâmica capturada em intervalos de tempo regular em uma seqüência de sinopse de quadros de vídeo mais curta, que descreve uma segunda cena dinâmica, em que o método compreende:

(a) a obtenção de um subconjunto de quadros de vídeo na dita primeira seqüência que mostram o movimento de pelo menos um objeto que compreende uma pluralidade de pixels localizados nas respectivas coordenadas x,y;

(b) a seleção das ditas partes do subconjunto que mostram aparecimentos não-espacialmente sobrepostos de pelo menos um objeto em cada quadro de vídeo; e

(c) a cópia das ditas partes de pelo menos três quadros de entrada diferentes em pelo menos dois quadros sucessivos da segunda seqüência sem mudar as respectivas coordenadas x,y dos pixels no dito objeto, e de maneira tal que pelo menos um dos quadros da segunda seqüência contenha pelo menos duas partes que aparecem em quadros diferentes na primeira seqüência.

De acordo com um segundo aspecto da invenção, é apresentado um sistema para transformar uma primeira seqüência de quadros de vídeo de uma primeira cena dinâmica em uma segunda seqüência de pelo menos dois quadros de vídeo que descrevem uma segunda cena dinâmica, em que o sistema compreende:

uma primeira memória para armazenar um subconjunto de quadros de vídeo na dita primeira seqüência que mostram o movimento de pelo menos um objeto que compreende uma pluralidade de pixels localizados nas respectivas coordenadas *x,y*,

uma unidade da seleção acoplada à primeira memória para selecionar das ditas partes do subconjunto que mostram aparecimentos não-espacialmente sobrepostos de pelo menos um objeto em cada quadro de vídeo,

um gerador de quadros para copiar as ditas partes de pelo menos três quadros de entrada diferentes em pelo menos dois quadros sucessivos da segunda seqüência sem mudar as respectivas coordenadas *x,y* dos pixels no dito objeto e de maneira tal que pelo menos um dos quadros da segunda seqüência contenha pelo menos duas partes que aparecem em quadros diferentes na primeira seqüência, e

uma segunda memória para armazenar os quadros da segunda seqüência.

A invenção compreende adicionalmente, de acordo com um terceiro aspecto, um portador de dados que incorpora tangivelmente uma seqüência de quadros de vídeo de saída que ilustram uma cena dinâmica, em que pelo menos dois quadros sucessivos dos ditos quadros de vídeo de saída compreendem uma pluralidade de pixels que têm as respectivas coordenadas *x,y* e são derivados de partes de um objeto de pelo menos três quadros de entrada diferentes sem mudar as respectivas coordenadas *x,y* dos pixels no dito objeto e de maneira tal

que pelo menos um dos quadros de vídeo de saída contenha pelo menos duas partes que aparecem em quadros de entrada diferentes.

A sinopse de vídeo dinâmica apresentada pela presente invenção é diferente das abordagens de abstração de vídeo precedentes revistas acima nas duas seguintes propriedades: (i) A sinopse de vídeo é ela própria um vídeo, expressando a dinâmica da cena. (ii) Para reduzir tanta redundância espaço-temporal quanto possível, o sincronismo relativo entre as atividades pode mudar.

Como um exemplo, consideremos o clip de vídeo esquemático representado como um volume de espaço-tempo na figura 1. O vídeo começa com uma pessoa caminhando na terra, e após um período de inatividade um pássaro está voando no céu. Os quadros inativos são omitidos na maior parte dos métodos de abstração de vídeo. A sinopse de vídeo é substancialmente mais compacta, ao rodar a pessoa e o pássaro simultaneamente. Isso constitui um uso ideal de regiões da imagem ao deslocar eventos de seu intervalo de tempo original a um outro intervalo de tempo quando nenhuma outra atividade ocorre nessa localização espacial. Tais manipulações relaxam a consistência cronológica dos eventos tal como foi apresentado primeiramente em [14].

A invenção também apresenta um método de baixo nível para produzir a sinopse de vídeo utilizando otimizações em Campos Randômicos de Markov [9].

Uma das opções fornecidas pela invenção é a capacidade de exibir múltiplos aparecimentos dinâmicos de um único objeto. Esse efeito é uma generalização dos retratos "estroboscópicos" utilizados na sinopse de vídeo tradicional de objetos móveis [6, 1]. Dois esquemas diferentes para fazer isso são apresentados. Em um primeiro esquema, os instantâneos do objeto em períodos de tempo diferentes são

apresentados no vídeo de saída de modo a fornecer uma indicação do progresso do objeto por todo o vídeo de uma localização inicial a uma localização final. Em um segundo esquema, o objeto não tem nenhuma localização inicial ou final definida, mas se move aleatória e imprevisivelmente. Neste caso, os instantâneos do objeto em períodos de tempo diferentes são apresentados outra vez no vídeo de saída, mas neste tempo dá a impressão de um número maior de objetos aumentados do que há realmente. O que ambos os esquemas compartilham em comum é que múltiplos instantâneos tirados em tempos diferentes de um vídeo de entrada são copiados em um vídeo de saída de uma maneira tal que é evitada a sobreposição espacial e sem copiar os dados de vídeo de entrada que não contribuem com o progresso dinâmico dos objetos de interesse.

Dentro do contexto da invenção e das reivindicações anexas, o termo "vídeo" é sinônimo de "filme" em seu termo mais geral contanto apenas que seja acessível como um arquivo de imagem de computador passível de pós-processamento e inclua qualquer tipo de arquivo de filme, por exemplo, digital, analógico. A câmera fica de preferência em uma posição fixa, o que significa que ela pode girar e efetuar zoom - mas não sujeitada a um movimento de translação tal como ocorre nas técnicas propostas até o presente. As cenas às quais a presente invenção diz respeito são dinâmicas em oposição, por exemplo, às cenas estáticas processadas na Patente U.S. n°. 6.665.003 [22] e outras referências relacionadas à exibição de imagens estereoscópicas qual não mostram uma cena dinâmica na qual quadros sucessivos dinâmicos têm continuidade espacial e temporal. De acordo com um aspecto da invenção, o problema é formulado como um problema de mini-corte simples que pode ser solucionado em tempo polinomial ao encontrar um fluxo máximo em um gráfico

[5].

A fim de descrever a invenção, será utilizada uma construção que é indicada como do "volume no espaço-tempo" para criar os vídeos panorâmicos dinâmicos. O volume no espaço-tempo pode ser construído a partir da seqüência de entrada das imagens ao empilhar sequencialmente todos os quadros ao longo do eixo do tempo. No entanto, deve ficar compreendido que até onde diz respeito à implementação real, não é necessário realmente construir o volume no espaço-tempo, por exemplo, ao empilhar realmente no tempo quadros bidimensionais de uma cena de fonte dinâmica. Mais tipicamente, os quadros da fonte são processados individualmente para construir quadros alvo, mas isso irá ajudar na compreensão da referência ao volume no tempo e espaço como se fosse uma construção física e não uma construção conceitual.

BREVE DESCRIÇÃO DOS DESENHOS

A fim de compreender a invenção e ver como ela pode ser executada na prática, uma realização preferida será descrita agora, apenas a título de exemplo não-limitador, com referência aos desenhos anexos, nos quais:

a Figura 1 é uma representação ilustrativa que mostra a abordagem da presente invenção para a produzir uma sinopse de vídeo compacto ao executar características temporalmente deslocadas simultaneamente;

as Figuras 2a e 2b são representações esquemáticas que ilustram as sinopses de vídeo geradas de acordo com a invenção;

as Figuras 3a, 3b e 3c são representações ilustrativas que mostram exemplos do rearranjo temporal de acordo com a invenção;

a Figura 4 é uma representação ilustrativa que mostra um único quadro de uma sinopse de vídeo utilizando um

efeito estroboscópico dinâmico ilustrado na Figura 3b;

as Figuras 5a, 5b e 5c são representações ilustrativas que mostram um exemplo quando uma sinopse curta pode descrever uma seqüência mais longa sem nenhuma perda de atividade e sem o efeito estroboscópico;

a Figura 6 é uma representação ilustrativa que mostra um exemplo adicional de uma sinopse de vídeo panorâmica de acordo com a invenção;

as Figuras 7a, 7b e 7c são representações ilustrativas que mostram detalhes de uma sinopse de vídeo de vigilância de rua;

as Figuras 8a e 8b são representações ilustrativas que mostram detalhes de uma sinopse de vídeo de vigilância de cerca;

a Figura 9 é uma representação ilustrativa que mostra a densidade de atividade crescente de um filme de acordo com uma realização adicional da invenção;

a Figura 10 é um diagrama esquemático do processo utilizado para gerar o filme mostrado na Figura 9;

a Figura 11 é um diagrama de blocos que mostra a funcionalidade principal de um sistema de acordo com a invenção; e

a Figura 12 é um fluxograma que mostra a operação principal executada de acordo com a invenção.

DESCRIÇÃO DETALHADA DAS REALIZAÇÕES

1. Detecção de Atividade

A invenção supõe que cada pixel de entrada foi etiquetado com seu nível de "importância". Embora a partir de agora o nível de "importância" será utilizado o nível de atividade, é evidente que qualquer outra medida pode ser utilizada para a "importância" com base no pedido requerido. A avaliação do nível de importância (ou de atividade) é suposta e não é ela própria uma característica da invenção.

Ela pode ser obtida utilizando um dentre vários métodos de detecção de irregularidades [4, 17], a detecção de objetos móveis, e o acompanhamento de objetos. Alternativamente, ela pode ser baseada em algoritmos de reconhecimento, tal como a detecção de rostos.

A título de exemplo, um indicador de atividade simples geralmente utilizado pode ser selecionado, onde um pixel de entrada $I(x,y,t)$ é etiquetado como "ativo" se a sua diferença de cor da média temporal na posição (x,y) for maior do que um determinado ponto inicial. Os pixels ativos são definidos pela função característica:

$$\chi(p) = \begin{cases} 1 & se \quad p \quad for \quad ativo \\ 0 & \quad\quad se \quad não \end{cases}$$

Para limpar o indicador de atividade do ruído, um filtro mediano é aplicado a χ antes de continuar com o processo de sinopse.

Embora seja possível utilizar uma medida de atividade contínua, os autores da presente invenção se concentraram no caso binário. Uma medida de atividade contínua pode ser utilizada com quase todas as equações na seguinte descrição detalhada com apenas pequenas mudanças [4, 17, 1].

Foram descritas duas realizações diferentes para a computação da sinopse de vídeo. Uma abordagem (Seção 2) utiliza a representação de gráfico e a otimização da função de custo utilizando cortes de gráfico. Uma outra abordagem (Seção 3) utiliza a segmentação e o acompanhamento de objetos.

2. Sinopse de Vídeo por Minimização de Energia

Deixar N quadros de uma seqüência de vídeo de entrada ser representados em um volume de espaço-tempo tridimensional $I(x,y,t)$, onde (x,y) são as coordenadas

espaciais desse pixel, e $1 \leq t \leq N$ é o número de quadros.

Seria desejável a geração de uma sinopse de vídeo $S(x, y, t)$ que tem as seguintes propriedades:

- A sinopse de vídeo S deve ser substancialmente mais curta do que o vídeo original I.
- A "atividade máxima" do vídeo original deve aparecer na sinopse de vídeo.
- O movimento dos objetos na sinopse de vídeo deve ser similar ao seu movimento no vídeo original.
- A sinopse de vídeo deve parecer bem, e emendas visíveis ou objetos fragmentados devem ser evitados.

A sinopse de vídeo S que tem as propriedades acima é gerado com um mapeamento M, atribuindo a cada coordenada (x,y,t) na sinopse S as coordenadas de um pixel de origem de I. Foi dado enfoque ao deslocamento de tempo dos pixels, mantendo as posições espaciais fixas. Desse modo, qualquer pixel de sinopse $S(x,y,t)$ pode advir de um pixel de entrada $I(x,y,M(x,y,t))$. O deslocamento de tempo M é obtido ao solucionar um problema de minimização de energia, onde a função de custo é fornecida por

$$E(M) = E_a(M) + \alpha E_d(M), \qquad (1)$$

onde $E_a(M)$ indica a perda na atividade, e $E_d(M)$ indica a descontinuidade através de emendas. A perda de atividade será o número de pixels ativos no vídeo de entrada I que não aparecem na sinopse de vídeo S,

$$E_a(M) = \sum_{(x,y,t)\in I} \chi(x,y,t) - \sum_{(x,y,t)\in S} \chi(x,y,M(x,y,t)). \qquad (2)$$

O custo de descontinuidade E_d é definido como a soma de diferenças de cores através das emendas entre os vizinhos espaço-temporais na sinopse de vídeo e os vizinhos correspondentes no vídeo de entrada (Uma formulação similar de A pode ser encontrada em [1]):

$$\sum_d (M) = \sum_{(x,y,t)\in S} \sum_i \left\| S((x,y,t)+e_i) - I((x,y,M(x,y,t))+e_i \right\|^2 \qquad (3)$$

onde e_i são os seis vetores unitários que representam os seis vizinhos espaço-temporais.

As figuras 2a e 2b são representações esquemáticas que ilustram as operações no espaço-tempo que criam uma sinopse de vídeo curta através da minimização da função de custo onde o movimento de objetos móveis é ilustrado pelas "tiras de atividade" nas figuras. A parte superior representa o vídeo original, ao passo que a parte inferior representa a sinopse de vídeo. Especificamente, na figura 2a a sinopse de vídeo mais curta *S* é gerada do vídeo de entrada *I* ao incluir os pixels mais ativos. Para assegurar a lisura, quando o pixel *A* em *S* corresponde ao pixel *B* em *I*, os seus vizinhos "além da fronteira" devem ser similares. Encontrar a minimização *M* ideal (3) é um problema de otimização muito grande. Uma solução aproximada é mostrada na figura 2b onde os pixels consecutivos na vídeo de sinopse são impedidos de vir dos pixels de entrada consecutivos.

Deve-se observar que a função de custo *E(M)* (Equação 1) corresponde a um campo aleatório tridimensional de Markov (MRF) onde cada nó corresponde a um pixel no volume tridimensional do filme de saída, e pode ser designado por qualquer valor de tempo que corresponde a um quadro de entrada. Os pesos nos nós estão determinados pelo custo da atividade, ao passo que as bordas entre os nós são determinadas de acordo com o custo da descontinuidade. A função de custo pode, portanto, ser minimizada por algoritmos tais como cortes de gráfico iterativos [9].

2.1. Solução Restringida Utilizando um Gráfico Bidimensional

A otimização da Equação (1), permitindo que cada pixel na sinopse de vídeo venha de qualquer tempo, é um problema de grande escala. Por exemplo, um vídeo de entrada

de três minutos que é resumido em uma sinopse de vídeo de cinco segundos resulta em um gráfico com aproximadamente 225 nós, cada um dos quais tem 5.400 etiquetas.

Foi mostrado em [2] que, para casos de texturas ou objetos dinâmicos que se movem em uma trajetória horizontal, MRFs tridimensionais podem ser solucionados eficientemente ao reduzir o problema a um problema unidimensional. Nesse trabalho, são visados os objetos que se movem de uma maneira mais geral, e, portanto, são utilizadas restrições diferentes. Os pixels consecutivos na sinopse de vídeo S são impedidos de vir dos pixels consecutivos no vídeo de entrada I. Sob essa restrição, o gráfico tridimensional é reduzido em um gráfico bidimensional onde cada nó corresponde a uma localização espacial no filme da sinopse. A etiqueta de cada nó $M(x,y)$ determina o número de quadros t em I mostrado no primeiro quadro de S, tal como ilustrado na figura 2b. Existe uma emenda entre duas localizações vizinhas (x_1,y_1) e (x_2,y_2) em S se $M(x_1,y_1) \neq M(x_2,y_2)$, e o custo da descontinuidade $E_d(M)$ ao longo da emenda for uma soma das diferenças de cores nessa posição espacial em todos os quadros em S.

$$E_i(M) = \sum_{x,y} \sum_{i} \sum_{t=1}^{K} \left\| S((x,y,t)+e_i) - I((x,y,M(x,y)+t)+e_i) \right\|^2 \qquad (4)$$

onde e_i são agora quatro vetores unitários que descrevem os quatro vizinhos espaciais.

O número de etiquetas para cada nó é N - K, onde N e K são os números de quadros nos vídeos de entrada e de saída, respectivamente. A perda da atividade para cada pixel é:

$$E_a(M) = \sum_{x,y} \left(\sum_{t=1}^{N} \chi(x,y,t) - \sum_{t=1}^{K} \chi(x,y,M(x,y)+t) \right).$$

3. Sinopse Baseada em Objetos

A abordagem de baixo nível para a sinopse de vídeo

dinâmico tal como descrito anteriormente é limitada para satisfazer propriedades locais tais como evitar emendas visíveis. As propriedades baseadas em objetos de nível mais elevado podem ser incorporadas quando os objetos podem ser detectados. Por exemplo, para evitar o efeito estroboscópico é requerida a detecção e o acompanhamento de cada objeto no volume. Essa seção descreve uma implementação da abordagem baseada em objetos para a sinopse de vídeo dinâmico. Existem diversos métodos de sumário de vídeo baseados em objetos na literatura (por exemplo, [7, 5, 16]), e todos eles utilizam os objetos detectados para a seleção de quadros significativos. Ao contrário desses métodos, a invenção desloca objetos a tempo e cria novos quadros de sinopse que nunca apareceram na seqüência de entrada a fim de fazer um uso melhor do espaço e do tempo.

Em uma realização, os objetos móveis são detectados tal como descrito acima ao comparar cada pixel à média temporal e ao calcular o limite dessa diferença. Isto é seguido pela limpeza de ruído utilizando um filtro mediano espacial, e ao agrupar os componentes conectados espaço-temporais. Deve ser apreciado o fato que há muitos outros métodos na literatura para a detecção e o acompanhamento de objetos que podem ser utilizados para essa tarefa (por exemplo, [7, 17, 21]. Cada processo de detecção e acompanhamento de objetos resulta em um jogo de objetos, onde cada objeto *b* é representado por sua função característica

$$\chi_b(x,y,t) = \begin{cases} 1 & se \quad (x,y,t) \in b \\ 0 & se \quad não, \end{cases} \qquad (5)$$

As figuras que 3a, 3b e 3c são representações ilustrativas que mostram exemplos do rearranjo temporal de acordo com a invenção. As partes superiores de cada figura representam o vídeo original, e as partes inferiores

representam a sinopse de vídeo onde o movimento de objetos móveis é ilustrado pelas "tiras de atividade" nas figuras. A figura 3a mostra dois objetos gravados em tempos diferentes deslocados ao mesmo intervalo de tempo na sinopse de vídeo. A figura 3b mostra um único objeto se movendo durante um período longo dividido em segmentos que tem intervalos mais curtos de tempo, que são então executados simultaneamente, criando um efeito estroboscópico dinâmico. A figura 3c mostra que a interseção dos objetos não perturba a sinopse quando os volumes do objeto são divididos em segmentos.

De cada objeto, os segmentos são criados ao selecionar subconjuntos de quadros em que o objeto aparece. Tais segmentos podem representar intervalos de tempo diferentes, tomados opcionalmente a taxas de amostragem diferentes.

A sinopse de vídeo S será construída a partir do vídeo de entrada I utilizando as seguintes operações:

(1) Os objetos $b_1 \ldots b_r$ são extraídos do vídeo de entrada I.

(2) Um jogo de segmentos não-sobrepostos B é selecionado dos objetos originais.

(3) Um deslocamento temporal M é aplicado a cada segmento selecionado, criando uma sinopse de vídeo mais curta enquanto se evita oclusões entre objetos e permite uma costura sem emendas. Isto é explicado na figura 1 e nas figuras 3a a 3c. A figura 4 é uma representação ilustrativa que mostra um exemplo onde um se obtém único quadro de uma sinopse de vídeo utilizando um efeito estroboscópico dinâmico tal como ilustrado na figura 3b.

As operações (2) e (3) são interrelacionadas, uma vez que seria desejável selecionar os segmentos e deslocar os mesmos no tempo para obter uma sinopse de vídeo curta e sem emendas. Deve-se apreciar o fato que a operação em (2) e (3) acima não precisa ser perfeita. Quando se refere a "segmentos

não-sobrepostos", uma sobreposição pequena pode ser permitida, e se refere a "se evita oclusão" uma sobreposição pequena entre os objetos deslocados no tempo pode ser permitida mas deve ser minimizada a fim de obter um vídeo visualmente apelativo.

Na representação baseada em objetos, um pixel na sinopse resultante pode ter fontes múltiplas (vir de objetos diferentes) e, portanto, foi adicionada uma etapa pós-processamento em que todos os objetos são costurados juntos. A imagem de fundo é gerada ao tomar um valor médio de pixel de todos os quadros da seqüência. Os objetos selecionados podem então ser misturados, utilizando pesos proporcionais à distância (no espaço de RGB) entre o valor do pixel em cada quadro e a imagem mediana. Esse mecanismo de costura é similar àquele utilizado em [6].

O jogo de todos os pixels que são mapeados para um único pixel da sinopse $(x,y,t) \in S$ como $src(x,y,t)$ foi definido, e é denotado o número de pixels (ativos) em um objeto (ou um segmento) b como

$$\#b = \sum_{x,y,t \in I} \chi_b(x,y,t)$$

É então definida uma função de energia que mede o custo para uma seleção de subconjunto de segmentos B e para um deslocamento temporal M. O custo de deslocamento inclui uma perda de atividade E_a, uma penalidade para oclusões entre os objetos E_o e um termo E_l que penaliza sinopses de vídeos longas:

$$E(M,B) = E_a + \alpha E_0 + \beta E_1 \qquad (6)$$

onde

$$E_a = \sum_b \#b - \sum_{b \in B} \#b \qquad (7)$$

$$E_o = \sum_{(x,y,t)\in S} Var\{src(x,y,t)\}$$

$$E_1 = comprimento(S)$$

3.1 Sinopse de Vídeo Com um Comprimento Predeterminado

É descrito agora o caso onde uma sinopse de vídeo curta de um comprimento predeterminado K é construída a partir de um vídeo mais longo. Nesse esquema, cada objeto é dividido em segmentos sobrepostos e consecutivos de comprimento K. Todos os segmentos são deslocados no tempo para começarem no tempo $t = 1$, e fica para decidir quais segmentos devem ser incluídos na sinopse de vídeo. Obviamente, com esse esquema alguns objetos não podem aparecer na sinopse de vídeo.

Foi definido primeiramente um custo de oclusão entre todos os pares de segmentos. Deixar b_i e b_j serem dois segmentos com tempos de aparecimento t_i e t_j; e deixar que o suporte de cada segmento seja representado pela sua função característica χ (tal como na Equação 5).

O custo entre esses dois segmentos é definido como sendo a soma de diferenças de cores entre os dois segmentos, depois de ter deslocado ao tempo $t = 1$.

$$v(b_i,b_j) = \sum_{x,y,t\in S} (I(x,y,t+t_i) - I(x,y,t+t_j))^2 \cdot \chi_{b_i}(x,y,t+t_i) \cdot \chi_{b_j}(x,y,t+t_j) \quad (8)$$

Para a sinopse de vídeo, foi selecionado um jogo parcial de segmentos B que minimiza o custo na Equação 6 onde agora E_1 é a constante K, e o custo de oclusão é fornecido por

$$E_o(B) = \sum_{i,j\in B} v(b_i,b_j) \quad (9)$$

Para evitar ter que mostrar o mesmo pixel espaço-temporal duas vezes (o que é admissível, mas um desperdício),

$v(b_i,b_j) = \infty$ foi ajustado para os segmentos b_i e b_j que se interceptam no filme original. Além disso, se o efeito estroboscópico for indesejável, ele pode ser evitado ao ajustar $v(b_i,b_j) = \infty$ para todo b_i e b_j que foram amostrados do mesmo objeto.

O arrefecimento simulado [8] é utilizado para minimizar a função de energia. Cada estado descreve o subconjunto de segmentos que são incluídos na sinopse, e os estados vizinhos são tomados para que sejam os jogos em que um segmento é removido, adicionado ou substituído por um outro segmento.

Após a seleção do segmento, um filme de sinopse de comprimento K é construído ao colar todos os segmentos deslocados uns aos outros. Um exemplo de um quadro de uma sinopse de vídeo que utiliza essa abordagem é fornecido na figura 4.

3.2 Sinopse de Vídeo Sem Perda

Para algumas aplicações, tais como a vigilância de vídeo, é possível que seja preferível uma sinopse de vídeo mais longa, mas na qual seja garantido que todas as atividades irão aparecer. Nesse caso, o objetivo não consiste em selecionar um jogo de segmentos de objetos tal como foi feito na seção precedente, mas, por outro lado, encontrar um rearranjo temporal compacto dos segmentos de objetos.

Outra vez, foi utilizado o arrefecimento simulado para minimizar a energia. Nesse caso, um estado corresponde a um jogo de deslocamentos do tempo para todos os segmentos, e dois estados são definidos como vizinhos se os seus deslocamentos do tempo diferirem para somente um único segmento. Há duas questões que devem ser anotadas neste caso:

- Os segmentos de objetos que aparecem no primeiro ou no último quadros devem permanecer assim na sinopse de vídeo;

(ou então eles podem de repente aparecer ou desaparecer). Cuidado foi tomado para que cada estado satisfaça essa restrição ao fixar os deslocamentos temporais de todos esses objetos de maneira correspondente.

- O arranjo temporal do vídeo de entrada é geralmente um mínimo local da função de energia, e, portanto, não é uma escolha preferível para inicializar o processo de arrefecimento. O arrefecimento simulado foi inicializado com um vídeo mais curto, onde todos os objetos se sobrepõem.

As figuras 5a, 5b e 5c são representações ilustrativas que mostram um exemplo dessa abordagem quando uma sinopse curta pode descrever uma seqüência mais longa sem nenhuma perda da atividade e sem o efeito estroboscópico. Três objetos podem ser deslocados no tempo para aparecerem simultaneamente. Especificamente, a figura 5a ilustra o diagrama esquemático de espaço-tempo do vídeo original (alto) e da sinopse de vídeo (fundo). A figura 5b ilustra três quadros de vídeo original; tal como visto no diagrama na figura 5a, no vídeo original cada pessoa aparece separadamente, mas na sinopse de vídeo todos os três objetos podem aparecer juntos. A figura 5c ilustra um quadro da sinopse de vídeo que mostra todas as três pessoas simultaneamente.

4. Sinopse de Vídeo Panorâmica

Quando uma câmera vídeo está fazendo a varredura de uma cena, muita redundância pode ser eliminada ao utilizar um mosaico panorâmico. Apesar disso, os métodos existentes constroem uma única imagem panorâmica, em que a dinâmica da cena é perdida. A dinâmica limitada pode ser representada por uma imagem estroboscópica [6, 1, 3], onde os objetos móveis são indicados em diversas localizações ao longo de suas trajetórias.

Uma sinopse de vídeo panorâmica pode ser criada ao

exibir simultaneamente as ações que ocorreram em tempos diferentes em regiões diferentes da cena. Uma condensação substancial pode ser obtida, uma vez que a duração da atividade para cada objeto é limitada ao tempo em que está sendo visto pela câmera. Um caso especial é quando a câmera segue um objeto tal como a leoa correndo mostrada na figura 6. Quando uma câmera segue a leoa correndo, a sinopse de vídeo é um mosaico panorâmico do fundo, e o primeiro plano inclui diversas cópias dinâmicas da leoa correndo. Neste caso, uma sinopse de vídeo curta pode ser obtida somente ao permitir o efeito estroboscópico.

A construção da sinopse de vídeo panorâmica é feita de uma maneira similar à sinopse de vídeo regular, com um estágio preliminar de alinhamento de todos os quadros em relação a algum quadro de referência. Após o alinhamento, as coordenadas da imagem dos objetos são tomadas de um sistema de coordenadas global, que pode ser o sistema de coordenadas de uma das imagens de entrada.

A fim de poder processar vídeos até mesmo quando a segmentação de objetos móveis não é perfeita, se tem oclusões penalizadas em vez de seu impedimento total. Essa penalidade de oclusão permite uma flexibilidade no arranjo temporal dos objetos, até mesmo quando a segmentação não é perfeita, e os pixels de um objeto podem incluir algum fundo.

Termos adicionais podem ser adicionados, que impelem o ordenamento temporal da sinopse de vídeo para o ordenamento do vídeo de entrada.

A minimização da energia acima em todas as seleções de segmentos possíveis B e um deslocamento temporal M é muito exaustiva devido ao grande número de possibilidades. No entanto, o problema pode ser reduzido de maneira significativamente ao restringir as soluções. Dois esquemas restringidos são descritos nas seguintes seções.

5. Exemplos de Vigilância

Uma aplicação interessante para a sinopse de vídeo pode ser o acesso a vídeos de vigilância armazenados. Quando se torna necessário examinar determinados eventos no vídeo, isso pode ser feito muito mais rapidamente com a sinopse de vídeo.

Conforme observado acima, a figura 5 mostra um exemplo do poder da sinopse de vídeo na condensação de toda a atividade em um período curto, sem perder nenhuma atividade. Isto foi feito ao utilizar um vídeo coletado de uma câmera que monitora uma estação de café. Dois exemplos adicionais são fornecidos de câmeras reais de vigilância. As figuras 8a, 8b e 8c são representações ilustrativas que mostram detalhes de uma sinopse de vídeo de vigilância de rua. A figura 8a mostra um quadro típico do vídeo original (22 segundos). A figura 8b ilustra um quadro de um filme da sinopse de vídeo (2 segundos) mostrando uma atividade condensada. A figura 8c ilustra um quadro de uma sinopse de vídeo mais curta (0,7 segundo), mostrando uma atividade ainda mais condensada. As imagens mostradas nessas figuras foram derivadas de um vídeo capturado por uma câmera que vigia uma rua da cidade, com os pedestres cruzando ocasionalmente o campo de visão. Muitas delas podem ser coletadas em uma sinopse muito condensada.

As figuras 8a e 8b são representações ilustrativas que mostram detalhes de uma sinopse de vídeo de vigilância de cerca. Há uma atividade muito pequena perto da cerca, e de tempo em tempo é possível ver um soldado rastejar na direção da cerca. A sinopse de vídeo mostra todos os casos de soldados rastejando e andando simultaneamente, ou opcionalmente tornando a sinopse de vídeo ainda mais curta ao ser apresentada estroboscopicamente.

6. Indexação de Vidro Através de Sinopse de Vídeo

A sinopse de vídeo pode ser utilizada para a

indexação de vídeo, fornecendo ao usuário links eficientes e intuitivos para ações de acesso nos vídeos. Isto pode ser feito ao associar com cada pixel de sinopse um ponteiro para o aparecimento do objeto correspondente no vídeo original. Na sinopse de vídeo, a informação de vídeo é projetada no "espaço de atividades", em que somente as atividades importam, independentemente de seu contexto temporal (embora ainda seja preservado o contexto espacial). Uma vez que as atividades são concentradas em um período curto, as atividades específicas no vídeo podem ser alcançadas com facilidade.

Ficará evidente a partir da descrição acima que, quando uma câmera de vídeo está fazendo a varredura de uma cena dinâmica, o "tempo cronológico" absoluto no qual uma região se torna visível no vídeo de entrada, não faz parte da dinâmica da cena. O "tempo local" durante o período de visibilidade de cada região é mais relevante para a descrição da dinâmica na cena, e deve ser preservado quando da construção de mosaicos dinâmicos. As realizações descritas acima apresentam um primeiro aspecto da invenção. De acordo com um segundo aspecto, será mostrado agora como criar mosaicos panorâmicos sem emendas, em que a costura entre imagens evita tanto quanto possível o corte de partes dos objetos na cena, até mesmo quando esses objetos podem estar se movendo.

7. Criação de Imagem Panorâmica Utilizando um Corte Mínimo Tridimensional

Deixar $I_1,\ldots,I_n$ ser os quadros da seqüência de entrada. Se supõe que a seqüência foi alinhada para um único quadro de referência utilizando um dos métodos existentes. Para fins de simplificação, será suposto que todos os quadros depois do alinhamento são do mesmo tamanho (os pixels fora do campo de visão da câmera estarão marcados como não-válidos).

Também irá se supor que a câmera está girando no sentido horário. (Os movimentos diferentes podem ser mantidos de uma maneira similar).

Deixar que $P(x,y)$ seja a imagem panorâmica construída. Para cada pixel (x,y) em P é preciso escolher o quadro $M(x,y)$ do qual esse pixel é tirado. (Isto é, se $M(x,y) = k$ então $P(x,y) = I_k(x,y)$). Obviamente, sob a suposição que a câmera está girando no sentido horário, a coluna da esquerda deve ser tirada do primeiro quadro, enquanto a coluna da direita dever ser tirada do último quadro. (Outras condições de limite podem ser selecionadas para produzir imagens panorâmicas com um campo de visão menor).

O objetivo aqui é a produção de uma imagem panorâmica sem emendas. Para obter isto, tentar-se-á evitar a costura dentro de objetos, particularmente se eles estiverem se movendo. Foi empregada uma contagem de emendas similar à contagem utilizada por [1], mas em vez de solucionar (com aproximação) um problema difícil de *NP*, será encontrada uma solução ideal para um problema mais restrito.

8. Formulação do Problema Como um Problema de Minimização de Energia

A diferença principal das formulações precedentes é o custo de costura, definido por:

$$E_{costura}(x,y,x',y') = \sum_{k=\min M}^{\max M-1} \frac{1}{2}\left\|I_{k+1} - I_{k+1}(x,y)\right\|^2 + \frac{1}{2}\left\|I_k(x',y')\right\|^2 \quad (10)$$

onde:

$minM = min(M(x,y), M(x',y'))$

$maxM = max(M(x,y), M(x',y'))$

Esse custo é razoavelmente supor que a atribuição dos quadros é contínua, o que significa que, se (x,y) e (x',y') forem pixels vizinhos, os seus quadros originais

$M(x,y)$ e $M(x',y')$ são próximos. A vantagem principal desse custo é que ele permite que seja solucionado o problema como um problema de corte mínimo em um gráfico.

A função que da energia nós minimizaremos é:

$$E(M) = \sum_{(x,y)(x',y')\in N(x,y)} \sum E_{costura}(x,y,x',y') + \sum_{(x,y)} (1 - Valid(x,y,M(x,y)) \cdot D, \quad (11)$$

onde:

$N(x,y)$ são os pixels na vizinhança de (x,y).

$E(x,y,x',y')$ é o custo da costura para cada um dos pixels vizinhos, tal como descrito na Equação 1.

$Valid(x,y,k)$ é $1 \Leftrightarrow I_k(x,y)$ é um pixel válido (isto é, no campo de visão da câmera).

D é um número muito grande (que representa o infinito).

9. Construção de um Panorama Simples

Será mostrado em seguida como converter o problema de múltiplas etiquetas bidimensional (que tem uma complexidade exponencial) em um problema binário tridimensional (que tem uma complexidade polinomial, e praticamente pode ser solucionado rapidamente). Para cada pixel x,y e o quadro de entrada k, é definido uma variável binária $b(x,y,k)$ que é igual a um *iff* $M(x,y) <= k$. ($M(x,y)$ é o quadro original do pixel (x,y)). Obviamente, $b(x,y,N) = 1$.

Deve-se observar se for determinado $b(x,y,k)$ para cada $1 \leq k \leq N$, é possível determinar $M(x,y)$ como o k mínimo para o qual $b(x,y,k) = 1$. Será escrito um termo de energia cuja minimização resulta em um panorama sem emendas. Para cada um dos pixels adjacentes (x,y) e (x',y') e para cada k, foi adicionado o termo de erro:

$$\left\| I_k(x,y) - I_{k+1}(x,y) \right\|^2 + \left\| I_k(x',y') - I_{k+1}(x',y') \right\|^2$$

para as atribuições em que $b(x,y,k) \neq b(x',y',k)$. (Este termo de erro é simétrico).

Também foi adicionada uma penalidade infinita para as atribuições em que $b(x,y,k) = 1$ mas $b(x,y,k+1) = 0$. (Y na vez que não é possível que $M(x,y) <= k$ mas $M(x,y) > k$).

Finalmente, se $I_k(x,y)$ for um pixel não-válido, é possível evitar escolher esse pixel ao aplicar uma penalidade infinita às atribuições b(x,y,k) = 1 ^ b(x,y,k+1) = 0 se $k > 1$ ou $b(x,y,k) = 1$ de $k = 1$. (Estas atribuições implicam que $M(x,y) = k$).

Todos os termos acima estão em pares de variáveis em uma grade tridimensional, e, portanto, é possível descrever como minimizar uma função de energia em um MRF binário tridimensional, e minimizar o mesmo em um tempo polinomial utilizando corte mínimo [9].

10. Criação de Filme Panorâmico Utilizando um Corte Mínimo Quadridimensional

Para criar um filme panorâmico (de comprimento L), é necessário criar uma seqüência de imagens panorâmicas. A construção de cada imagem panorâmica independentemente não é boa, uma vez que nenhuma consistência temporal é obtida. Uma outra maneira consiste em começar com uma imagem de mosaico inicial como primeiro quadro, e para as imagens de mosaico consecutivas tomar cada pixel do quadro consecutivo utilizado do mosaico precedente ($M_t(x,y) = M(x,y)+1$). Essa possibilidade é similar àquela que foi descrita acima com referência à figura 2b dos desenhos.

De acordo com o segundo aspecto da invenção, é preferivelmente utilizada uma formulação diferente, que dá à costura uma oportunidade de mudar de um quadro panorâmico a outro, que é muito importante para a costura de objetos móveis bem sucedida.

Foi construído um gráfico quadridimensional que consista em L casos do gráfico tridimensional descrito anteriormente:

$b(x,y,k,l) = 1 \Leftrightarrow M_l(x,y)k.$

Para reforçar a consistência temporal, foi aplicada uma penalidade infinita às atribuições $b(x,y,N,l) = 1$ para cada $1 < L$, e uma penalidade infinita para as atribuições $b(x,y,1,l) = 0$ para cada $l>1$.

Além disso, para cada (x,y,k,l) $(1 \leq 1 \leq L - 1, 1 \leq k \leq N - 1)$ a função de custo é ajustada:

$$E_{temp} = \frac{1}{2}\left\|I_k(x,y) - I_{k+1}(x,y)\right\|^2 + \frac{1}{2}\left\|I_{k+1}(x,y) - I_{k+2}(x,y)\right\|^2 \qquad (12)$$

para as atribuições $b(x,y,k,l) = 1 \neq b(x,y,k+1,l+1)$. (Para $k = N-1$ só é utilizado o termo esquerdo do custo). Esse custo incentiva a exibição de pixels consecutivos (temporais) no filme resultante (a menos que, por exemplo, esses pixels estejam no fundo).

Uma variante desse método consiste em conectar cada pixel (x,y) não ao mesmo pixel no quadro consecutivo, mas ao pixel correspondente $(x+u,y+v)$ de acordo com o fluxo óptico nesse pixel (u,v). Os métodos apropriados para computar o fluxo óptico podem ser encontrados, por exemplo, em [19]. O uso do fluxo óptico lida melhor com o caso de objetos móveis.

Outra vez, é possível minimizar a função de energia ao utilizar cortes mínimos no gráfico quadridimensional, e a solução binária define um filme panorâmico que se reduz os problemas de costura.

11. Melhorias Práticas

Pode ser necessária uma quantidade enorme de memória para salvar o gráfico quadridimensional. Portanto, são empregadas diversas melhorias que reduzem os requisitos de memória e o tempo operacional do algoritmo:

- Conforme mencionado anteriormente, a energia pode ser minimizada sem salvar explicitamente vértices para pixels não-válidos. O número de vértices é reduzido desse modo ao

número de pixels no vídeo de entrada, multiplicado pelo número de quadros no vídeo de saída.

• Em vez da solução para cada quadro no vídeo de saída, só pode ser solucionado para um jogo amostrado de quadros de saída, e interpolada a função de costura entre eles. Essa melhoria é baseada na suposição que o movimento na cena não é muito grande.

• É possível restringir cada pixel para que advenha somente de um jogo parcial de quadros de entrada. Isso faz sentido especialmente para uma seqüência de quadros tomada de um vídeo, onde o movimento entre cada par de quadros consecutivos é muito pequeno. Nesse caso, não se perde muito ao amostrar o jogo de quadros originais para cada pixel. Mas é aconselhável amostrar os quadros originais de uma maneira consistente. Por exemplo, se o quadro k for uma fonte possível para o pixel (x,y) nos l-*ésimo* quadro de saída, então o quadro $k+1$ deve ser um quadro original possível para o pixel (x,y) no $(l+1)$-ésimo quadro de saída.

• É utilizada uma estrutura de multi-resolução (tal como foi feito, por exemplo, em [2]), onde uma solução grosseira é encontrada para imagens de baixa resolução (depois de borrar e de sub-amostrar), e a solução só é refinada nos limites.

12. Combinação de Vídeos com a Contagem de Interesse

Será descrito agora um método para combinar filmes de acordo com uma contagem de interesse. Há diversas aplicações, tais como a criação de um filme com atividade mais densa (ou mais escassa), ou até mesmo o controle da cena em uma maneira especificada pelo usuário.

O panorama dinâmico descrito em [14] pode ser considerado como um caso especial, onde as partes diferentes do mesmo filme são combinadas para se obter um filme com um campo de visão maior: nesse caso, é definida uma contagem de interesse de acordo com a "visibilidade" de cada pixel em

cada tempo. De maneira mais geral, a combinação de partes diferentes (deslocamentos no tempo ou no espaço) do mesmo filme pode ser utilizada em outros casos. Por exemplo, para tornar a atividade no filme mais densa, é possível combinar a parte diferente do filme onde a ação ocorre, a um filme novo com muita ação. A realização descrita acima com referência às figuras 1 a 8 descreve o exemplo especial de maximização da atividade, e utiliza uma metodologia diferente.

Duas questões que devem ser solucionadas são:

1. Como combinar os filmes em um filme "de boa aparência". Por exemplo, é desejável evitar problemas de costura.
2. Maximização da contagem de interesse.

Para começar, são descritas as contagens diferentes que podem ser utilizadas, e então é descrito o esquema utilizado para combinar os filmes.

Uma das características principais que podem ser utilizadas como uma função de interesse para filmes é o nível de "importância" de um pixel. Nas experiências feitas foi considerado que a "atividade" em um pixel indica a sua importância, mas outras medidas da importância também são apropriadas. A avaliação do nível de atividade não é ela própria uma característica da presente invenção e pode ser feita ao utilizar um de vários métodos tal como indicado acima na Seção 1 (Detecção da Atividade).

13. Outras Contagens

Outras contagens que podem ser utilizadas para combinar filmes:

• Contagem da Visibilidade: Quando a câmera está se movendo, ou se alguém tenta preencher uma lacuna em um vídeo, há pixels que não são visíveis. É possível penalizar (não necessariamente com uma contagem infinita) os pixels não-válidos. Desta maneira, é possível incentivar o preenchimento de lacunas (ou o aumento do campo de visão), mas pode ser

preferível não preencher a lacuna, ou utilizar o campo de visão menor se resultar em uma costura má.

• Orientação: A medida da atividade pode ser substituída por uma medida direcional. Por exemplo, é possível favorecer as regiões que se movem horizontalmente em relação às regiões que se movem verticalmente.

• Especificada pelo usuário: O usuário pode especificar uma função favorita de interesse, tal como a cor, a textura, etc. Além disso, o usuário pode especificar regiões (e momentos no tempo) manualmente com contagens diferentes. Por exemplo, ao desenhar uma máscara onde 1 denota que a atividade máxima é desejada, ao passo que 0 denota que nenhuma atividade é desejada, o usuário pode controlar a dinâmica na cena, isto é, para ocorrer em um lugar específico.

14. O Algoritmo

É empregado um método similar àquele utilizado por [20], com as seguintes mudanças:

• Foi adicionada uma contagem de interesse para que cada pixel seja escolhido de um filme ou de um outro. Essa contagem pode ser adicionada utilizando bordas de cada pixel de cada filme para os vértices terminais (fonte e fundo), e os pesos nessas bordas são as contagens de interesse.

• É (opcionalmente) computado o fluxo óptico entre cada par consecutivo de quadros. Então, para reforçar a consistência, é possível substituir as bordas entre vizinhos temporais $((x,y,t)(x,y,t+1))$ pelas bordas entre vizinhos de acordo com o fluxo óptico $((x,y,t)$ a $(x + u(x,y),y + v(x,y),t+1))$. Isso realça a transição entre os filmes costurados,uma vez que incentiva a costura para seguir o fluxo que é menos visível.

• Deve se levar em consideração não somente o custo da costura, mas também a contagem de interesse quando se decide sobre quais as partes de um filme (ou quais filmes) se deve combinar. Por exemplo, ao criar um filme com o nível de

atividade mais denso, é escolhido um jogo de filmes S que maximiza a contagem:

$$\sum_{x,y,t} \bigcup_{b \in S} \chi_b(x,y,t)$$

A figura 9b é uma representação ilustrativa que demonstra esse efeito como a densidade aumentada da atividade de um filme, um quadro original do qual é mostrado na figura 9a. Quando mais de dois filmes são combinados, é utilizada uma abordagem iterativa, onde em cada iteração um novo filme é combinado no filme resultante. Para fazer isso corretamente, devem ser consideradas as emendas e as contagens antigas que resultaram das iterações precedentes. Esse esquema, embora sem as contagens de interesse, é descrito por [20]. Um quadro de amostra do vídeo resultante é mostrado na figura 9b.

A figura 10 é um diagrama esquemático do processo. Neste exemplo, um vídeo é combinado com uma versão temporalmente deslocada dele mesmo. A combinação é feita ao utilizar um corte mínimo de acordo com os critérios descritos acima, isto é, ao maximizar a contagem de interesse enquanto é minimizado o custo da costura.

Com referência agora à figura 11, é mostrado um diagrama de blocos de um sistema 10 de acordo com a invenção para transformar uma primeira seqüência de quadros de vídeo de uma primeira cena dinâmica capturada por uma câmera 11 em uma segunda seqüência de pelo menos dois quadros de vídeo que ilustram uma segunda cena dinâmica. O sistema inclui uma primeira memória 12 para armazenar um subconjunto de quadros de vídeo na primeira seqüência que mostram o movimento de pelo menos um objeto que compreende uma pluralidade de pixels localizados nas respectivas coordenadas x,y. Uma unidade de seleção 13 é acoplada à primeira memória 12 para selecionar

das partes do subconjunto quais aquelas que mostram aparecimentos não-espacialmente sobrepostos de pelo menos um objeto na primeira cena dinâmica. Um gerador de quadros 14 copia as partes de pelo menos três quadros de entrada diferentes em pelo menos dois quadros sucessivos da segunda seqüência sem mudar as respectivas coordenadas x,y dos pixels no objeto e de maneira tal que pelo menos um dos quadros da segunda seqüência contém pelo menos duas partes que aparecem em quadros diferentes na primeira seqüência. Os quadros da segunda seqüência são armazenados em uma segunda memória 15 para processamento subseqüente ou exibição por uma unidade de exibição 16. O gerador de quadros 14 pode incluir uma unidade de deformação 17 para deformar espacialmente pelo menos duas das partes antes de copiar na segunda seqüência.

O sistema 10 pode ser executado na prática por um computador apropriadamente programado que tem um cartão de gráficos ou uma estação de trabalho e periféricos apropriados, tudo tal como é bem conhecido no estado da técnica.

No sistema 10, pelo menos três quadros de entrada diferentes podem ser temporalmente contíguos. O sistema 10 também pode incluir uma unidade de alinhamento 18 opcional acoplada à primeira memória para pré-alinhar a primeira seqüência de quadros de vídeo. Neste caso, a câmera 11 será acoplada à unidade de alinhamento 18 de modo a armazenar os quadros de vídeo pré-alinhados na primeira memória 12. A unidade de alinhamento 18 pode operar por meio:

da computação dos parâmetros de movimento da imagem entre os quadros na primeira seqüência

a deformação dos quadros de vídeo na primeira seqüência de modo que os objetos estacionários na primeira cena dinâmica fiquem estacionários no vídeo.

Do mesmo modo, o sistema 10 também pode incluir um

gerador de fatias do tempo 19 opcional acoplado à unidade de seleção 13 para varrer o volume de espaço-tempo alinhado por uma "frente do tempo" e gerar uma seqüência de fatias de tempo.

Essas características opcionais não são descritas em detalhes, uma vez que elas, bem como os termos uma "frente de tempo" e "fatias de tempo", são descritas integralmente no pedido de patente WO2006/048875 acima mencionado ao qual é feita referência.

Para fins de integralidade, a figura 12 é um fluxograma que mostra as operações principais executadas pelo sistema 10 de acordo com a invenção.

15. Discussão

A sinopse de vídeo foi proposta como uma abordagem para condensar a atividade em um vídeo em um período de tempo muito curto. Essa representação condensada pode permitir o acesso eficiente às atividades nas seqüências de vídeo. Duas abordagens foram apresentadas: uma abordagem utiliza a otimização de gráfico de baixo nível, onde cada pixel na sinopse de vídeo é um nó nesse gráfico. Essa abordagem tem o benefício de obter a sinopse de vídeo diretamente do vídeo de entrada, mas a complexidade da solução pode ser muito grande. Uma abordagem alternativa consiste em detectar primeiramente os objetos móveis, e executar a otimização nos objetos detectados. Embora uma etapa preliminar de segmentação do movimento seja necessária na segunda aproximação, ela é muito mais rápida, e são possíveis as restrições baseadas em objetos. A atividade na sinopse de vídeo resultante é muito mais condensada do que a atividade em qualquer vídeo comum, e a visualização de tal sinopse pode parecer esquisita ao observador inexperiente. Mas quando o objetivo consiste em observar muita informação em um tempo curto, a sinopse de vídeo atinge esse objetivo. Uma atenção especial deve ser

prestada à possibilidade de obter estroboscopia dinâmica. Embora permita uma redução adicional no comprimento da sinopse de vídeo, a estroboscopia dinâmica pode precisar de uma adaptação adicional do usuário. Leva algum tempo para treinar para se dar conta que as ocorrências espaciais múltiplas de um único objeto indicam um tempo de atividade mais longo. Embora tenha sido detalhada uma execução específica para a sinopse de vídeo dinâmica, muitas extensões são diretas. Por exemplo, ao invés de ter um indicador de "atividade" binário, o indicador de atividade pode ser contínuo. Uma atividade contínua pode estender as opções disponíveis para criar a sinopse de vídeo, por exemplo, ao controlar a velocidade dos objetos exibidos com base em seus níveis de atividade. A sinopse de vídeo também pode ser aplicada para os filmes longos que consistem em muitas tomadas. Teoricamente, o presente algoritmo não irá ajuntar as partes das cenas diferentes devido à penalidade de oclusão (ou a descontinuidade). Neste caso, o modelo de fundo simples utilizado para uma única tomada tem que ser substituído por um estimador de fundo ajustável. Uma outra abordagem que pode ser aplicada em filmes longos consiste no emprego de um método existente para a detecção do limite de tomada e a criação da sinopse de vídeo em cada um tomada separadamente.

Também deve ficar compreendido que o sistema de acordo com a invenção pode ser um computador apropriadamente programado. Do mesmo modo, a invenção contempla um programa de computador que pode ser lido por um computador para executar o método da invenção. A invenção contempla adicionalmente uma memória que pode ser lida por máquina que incorpora tangivelmente um programa de instruções executável pela máquina para executar o método da invenção.

REIVINDICAÇÕES

1. MÉTODO PARA A CRIAÇÃO DE UMA SINOPSE DE VÍDEO, ao transformar uma seqüência de origem de quadros de vídeo de uma primeira cena dinâmica capturada por uma câmera de vídeo a uma seqüência de sinopse mais curta dos quadros de vídeo que ilustram uma segunda cena dinâmica, no qual o método compreende a obtenção de um subconjunto de quadros de vídeo na dita seqüência de origem que mostram o movimento de pelo menos um objeto, em que cada objeto é um subconjunto conectado de pixels de pelo menos três quadros diferentes do vídeo de origem; em que o método é caracterizado pelo fato de compreender:

a seleção de pelo menos três objetos da dita seqüência de origem, e a amostragem de cada objeto de origem selecionado de um ou mais objetos de sinopse pela amostragem temporal;

a determinação, para cada objeto de sinopse, de um respectivo tempo de exibição para iniciar a sua exibição no vídeo de sinopse; e

a geração do vídeo de sinopse através da exibição dos objetos de sinopse selecionados, cada um dos quais em seu tempo respectivo tempo de exibição predeterminado sem mudar a localização espacial dos ditos objetos na primeira cena dinâmica de maneira tal que pelo menos três pixels, cada um deles derivado de respectivos tempos diferentes na seqüência de origem, são exibidos simultaneamente no vídeo de sinopse.

2. MÉTODO, de acordo com a reivindicação 1, caracterizado pelo fato de que um dos objetos é um objeto de fundo.

3. MÉTODO, de acordo com a reivindicação 2, caracterizado pelo fato de incluir a costura dos objetos e do fundo em um vídeo sem costura.

4. MÉTODO, de acordo com qualquer uma das

reivindicações 1 a 3, caracterizado pelo fato de que os objetos de origem são selecionados, e um respectivo tempo para iniciar a exibição de cada objeto de sinopse é determinado de modo a otimizar uma função de custo.

5. MÉTODO, de acordo com qualquer uma das reivindicações 1 a 4, caracterizado pelo fato de que a seqüência de origem é capturada por uma câmera que é girada em relação a um eixo geométrico em uma posição fixa, e inclui a deformação espacial de pelo menos duas das ditas partes antes de copiar para a seqüência de sinopse.

6. MÉTODO, de acordo com qualquer uma das reivindicações 1 a 4, caracterizado pelo fato de que a seqüência de origem é capturada por uma câmera estática em uma posição fixa.

7. MÉTODO, de acordo com qualquer uma das reivindicações 1 a 6, caracterizado pelo fato de que pelo menos três quadros de origem diferentes são temporalmente contíguos.

8. MÉTODO, de acordo com qualquer uma das reivindicações 1 a 7, caracterizado pelo fato de que as partes selecionadas são espacialmente contíguas na primeira cena dinâmica.

9. MÉTODO, de acordo com qualquer uma das reivindicações 1 a 8, caracterizado pelo fato de que dois eventos que ocorrem simultaneamente na seqüência de vídeo de origem são exibidos em momentos diferentes na seqüência de sinopse de vídeo.

10. MÉTODO, de acordo com qualquer uma das reivindicações 1 a 9, caracterizado pelo fato de ser utilizado para qualquer um dos seguintes: sinopse de vídeo para vigilância; aumento da densidade da atividade de um filme; indexação de vídeo.

11. MÉTODO, de acordo com a reivindicação 10,

caracterizado pelo fato de incluir a manutenção, para cada pixel na seqüência de sinopse, de um ponteiro para um pixel correspondente na seqüência de origem.

12. MÉTODO, de acordo com qualquer uma das reivindicações 1 a 11, caracterizado pelo fato de incluir o pré-alinhamento da seqüência de origem para obter uma seqüência de origem alinhada, por meio da:

(a) computação de parâmetros de movimento da imagem entre quadros na seqüência de origem; e

(b) deformação dos quadros de vídeo na seqüência de origem de modo que os objetos estacionários na primeira cena dinâmica sejam estacionários na seqüência de origem alinhada.

13. SISTEMA PARA TRANSFORMAR UMA SEQÜÊNCIA DE ORIGEM DE QUADROS DE VÍDEO DE UMA PRIMEIRA CENA DINÂMICA EM UMA SEQÜÊNCIA DE SINOPSE DE PELO MENOS DOIS QUADROS DE VÍDEO QUE ILUSTRAM UMA SEGUNDA CENA DINÂMICA, em que o sistema compreende:

uma primeira memória (12) para armazenar um subconjunto de quadros de vídeo na dita seqüência de origem que mostram o movimento de pelo menos um objeto, e cada objeto é um subconjunto conectado dos pixels de pelo menos três quadros de origem diferentes,

em que o sistema é caracterizado pelo fato de compreender:

uma unidade de seleção (13) acoplada à primeira memória (12) para selecionar pelo menos três objetos de origem da dita seqüência de origem, e para amostrar de cada objeto de origem selecionado um ou mais objetos de sinopse através de amostragem temporal,

um gerador de quadros (14) para determinar para cada objeto de sinopse um respectivo tempo de exibição para iniciar a sua exibição no vídeo de sinopse e gerar o vídeo de sinopse ao exibir objetos de sinopse selecionados ou objetos

derivados dos mesmos, cada um dos quais em seu respectivo tempo de exibição predeterminado sem mudar a posição espacial dos ditos objetos ou dos respectivos objetos derivados dos mesmos na primeira cena dinâmica, de maneira tal que pelo menos três pixels, cada um deles derivado de respectivos tempos diferentes na seqüência de origem, são exibidos simultaneamente no vídeo de sinopse,

uma segunda memória (15) acoplada ao gerador de quadros para armazenar quadros da seqüência de sinopse, e

um meio para acoplar um dispositivo de exibição (16) à segunda memória (15) para exibir a segunda cena dinâmica.

14. SISTEMA, de acordo com a reivindicação 13, caracterizado pelo fato de que o gerador de quadros (14) inclui uma unidade de deformação (17) para deformar espacialmente pelo menos duas das ditas partes antes de copiar para a seqüência de sinopse.

15. PRODUTO DE PROGRAMA DE COMPUTADOR, caracterizado pelo fato de compreender o código do programa de computador para executar o método de acordo com qualquer uma das reivindicações 1 a 12 quando o dito programa for rodado em um computador.

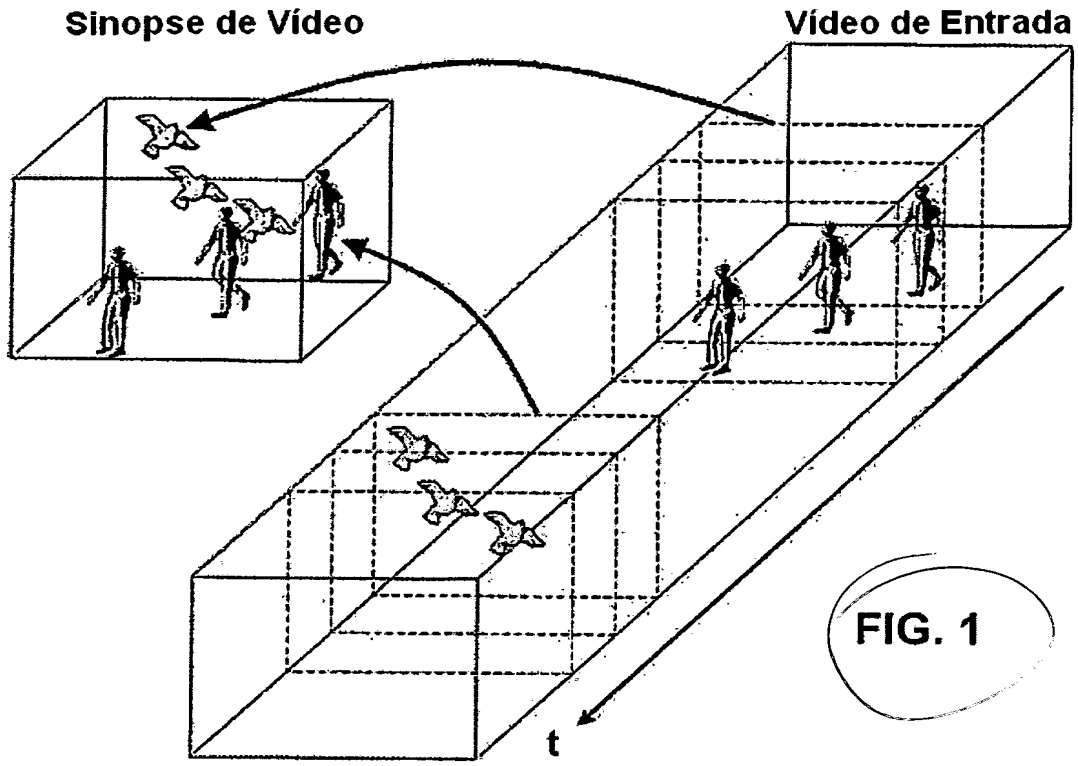


FIG. 1

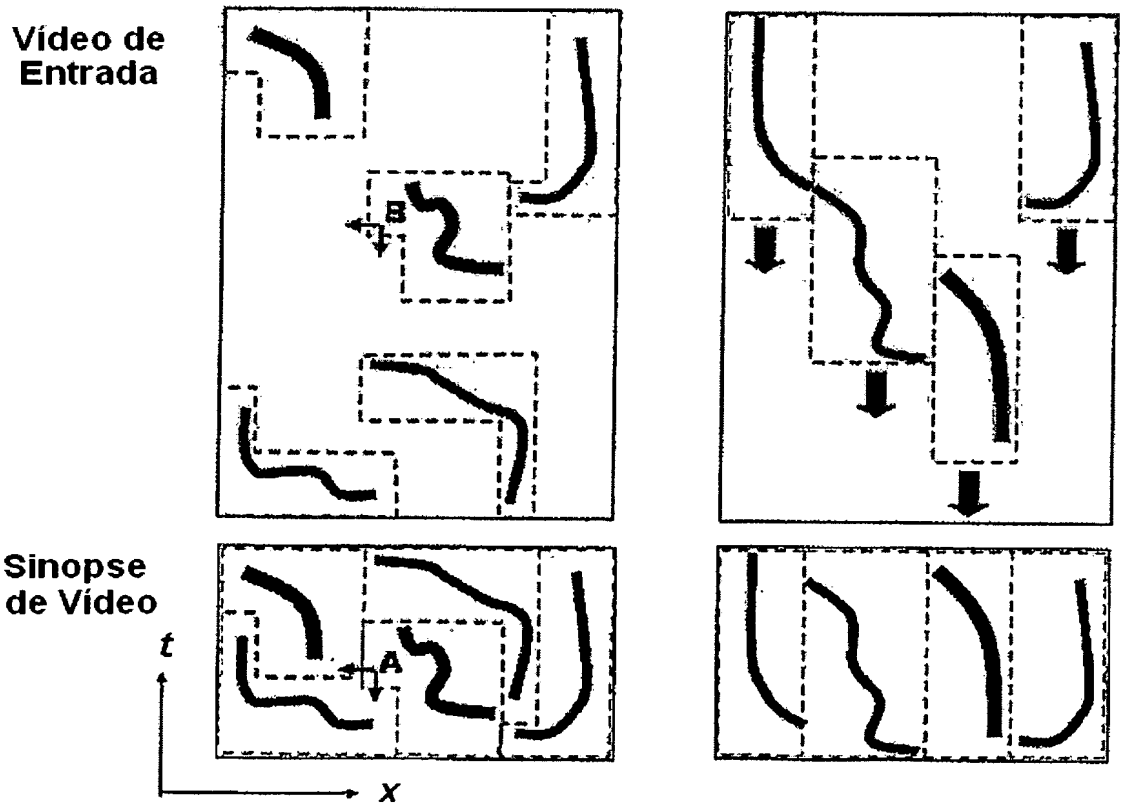


FIG. 2a

FIG. 2b

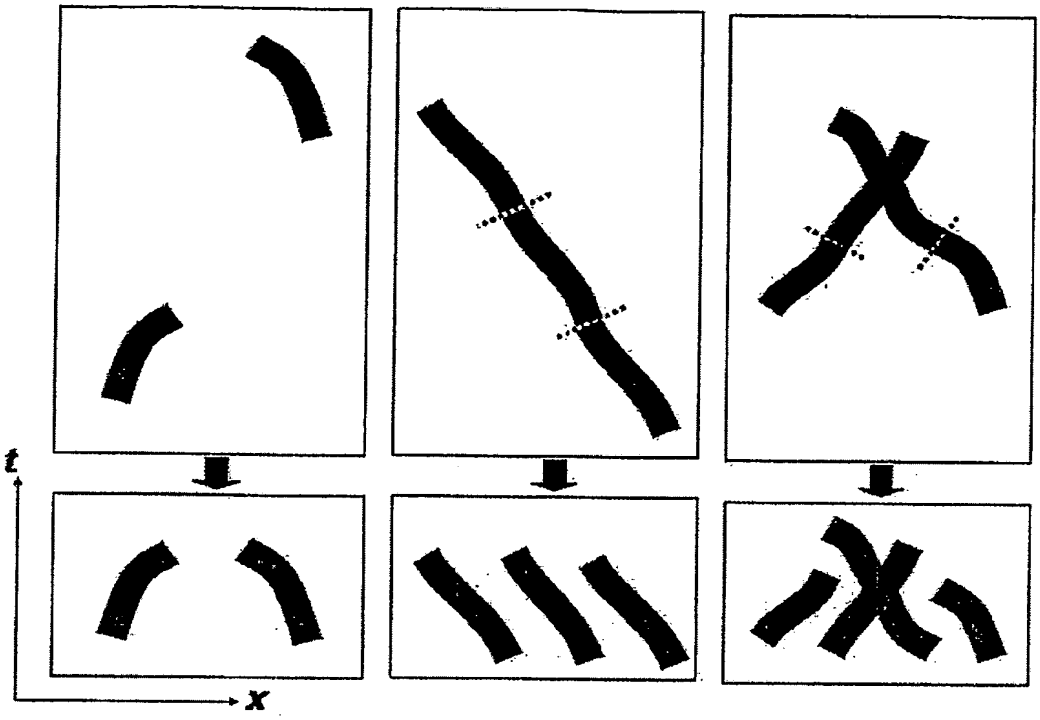


FIG. 3a FIG. 3b FIG. 3c

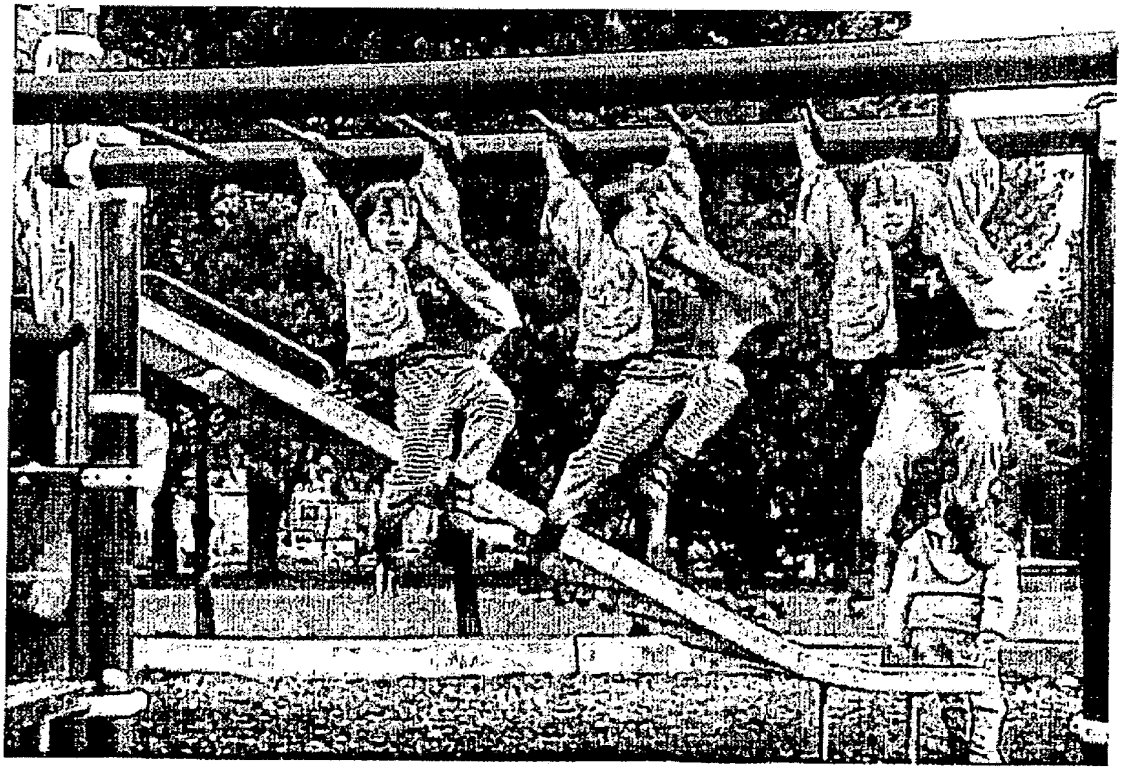


FIG. 4

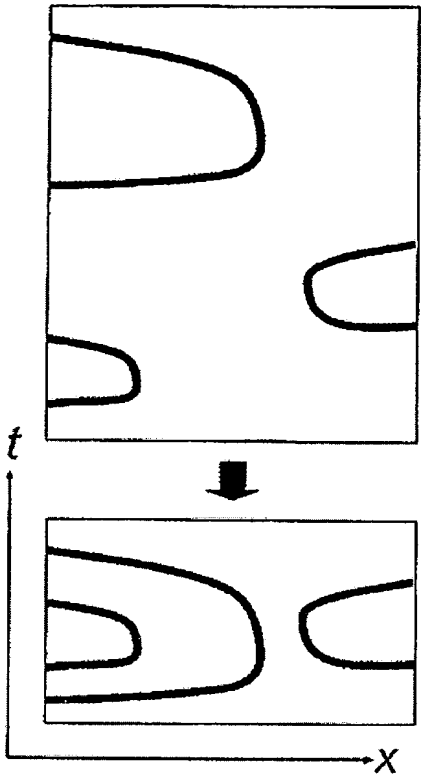


FIG. 5a

FIG. 5b

FIG. 5c

FIG. 6

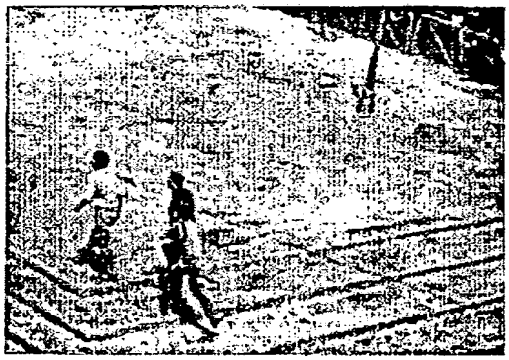


FIG. 7a

FIG. 7b

FIG. 7c

FIG. 8a

FIG. 8b

FIG. 9a

FIG. 9b

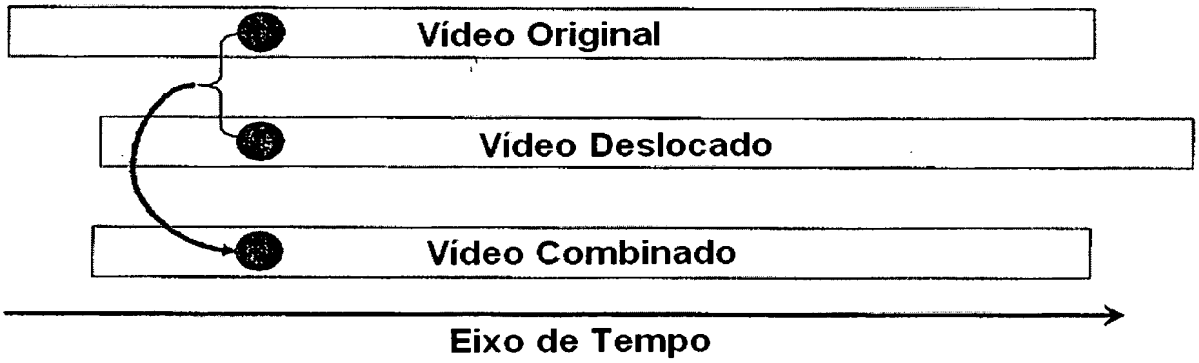


FIG. 10

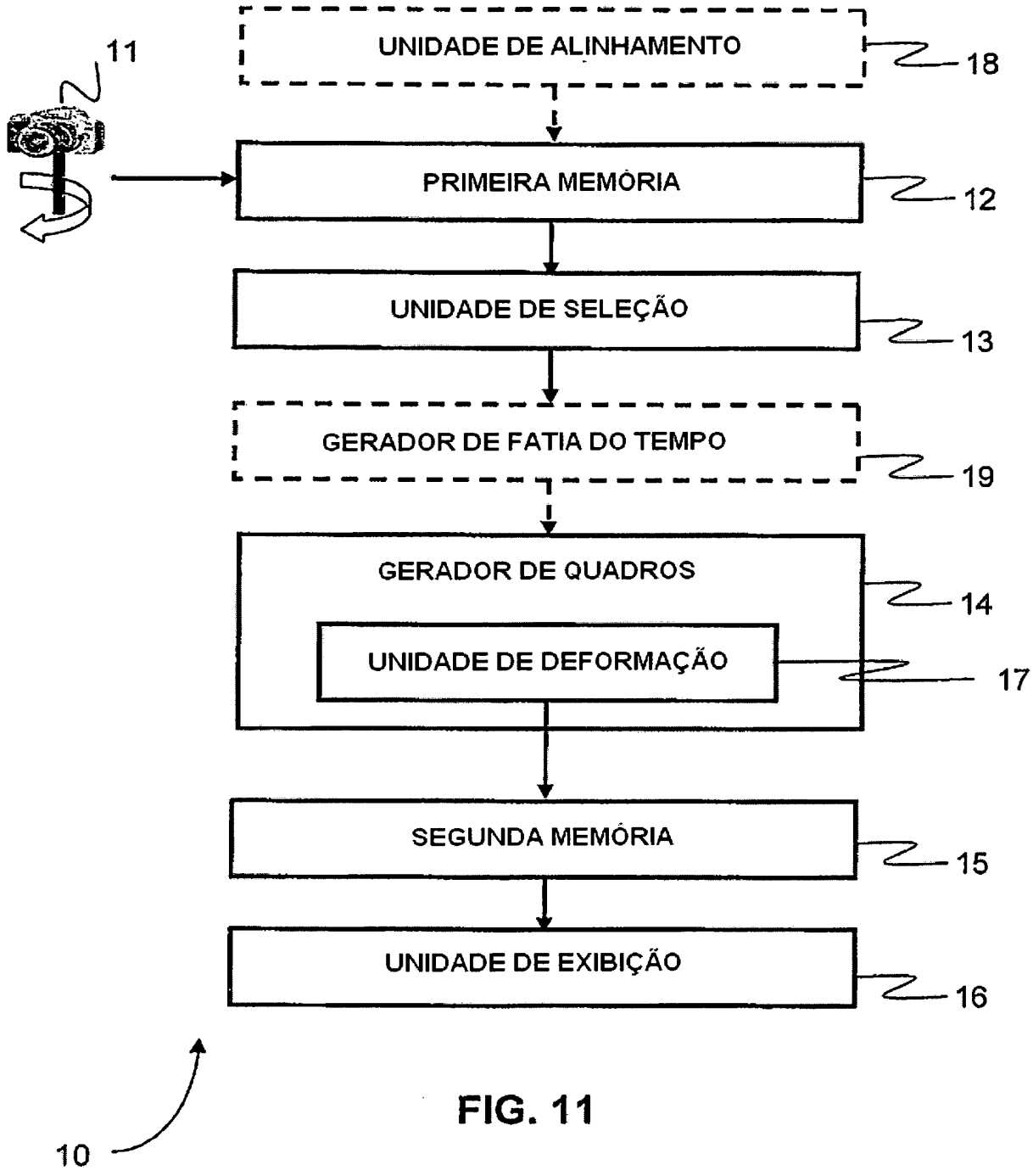

11
UNIDADE DE ALINHAMENTO
18
PRIMEIRA MEMÓRIA
12
UNIDADE DE SELEÇÃO
13
GERADOR DE FATIA DO TEMPO
19
GERADOR DE QUADROS
14
UNIDADE DE DEFORMAÇÃO
17
SEGUNDA MEMÓRIA
15
UNIDADE DE EXIBIÇÃO
16
10

FIG. 11

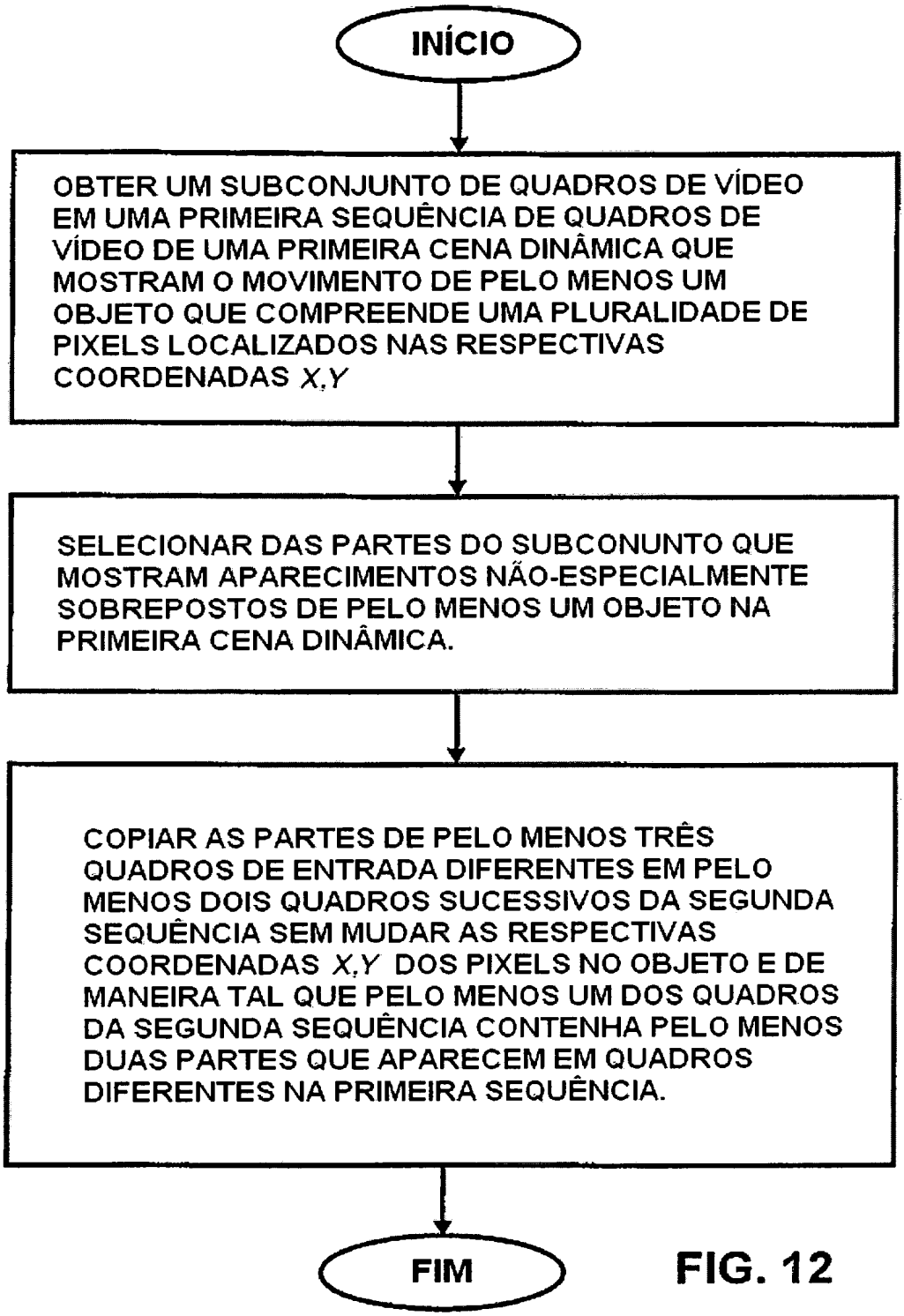

INÍCIO
OBTER UM SUBCONJUNTO DE QUADROS DE VÍDEO EM UMA PRIMEIRA SEQUÊNCIA DE QUADROS DE VÍDEO DE UMA PRIMEIRA CENA DINÂMICA QUE MOSTRAM O MOVIMENTO DE PELO MENOS UM OBJETO QUE COMPREENDE UMA PLURALIDADE DE PIXELS LOCALIZADOS NAS RESPECTIVAS COORDENADAS X,Y
SELECIONAR DAS PARTES DO SUBCONUNTO QUE MOSTRAM APARECIMENTOS NÃO-ESPECIALMENTE SOBREPOSTOS DE PELO MENOS UM OBJETO NA PRIMEIRA CENA DINÂMICA.
COPIAR AS PARTES DE PELO MENOS TRÊS QUADROS DE ENTRADA DIFERENTES EM PELO MENOS DOIS QUADROS SUCESSIVOS DA SEGUNDA SEQUÊNCIA SEM MUDAR AS RESPECTIVAS COORDENADAS X,Y DOS PIXELS NO OBJETO E DE MANEIRA TAL QUE PELO MENOS UM DOS QUADROS DA SEGUNDA SEQUÊNCIA CONTENHA PELO MENOS DUAS PARTES QUE APARECEM EM QUADROS DIFERENTES NA PRIMEIRA SEQUÊNCIA.
FIM

FIG. 12

PI0620497-0

RESUMO

MÉTODO PARA A CRIAÇÃO DE UMA SINOPSE DE VÍDEO, SISTEMA PARA TRANSFORMAR UMA SEQÜÊNCIA DE ORIGEM DE QUADROS DE VÍDEO DE UMA PRIMEIRA CENA DINÂMICA EM UMA SEQÜÊNCIA DE SINOPSE DE PELO MENOS DOIS QUADROS DE VÍDEO QUE ILUSTRAM UMA SEGUNDA CENA DINÂMICA, E, PRODUTO DE PROGRAMA DE COMPUTADOR

Trata-se de um método e um sistema implementado por computador que transformam uma primeira seqüência de quadros de vídeo de uma primeira cena dinâmica em uma segunda seqüência de pelo menos dois quadros de vídeo que apresentam uma segunda cena dinâmica. Um subconjunto de quadros de vídeo na primeira seqüência é obtido, o qual mostra o movimento de pelo menos um objeto que tem uma pluralidade de pixels localizados nas respectivas coordenadas *x,y* e são selecionadas partes do subconjunto que mostram os aparecimentos não-espacialmente sobrepostos de pelo menos um objeto na primeira cena dinâmica. As partes são copiadas de pelo menos três quadros de entrada diferentes em pelo menos dois quadros sucessivos da segunda seqüência sem mudar as respectivas coordenadas *x,y* dos pixels no objeto e de maneira tal que pelo menos um dos quadros da segunda seqüência contenha pelo menos duas partes que aparecem em quadros diferentes na primeira seqüência.