República Federativa do Brasil
Ministério da Economia
Instituto Nacional da Propriedade Industrial

(11) PI 1011189-1 B1

(22) Data do Depósito: 07/05/2010

(45) Data de Concessão: 24/08/2021

(54) Título: SISTEMA BASEADO EM COMPUTADOR PARA SELECIONAR PONTOS DE VISUALIZAÇÃO ÓTIMOS E MEIO DE ARMAZENAMENTO DE SINAL LEGÍVEL POR MÁQUINA NÃO TRANSITÓRIA

(51) Int.Cl.: H04N 5/262; H04N 5/268.

(30) Prioridade Unionista: 07/05/2009 GB 09 07870.0.

(73) Titular(es): UNIVERSITE CATHOLIQUE DE LOUVAIN.

(72) Inventor(es): CHRISTOPHE DE VLEESCHOUWER; FAN CHEN.

(86) Pedido PCT: PCT BE2010000039 de 07/05/2010

(87) Publicação PCT: WO 2010/127418 de 11/11/2010

(85) Data do Início da Fase Nacional: 07/11/2011

(57) Resumo: SISTEMAS E MÉTODOS PARA A PRODUÇÃO AUTÔNOMA DE VÍDEOS A PARTIR DE DADOS MULTIDETECTADOS. A presente invenção refere-se a um método e a um sistema baseado em computador autônomo para a produção personalizada de vídeos tais como vídeos esportivos, como video limitada. Modalidades da presente invenção referem-se à seleção de uma visualização para exibição a partir dos múltiplos fluxos de vídeo capturados pela rede de câmeras. São providas soluções técnicas para proporcionar um conforto perceptivo bem como uma integração eficiente de informação contextual, a qual é implementada, por exemplo, pela suavização de sequências geradas de pontos de visualização de câmera para suavizar os artefatos visuais cintilantes e artefatos de contar histórias descontínuos. Um projeto e uma implementação de processos de seleção de ponto de visualização são descritos, que têm sido verificados por experimentação, que mostram que o método e o sistema da presente invenção distribuem de forma eficiente a carga de processamento através das câmeras, e eficientemente seleciona pontos de visualização que cobrem a ação do time disponível enquanto que evita artefatos perceptivos maiores.

Relatório Descritivo da Patente de Invenção para **"SISTEMA BASEADO EM COMPUTADOR PARA SELECIONAR PONTOS DE VISUALIZAÇÃO ÓTIMOS E MEIO DE ARMAZENAMENTO DE SINAL LEGÍVEL POR MÁQUINA NÃO TRANSITÓRIA".**

Campo da Invenção

[0001] A presente invenção refere-se à integração de informações de múltiplas câmeras em um sistema de vídeo, por exemplo, uma produção de televisão ou sistema de vigilância inteligente e para a produção automática de conteúdo de vídeo, por exemplo, para tornar uma ação que envolve uma ou diversas pessoas e/ou objetos de interesse.

Antecedentes da Técnica

[0002] O projeto APIDIS (Produção Autônoma de Imagens com base em Detecção Inteligente e Distribuída) tenta fornecer uma solução para gerar conteúdos personalizados para representação visual melhorada e de baixo custo de cenários controlados tal como televisão de esportes, em que a qualidade de imagem e o conforto de percepção são tão essenciais quanto a integração eficiente de informações contextuais [I].

[0003] No contexto APIDIS, múltiplas câmeras são distribuídas em redor da ação de interesse, e a produção autônoma de conteúdo envolve três questões técnicas principais com respeito a tais câmeras:

(i) como selecionar pontos de visualização ideais, isto é, parâmetros de enquadramento em uma dada câmera, de modo que estes sejam ajustados para uma resolução de exibição limitada,

(ii) como selecionar a câmera certa para renderizar a ação em um dado momento, e

(iii) como suavizar as sequências de câmera/de ponto de visualização para remover os artefatos de produção.

[0004] Os artefatos de produção consistem tanto em artefatos

visuais, o que pode significar principalmente efeitos de tremulação devido a estremecimentos ou a aumento/diminuição de aproximação rápida de pontos de visualização, quanto em artefatos de contar histórias tal como a descontinuidade da história causada pela mudança rápida de câmera e movimentos dramáticos de ponto de visualização.

[0005] A fusão de dados de múltiplas câmeras foi amplamente discutida na literatura. Esses trabalhos anteriores podem ser, em termos gerais , classificados em três categorias principais de acordo como os vários propósitos dos mesmos. Os métodos na primeira categoria lidam com calibragem de câmera e câmera inteligente controladas pelas informações contextuais de integração do ambiente de câmeras múltiplas [4]. A reconstrução da cena 3D [5] ou síntese de vídeo de ponto de visualização arbitrário [2] de múltiplas câmeras também é um tópico popular. A terceira categoria usa múltiplas câmeras para resolver certos problemas tal como a oclusão em várias aplicações, por exemplo, rastreamento de pessoas [6]. Todos esses trabalhos se focam bastante na extração de informações contextuais 3D importantes, mas com pouca consideração nas questões da técnica mencionadas acima sobre produção de vídeo.

[0006] No que diz respeito à produção de vídeo autônoma, existem alguns métodos propostos na literatura para selecionar a área mais representativa de uma imagem independente. Suh et al.[7] definiram a região de enquadramento ideal como no retângulo mínimo que continha uma saliência através de um dado limiar, em que a saliência foi computada pelo modelo de atenção visual [8]. Na Referência [9], outro método com base no modelo de atenção foi proposto, em que foi mais discutida uma trajetória ideal de mudança de atenção do que a decisão de ponto de visualização. Sabe-se também sobre a exploração de uma rede distribuída de câmeras para aproximar as imagens que seriam capturadas por um sensor virtual localizado em uma posição arbitrária,

com cobertura de ponto de visualização arbitrária. Para poucas câmeras com lentes inteiramente heterogêneas e cobertura de cena, a maioria dos métodos de última geração de ponto de visualização livre produzem resultados embaçados [2] [3].

[0007] Na Referência [10], um sistema de produção automática para vídeos de esporte de futebol são propostos e a seleção de ponto de visualização com base no entendimento da cena também foi discutida. No entanto, esse sistema apenas troca pontos de visualização entre três tamanhos de captura fixos de acordo com diversas regras fixas, o que leva a artefatos visuais desconfortáveis devido à mudança dramática de tamanho de captura. Ademais, estes são apenas discutidos no caso de câmera única.

[0008] Adicionalmente à pesquisa da literatura acima, diversos pedidos de patente consideraram sistemas de câmeras múltiplas (multidirecional) para produzir e editar o conteúdo de vídeo de uma maneira semiautomática. Três categorias principais de sistemas podem ser identificadas.

[0009] A primeira categoria seleciona uma visualização (isto é, um vídeo) entre aquelas cobertas por um conjunto de câmeras predefinido, com base em algum mecanismo de detecção de atividade. Em [15], cada câmera é ativada com base em algum dispositivo externo, que desencadeia a aquisição de vídeo cada vez que um evento em particular é detectado (por exemplo, um objeto que entra no campo de visualização). Em [16], sensores de áudio são usados para identificar a direção em que o vídeo deveria ser capturado. A segunda categoria captura um sinal visual rico, ou com base em câmeras multidirecionais ou em definição de câmeras múltiplas de ângulo amplo, de modo a oferecer alguma flexibilidade na maneira que a cena é renderizada na extremidade receptora. Por exemplo, os sistemas em [17] e [18] consideram respectivamente câmeras múltiplas e sistemas de

visualização multidirecional para capturar e difundir fluxos de vídeo de ângulo amplo. Em [17], uma interface permite que o espectador monitore o(s) fluxo(s) de vídeo de ângulo amplo para selecionar que parte do vídeo abrir em tempo real. Ademais, o operador pode parar a reprodução e controlar efeitos de aproximação de movimento horizontal/vertical em um quadro particular. Em [18], a interface é melhorada com base na detecção automática de áreas do vídeo em que o participante de um evento esteja presente. Por conseguinte, o espectador tem a oportunidade de escolher de modo interativo para qual participante do evento gostaria de olhar.

[00010] De maneira similar, [19 a 21] detectar pessoas de interesse em uma cena (tipicamente um palestrante ou a participante de videoconferência). No entanto, a melhoria através de [18] é dupla. Primeiramente, em [19 a 21], métodos são propostos para definir automaticamente um conjunto de projeções candidatas com base em análise automática da cena. A seguir, mecanismos são definidos para selecionar automaticamente uma captura entre as projeções candidatas. Em [19], a definição de captura depende na detecção e rastreamento do palestrante, e regras probabilísticas são usadas para trocar pseudoaleatoriamente da câmara da audiência para a câmera de palestrante durante uma palestra. Em [20] e [21], uma lista de projeções candidatas também é definida com base na detecção de algum objeto de interesse em particular (tipicamente um rosto), mas efeitos de edição mais sofisticados são considerados para criar uma renderização dinâmica (videoconferência). Por exemplo, uma captura pode mudar de uma pessoa para a outra, ou diversos rostos podem ser colados próximos um do outro em uma captura única. A saída de vídeo editada é assim construída pela seleção de uma captura melhor entre as projeções candidatas para cada cena (em [20] e [21], a cena corresponde a um período de tempo particular). A melhor captura é

selecionada com base em um conjunto predefinido de regras cinematográficas, por exemplo, para evitar muitas da mesma captura em uma sequência.

[00011] É digno de nota que os parâmetros de captura (isto é os parâmetros de enquadramento na visualização atual) permanecem fixos até que a câmera seja trocada. Ademais, em [19 a 21] a captura é diretamente associada com um objeto, de modo que no final, a seleção de captura termina na seleção do(s) objetos(s) para renderizar, que pode ser difícil e irrelevante em contextos que sejam mais complexos que uma videoconferência ou uma palestra. Especificamente, [19 a 21] não selecionam a captura com base no processamento conjunto das posições de objetos múltiplos. A terceira e última categoria de sistemas de produção de vídeo semiautomáticos diferencia as câmeras que são dedicadas a análise da cena daquelas que são usadas para capturar as sequências de vídeo. Em [22], uma rede de câmeras é usada para propósitos de análise da cena de esporte. As saídas do módulo de análise são então exploradas para computar estatísticas sobre o jogo, mas também para controlar as câmeras de aproximação de ajuste vertical/horizontal que coletam vídeos dos jogadores de interesse (tipicamente aquele que segura o disco ou a bola). [22] precisa implantar toda a análise dos algoritmos de cena em tempo real, desde que esse visa controlar os parâmetros PTZ da câmera instantaneamente, como uma função da ação observada na cena. Mais importante e fundamentalmente, [22] seleciona os parâmetros PTZ para capturar objetos específicos detectados e não para oferecer renderização apropriada de uma ação de tempo, potencialmente composta de objetos múltiplos de interesse. Nisto é similar a [19 a 21]. Da mesma forma, quando vídeos múltiplos são coletados, [22] não fornece qualquer solução para selecionar um deles. Este apenas encaminha todos os vídeos para uma interface que os apresenta de uma maneira integrada

para um operador humano. Isso é a fonte de uma obstrução quando muitas câmeras fonte são consideradas.

[00012] O documento US2008/0129825 apresenta o controle de câmera motorizada para capturar imagens de objetos individuais rastreados, por exemplo, para esportes individuais como competições de atletismo. O usuário seleciona a câmera através de uma interface de usuário. As unidades de localização são fixadas aos objetos. Por conseguinte estes são intrusivos.

[00013] O documento GB2402011 apresenta um controle de câmera mecanizado que usa parâmetros de evento. Com base no rastreamento de jogador e um conjunto de regras de desencadeamento, o campo de visualização de câmera é adaptado e trocado entre visualização próxima, média, distante. Uma câmera é selecionada com base nos eventos desencadeadores. Um evento desencadeador tipicamente corresponde a ações ou movimentos específicos de esportistas, por exemplo, o serviço de um jogador de tênis, ou as atualizações das informações de placar.

[00014] O documento US2004/0105004A1 refere-se à renderização de discussões ou encontros. As câmeras de rastreamento são exploradas para renderizar o apresentador ou um membro da audiência que pergunta uma questão. O apresentador e os membros de audiência são rastreados com base na localização de fonte de som, que usa uma matriz de microfones. Dada a posição do alvo da câmera de rastreamento, os parâmetros PTZ da câmera motorizada são controlados de forma a fornecer um vídeo editado suave do alvo. O método descrito e o sistema são adequados apenas para seguir uma única pessoa individual. Com respeito à seleção da câmera, a troca é apresentado entre um conjunto de visualizações muito distintas (uma visualização geral do ambiente, uma visualização das lâminas, uma visualização próxima do apresentador, e uma visualização próxima um

membro falante de uma audiência). O processo de seleção de câmera é controlado com base na detecção de evento (por exemplo o aparecimento de uma nova lâmina, ou um membro falante de uma audiência) e regras de videografia definida por professionais, para emular um tempo humano de produção de vídeo.

REFERÊNCIAS

[1] Página do projeto APIDIS. *http://www.apidis.org/*

Demo vídeos relacionados a este artigo: *http://www.apidis.org/Initial Results/APIDISP%20Initial%20Results.htm*

[2] S. Yaguchi, and H. Saito, Arbitrary viewpoint video synthesis from multiple uncalibrated cameras, IEEE Trans. Syst. Man. Cybern. B, 34(2004)

[3] N. Inamoto, and H. Saito, Free viewpoint video synthesis and presentation from multiple sporting videos, Electronics and Communications in Japan (Part III: Fundamental Electronic Science), 90(2006) 40-49.

[4] LH. Chen, and SJ. Wang, An efficient approach for the calibration of multiple PTZ cameras, IEEE Trans. Automation Science and Engineering, 4(2007) 286-293.

[5] P. Eisert, E. Steinbach, and B. Girod, Automatic reconstruction of stationary 3-D objects from multiple uncalibrated camera views, IEEE

Trans. Circuits and Systems for Video Technology, Special Issue on 3D Video Technology, 10(1999) 261-277.

[6] A. Tyagi, G. Potamianos, J. W. Davis, and S.M. Chu, Fusion of Multiple camera views for kernel-based 3D tracking, WMVC'07, 1(2007) 1-1.

[7] B. Suh, H. Ling, B. B. Bederson, and D.W. Jacobs, Automatic thumbnail cropping and its effectiveness, Proc. ACM UIST 2003, 1(2003) 95-104.

[8] L. Itti, C. Koch, and E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, 20(1998) 1254-1259.

[9] X. Xie, H. Liu, W. Y. Ma, H.J. Zhang, "Browsing large pictures under limited display sizes, IEEE Trans. Multimedia, 8(2006) 707-715.

[10] Y. Ariki, S. Kubota, and M. Kumano, Automatic production system of soccor sports video by digital camera work based on situation recognition, ISM'06, 1(2006) 851-860.

[11] J. Owens, Television sports production, 4th Edition, Focal Press, 2007.

[12] J. W. Gibbs, Elementary principles in statistical mechanics, Ox Bow Press, 1981.

[13] D. Chandler, Introduction to modern statistical mechanics, Oxford University Press, 1987.

[14] C. De Vleeschouwer, F. Chen, D. Delannay, C. Parisot, C. Chaudy, E. Martrou, and A. Cavallaro, Distributed video acquisition and annotation for sport-event summarization, NEM summit, (2008).

[15] Documento EP1289282 (Al) Video sequence automatic production method and system Inventor: AYER SERGE [CH] ; MOREAUX MICHEL [CH] (+1); Applicant: DARTFISH SA [CH]; EC: H04N5/232 IPC: H04N5/232; H04N5/232;; (IPC 1-7): H04N5/232

[16] Documento US20020105598, EP1352521 AUTOMATIC MULTI-CAMERA VIDEO COMPOSITION; INTEL CORP

[17] Documento US6741250 Method and system for generation of multiple viewpoints into a scene viewed by motionless cameras and for presentation of a view path; BE HERE CORP

[18] US20020191071 Automated online broadcasting system and method using an omni-directional camera system for viewing meetings over a computer network; MICROSOFT CORP

[19] Documento US20020196327 Automated video production system and method using expert video production rules for online publishing of lectures; MICROSOFT CORP; Microsoft Corporation

[20] Documento US20060251382 Al System and method for automatic video editing using object recognition MICROSOFT CORP

[21] Documento US20060251384 Automatic video editing for real-time multi -point video conferencing; MICROSOFT CORP

[22] Documento WO200599423 AUTOMATIC EVENT VIDEOING, TRACKING AND CONTENT GENERATION SYSTEM; AMAN JAMES A; BENNETT PAUL MICHAEL

Aspectos da Presente invenção

[00015] O objetivo da presente invenção é o de fornecer computadores com métodos e sistemas de base para a produção autônoma de um vídeo editado, composto com base nos fluxos múltiplos de vídeo capturados por uma rede de câmeras, distribuídas em redor da cena de interesse.

[00016] A presente invenção fornece um computador autônomo com base em no método e no sistema para a produção personalizada de vídeos tal como vídeos de times de esporte tal como vídeos de basquetebol de dados de sensores múltiplos sob resolução de exibição limitada. No entanto a invenção tem uma faixa de aplicação mais ampla e não é limitada apenas para esse exemplo. As modalidades da presente invenção relacionam-se a seleção de uma visualização para exibir entre os fluxos múltiplos de vídeo capturados pela rede de câmeras. As soluções da técnica são fornecidas para fornecer conforto de percepção assim como uma integração eficiente das informações contextuais, que são implantadas, por exemplo, pela suavização gerada para as sequências de ponto de visualização/câmera para aliviar artefatos visuais tremeluzentes e artefatos de contar histórias

descontínuos. Um projeto e uma implantação do processo de seleção de ponto de visualização são apresentados os quais foram verificados por experimentos, que mostram que o método e o sistema da presente invenção distribuem eficientemente a carga de processamento através das câmeras, e selecionam efetivamente pontos de visualização que cubram a ação de equipe próxima enquanto evita artefatos de percepção principais.

[00017] Consequentemente a presente invenção fornece um computador com base método para a produção autônoma de um vídeo editado dos fluxos múltiplos de vídeo capturados por uma pluralidade de câmeras distribuídas em redor de uma cena de interesse, o método compreende:

- detectar objetos/pessoas de interesse nas imagens dos fluxos de vídeo, por exemplo, sabendo as coordenadas mundiais em 3D atuais,

- selecionar para cada câmera o campo de visualização que renderiza a cena de interesse de uma maneira que (permita que o espectador) siga a ação executada pelos múltiplos objetos/pessoas que estão interagindo que tenham sido detectados. O parâmetros de campo de visualização se referem, por exemplo, a janela de enquadramento em uma câmera estática, e/ou para os parâmetros de posição e aproximação de ajuste horizontal/vertical em uma câmera motorizada e móvel. O conceito de ação seguinte pode ser quantificado pela medição da quantidade de pixels associados a cada objeto/pessoa de interesse na imagem exibida. Acompanhamento preciso da ação resulta em renderização completa e próxima, em que a conta de totalização da quantidade de objetos/pessoas na imagem exibida, enquanto medição de proximidade a quantidade de pixels disponível para descrever cada objeto.

- montar o vídeo editado pela seleção e concatenação de

segmentos vídeo fornecidos por um ou mais câmeras individuais, de uma maneira que maximize a métrica de totalização e a proximidade de uma maneira que maximize a métrica de totalização e a proximidade, enquanto suaviza a sequência de parâmetros de renderização associados aos segmentos concatenados.

[00018] A seleção de parâmetros de renderização pode ser objetivos ou objetivos de interesse simultaneamente. O conhecimento sobre a posição dos objetos nas imagens pode ser explorado para decidir como renderizar uma ação capturada. O método pode selecionar parâmetros de campo de visualização para a câmera que renderiza a ação como uma função de tempo com base em um equilíbrio ideal entre as métricas de proximidade e de totalização. Ou como exemplo, os parâmetros de campo de visualização de referem ao enquadramento na visualização da câmera de imagens estáticas ou para os parâmetros de aproximação ou deslocamento de ajuste vertical/horizontal para dinamizar e potencializar as câmeras móveis.

[00019] As métricas de proximidade e de totalização podem ser adaptadas de acordo com as preferências de usuário e/ou os recursos. Por exemplo, um recurso de usuário pode ser a resolução de codificação. A preferência de usuário pode ser pelo menos um dos objetos preferenciais, ou da câmera preferencial. Imagens de todas as visualizações de todas as câmeras podem ser mapeadas para as mesmas coordenadas temporais absolutas com base em referências comuns temporais únicas para todas as visualizações de câmera. Em cada instante, e para cada visualização de câmera, parâmetros de campo de visualização são selecionados de modo que melhorem a troca entre a totalização e a proximidade. O ponto de visualização selecionado em cada visualização de câmera pode ser classificada de acordo com a qualidade da troca de totalização/proximidade, e para o grau de suas oclusões. Para o segmento temporal atual, os parâmetros

de uma câmera virtual ideal que ajusta, aproxima e troca através de visualizações que podem ser trocadas para conservar uma alta avaliação dos pontos de visualização selecionados e minimizando a quantidade de movimentos de virtual câmera.

[00020] O método pode incluir a seleção do campo de visualização ideal em cada câmera, em um dado instante.

[00021] Um campo de visualização v_k na k-ésima visualização de câmera é definida pelo tamanho S_k e o centro c_k da janela que é enquadrada na k-ésima visualização para exibição real. É selecionado para incluir os objetos de interesse e fornecer uma alta descrição de resolução dos objetos, e um campo de visualização ideal v_k^* é selecionado para maximizar uma soma pesada de objetos de interesse como se segue:

$$\mathbf{v}_k^* = \arg\max_{\{S_k, \mathbf{c}_k\}} \sum_{n=1}^{N} I_n \cdot \alpha(S_k, \mathbf{u}) \cdot m\left(\frac{\left\| \mathbf{x}_{n,k} - \mathbf{c}_k \right\|}{S_k} \right)$$

em que, na equação acima:

- I_n denota o nível de interesse assinalado para os n-ésimos objetos detectados na cena.

- $x_{n,k}$ denota a posição dos n-ésimos objetos na visualização de câmera k.

- A função m(...) modula o peso dos n-ésimos objetos de acordo com a distância do centro da janela de ponto de visualização, comparado ao tamanho da janela.

- O vetor u reflete as preferências de usuário, em particular, o componente u_{res} define a resolução do fluxo de saída, que é geralmente limitado pela largura de banda de construção ou resolução de dispositivo de usuário final.

- A função α(.) reflete a penalidade induzida pelo fato de que um sinal capturado nativo pela k-ésima câmera que tem que ser

subtestada uma vez que o tamanho do ponto de visualização se torne maior que a resolução máxima u_{res} permitida pelo usuário.

[00022] Preferencialmente α(....) diminui com S_k e a função α(....) é igual a um quando $S_k < u_{res}$, e diminui em seguida. α(....) é definido por:

$$\alpha(S,\mathbf{u}) = \left[\min\left(\frac{u_{res}}{S},1\right)\right]^{u_{próximo}},$$

em que o exponente u_{close} é maior que 1, e aumenta conforme a preferência de usuário de renderização de resolução total de aproximação na área, comparado aos pontos de visualização grandes mas sub testados.

[00023] O método inclui avaliação do ponto de visualização associado a cada câmera de acordo com a qualidade da troca de totalização/proximidade, e para o grau das oclusões. A taxa mais alta deveria corresponder a uma visualização que (1) torna a maioria do objeto de interesse visível, e (2) é próximo à ação, significando que apresenta objetos importantes com muitos detalhes, isto é uma resolução alta. Formalmente, dado ao n-ésimo interesse de cada jogador, a taxa $h_k(X_n)$ associada com a k-ésima visualização de câmera é definido como se segue:

$$I_k(\mathbf{v_k},\mathbf{u}) = \sum_{n=1}^{N} I_n \cdot o_k(\mathbf{x}_n|\bar{\mathbf{x}}) \cdot h_k(\mathbf{x}_n) \cdot \beta_k(S_k,\mathbf{u}).m\left(\frac{\|\mathbf{x}_{n,k} - \mathbf{c}_k\|}{S_k}\right)$$

em que, a equação acima:

- I_n denota o nível de interesse atribuído ao n-ésimo objeto detectado na cena.

- x_n denota da posição do n-ésimo objeto no espaço 3D;

- $o_k(x_{nl}\ x)$ mede a razão de oclusão do n-ésimo objeto na visualização de câmera k, sabendo que a posição de todos os outros

objetos, a razão de oclusão de um objetos que é definido para ser a fração de pixels dos objetos que são escondidos por outros objetos quando projetados no sensor de câmera;

- A altura $h_k(x_n)$ é definida para ser a altura nos pixels da projeção na visualização k de uma altura de referência de objetos de referência localizados em X_n. O valor de $h_k(x_n)$ é diretamente computado com base na calibragem de câmera, ou quando a calibragem não está disponível, esta pode ser estimada com base na altura dos objetos detectados na visualização k.

- A função $\beta_k(.)$ reflete o impacto das preferências de usuário em termos de visualização de câmera e resolução de exibição. $\beta_k(.)$ é definido como

$$\beta_k(S,\mathbf{u}) = u_k \cdot \alpha(S,\mathbf{u}),$$

em que u_k denota o peso designado para a k-ésima câmera, e é definido acima.

[00024] O método pode compreender a suavização da sequência de índices de câmera e os parâmetros de ponto de visualização correspondentes, em que o processo de suavização é, por exemplo, implantado com base nos dos Campos Aleatórios de Markov, mecanismo de filtragem de baixa passagem linear ou não linear, ou através de um formalismo de modelo gráfico, resolvido com base no algoritmo de Viterbi convencional.

[00025] A captura de múltiplos fluxos de vídeo pode ser câmeras estáticas ou dinâmicas.

[00026] A presente invenção também inclui um computador com base no sistema que compreende um mecanismo de processamento e memória para a produção autônoma de um vídeo editado de fluxos múltiplos de vídeo capturados por uma pluralidade de câmeras distribuída ao redor a cena de interesse, adaptado para executar

qualquer dos métodos da presente invenção. O sistema pode compreender: um detector para detectar objetos nas imagens dos fluxos de vídeo, significa primeiro para selecionar uma ou mais câmeras de ponto de visualização com base no processamento conjunto de posições de objetos múltiplos que tenham sido detectados, e também significa selecionamento de parâmetros de renderização que maximizem e suavizem as métricas de proximidade e de totalização pela concatenação de segmentos nos fluxos de vídeo fornecidos por uma ou mais câmeras individuais.

[00027] O computador com base no sistema pode ter

- meios para detectar objetos/pessoas de interesse nas imagens dos fluxos de vídeo, por exemplo, sabendo as coordenadas mundiais em 3D atuais,

- meios para selecionar cada câmera de campo de visualização que renderizam a cena de interesse de uma maneira que (permita que o espectador) siga a ação executada pelos múltiplos objetos/pessoas que estão interagindo que tenham sido detectados. Os parâmetros de campo de visualização se referem, por exemplo, a janela de enquadramento em uma câmera estática, e/ou a horizontal-inclinação-ampliação (*zoom*) e parâmetros de posição em uma câmera motorizada e móvel. O conceito de ação seguinte pode ser quantificada pela medição da quantidade de pixels associados a cada um(a) do(a)s objetos/pessoas de interesse na imagem exibida. O acompanhamento preciso da ação resulta de renderização completa e próxima, em que a totalização conta o número de objetos/pessoas na imagem exibida, enquanto a proximidade mede a quantidade de pixels disponível para cada objeto,

- meios para montar o vídeo editado pela seleção e a concatenação de segmentos de vídeo fornecidos por uma ou mais câmeras individuais, de uma maneira que maximize a métrica de

totalização e aproxime ao longo do tempo, enquanto a suaviza a sequência de parâmetros de renderização associados a segmentos concatenados.

[00028] A presente invenção também fornece um produto de programa de computador que compreende segmentos de código que quando executados em um mecanismo de processamento executam qualquer métodos da invenção ou implantação de qualquer sistema de acordo com a invenção.

[00029] A presente invenção também inclui a meio de armazenagem de sinal legível por máquina não transitória armazenando o produto de programa de computador.

[00030] A presente invenção pode lidar com as cenas que envolvem diversas pessoas se movendo e interagindo/objetos de interesse. No seguinte, as cenas são denotadas como ações de time, e tipicamente correspondem a cenas encontradas no contexto de esportes de time.

[00031] A automatização de processos de produção permite:

- reduzir os custos de produção, evitando processos manuais longos e tediosos, tanto para controle de câmera e seleção de câmera;

- aumentar a produção e qualidade de largura de banda, por manusear potencialmente um número infinito de câmeras simultaneamente;

- criar conteúdo personalizado, pelo processo de produção diversas vezes, com parâmetros distintos.

[00032] Um objetivo da presente invenção tem como alvo a produção de semanticamente significante, isto é mostrar a ação de interesse, e conteúdos confortáveis perceptualmente de dados multidetectados brutos. O sistema de acordo com a presente invenção é um computador com base que inclui uma memória e um mecanismo de processamento e que é um sistema de produção eficiente computacionalmente, por exemplo, com base em um paradigma de dividir e conquistar (vide a Fig.

15).

[00033] Em modalidades, o melhor campo de visualização é primeiramente computado para cada câmera individual, e então a melhor câmera para renderizar a cena é selecionada. Junto com o índice de câmera e o campo de visualização define o ponto de visualização para renderizar a ação. Quando a câmera é fixada, a definição de campo de visualização é limitada ao enquadramento da imagem capturada pela câmera. Quando a câmera é motorizada, o campo de visualização resulta diretamente dos parâmetros da câmera de horizontal-inclinação-ampliação, e pode assim capturar uma parte retangular arbitrária do campo de luz que alcança o centro da câmera.

[00034] Para definir de uma maneira quantitativa a noção do melhor campo de visualização ou do melhor índice câmera, a presente invenção introduz três conceitos importantes, que são "totalização", "proximidade" e "suavização". A Totalização significa a totalidade da ação de renderização. No contexto da ação de renderização de tempo, a totalização mede que bem os objetos/as pessoas de interesse na cena (tipicamente os jogadores que participam de um time de esportes) são incluídos na imagem exibida. A Proximidade define a riqueza de descrição de detalhes (tipicamente quantidade média de pixels que estão disponíveis para renderizar as pessoas/objetos de interesse), e a suavização é um termo que se refere a continuidade de seleção de ponto de visualização . Pela troca entre estes fatores, os métodos são fornecidos para selecionar (como uma função de tempo) pontos de visualização ideais para caber a resolução de exibição e outras preferências de usuário, e para a suavização dessas sequências para contar as histórias continuamente e com elegância. A presente invenção é completamente autônoma e se governa por si só, no sentido de que esta pode selecionar os pixels para exibir sem qualquer intervenção humana, com base em um conjunto padrão de parâmetros de produção

e no resultado dos sistemas de detecção. Mas a invenção também pode lidar com as preferências de usuário, tais como perfil narrativo de usuário, e capacidades do dispositivo. As preferências de narrativa podem ser resumidas em quatro descrições, isto é, preferência de usuário por grupo de objetos ou "time", preferência de usuário por objetos ou "jogador", preferência de usuário por "tipo de visualização" (por exemplo, visualizações de aproximação de ampliação ou redução), e preferência de usuário por "câmera". Todas as restrições de dispositivo, tal como resolução de exibição, velocidade da rede, desempenho de decodificação, são abstraídas como o parâmetro de resolução de saída, que denota a resolução na qual o vídeo de saída é codificado para ser conduzida e exibida no hospedeiro final.

[00035] A capacidade de levar em consideração essas preferências depende do conhecimento capturado à cerca da cena, por exemplo, através de ferramentas de análise de vídeo. Por exemplo, uma modalidade da presente invenção foi implantada em "Detection and Recognition of Sports(wo)men from Multiple Views", D. Delannay, N. Danhier e C. De Vleeschouwer, Third ACM/IEEE International Conference on Distributed Smart Cameras, Como, Itália em setembro de 2009 para rastrear automaticamente e reconhecer os jogadores em movimento na cena de interesse. Esse documento está incluso como Apêndice 2.

[00036] Em primeiro lugar, em modalidades da presente invenção um conjunto de câmeras que cobrem (parcialmente) a mesma área é considerado, o qual tem a possibilidade de ser ativado simultaneamente com base em qualquer mecanismo de detecção de atividade o que é outra importante vantagem da presente invenção sobre a técnica antecedente. O propósito da invenção não é, portanto selecionar uma visualização de câmera com base no fato de que alguma atividade foi detectada na visualização, mas sim selecionar ao longo do tempo da

visualização de câmera e suas variações correspondentes em parâmetros como reenquadramento ou parâmetros PTZ, para melhor renderizar a ação que ocorre na área coberta. Nestas, a qualidade de renderização refere-se à otimização de uma compensação entre medidas de proximidade, totalidade e suavidade.

[00037] Em segundo lugar, a presente invenção tem uma vantagem de se adaptar dinamicamente suavizar os parâmetros de ponto de visualização com o tempo, o que é um aprimoramento sobre os sistemas da técnica antecedente na qual os parâmetros de captura (por exemplo, os parâmetros de reenquadramento na visualização disponível) ficam fixos até que a câmera seja trocada.

[00038] Em terceiro lugar, nas modalidades da presente invenção uma escolha entre um objeto ou outro não é feita, mas sim uma seleção é feita do ponto de visualização com base no processamento conjunto das posições dos múltiplos objetos que foram detectados. De acordo com as modalidades da presente invenção, uma seleção é feita da sequência pontos de visualização que é ideal pela maneira que maximiza e suaviza as métricas de proximidade e totalidade, por exemplo, para todos os objetos simultaneamente.

[00039] Essas diferenças comparadas à técnica antecedente trazem benefícios significativos em relação ao problema de produção de conteúdo, por exemplo, em um contexto de um esporte de equipe. Permite-se, primeiramente, seguir a ação de jogadores que interagem e se movem, o que não era possível com base nos métodos da técnica antecedente.

[00040] Preferencialmente, os métodos e sistemas da presente invenção capturam e produzem conteúdo automaticamente, sem a necessidade de processos manuais dispendiosos (nenhuma equipe técnica ou operador de câmera é necessário).

[00041] Como consequência de seu custo-benefício, a presente

invenção tem como objetivo manter a produção de conteúdo lucrativa mesmo para públicos de tamanho pequeno ou médio. Assim, a mesma promove a emergência de novos mercados e oferece ampla variedade de escolha de conteúdos que são de interesse para um número relativamente pequeno de usuários (por exemplo, o sumério de um evento esportivo regional, uma aula de universidade ou um dia na enfermagem).

[00042] Adicionalmente, ao automatizar a produção é possível a personalização de acesso a conteúdo. Gerar um vídeo personalizado consiste simplesmente em (re) executar o processo de produção com parâmetros de entrada que correspondem às preferências específicas ou restrições expressa pelo usuário.

[00043] Um propósito da presente invenção é produzir um relatório de vídeo de um evento com base na concatenação de segmentos de vídeo (e, opcionalmente, do áudio correspondente) capturados por um conjunto de câmeras. Na prática, tanto as câmeras estáticas quanto as dinâmicas podem ser manipuladas pela presente invenção:

[00044] º usar sensores estáticos adiciona ao custo-benefício, pois permite armazenar todo conteúdo relevante e processá-lo offline, selecionar os fragmentos de fluxos que são válidos para a apresentação ao expectador.

[00045] Os princípios de produção autônoma descritos abaixo poderiam também ser usados para controlar uma (conjunto de) câmera(s) PTZ dinâmica(s) . Nesse caso, as informações sobre a localização do objeto de interesse tem que ser fornecida em tempo real, por exemplo, com base na análise em tempo real do sinal capturado por alguns sensores audiovisuais (como feito em [ref]), ou com base nas informações coletadas dos transmissores embutidos. Além disso, o espaço de campos de visualização candidatos é definido pela posição e parâmetros de controle da câmera PTZ e não pela imagem

reenquadrada no ângulo de visualização coberto pela câmera estática.

[00046] O principal pressuposto por trás dos ajustes de aquisição de rede é a existência de uma referência temporal única comum para todas as visualizações de câmera, de maneira que as imagens de todas as câmeras possam ser mapeadas às mesmas coordenadas temporais absolutas da cena disponível. Assume-se, então, que as câmeras estejam sincronizadas de maneira frouxa, mas não necessariamente de maneira precisa. Assim, a sincronização frouxa se refere a um conjunto de câmeras que capturam imagens de maneira independente e que se baseia em marcações do tempo para associar as imagens que foram capturadas de maneira similar, mas não necessariamente nos instantes idênticos. Em contraste, uma sincronização precisa seria em referência a uma captura sincronizada das imagens pelas câmeras, como feito quando a aquisição é controlada por um sinal de acionamento comum.

[00047] Para decidir sobre como renderizar a ação da equipe disponível, a invenção precisa saber sobre a posição dos objetos de interesse na cena. Esse conhecimento pode ser uma estimativa (propensa a erros) e pode se referir tanto à posição de objetos na cena 3D ou à posição de objetos em cada uma das visualizações de câmera.

[00048] Essas informações podem ser fornecidas com base nos transmissores que são carregados pelos objetos a serem rastreados na cena de interesse. Esse conhecimento pode ser fornecido por uma alternativa não intrusiva, por exemplo, pela exploração de um conjunto de sinais de vídeo capturados por uma rede de câmeras estáticas, por exemplo, aquelas usadas para relatório de produção de vídeo, para detectar e rastrear os objetos de interesse. O método é descrito em "Detection e Recognition de Sports(wo)men from Múltiplos Views, D. Delannay, N. Danhier, e C. De Vleeschouwer, Third ACM/IEEE International Conference on Distributed Smart Cameras, Como, Itália, setembro de 2009" o qual é incorporado no presente documento por

referência em sua totalidade. O mesmo se baseia em um modelo de referência de segundo plano para identificar os pixels que mudam em cada visualização. Quando as múltiplas visualizações são calibradas, por exemplo, através de um processo offline, as máscaras de detecção de mudança que são coletadas em cada visualização podem ser intercaladas, por exemplo, em uma máscara de ocupação de fundo, para identificar a posição de objetos de interesse no espaço 3D (vide, por exemplo, a abordagem retratada na figura 16). Filtros de partícula ou técnicas com base em gráfico podem, então, ser usadas para ligar ocorrências do mesmo objeto ao longo da linha do tempo. Observa-se que tais técnicas de rastreamento e detecção são bastante conhecidas para aqueles versados na técnica e não serão descritas em detalhes no presente documento. A modalidade desses algoritmos que têm sido implantados é descrita na referência acima e oferece a vantagem de lidar com oclusões de uma maneira computacionalmente eficiente.

[00049] Uma vez que as posições dos objetos de interesse são conhecidas, a invenção suporta produção autônoma (= seleção de pontos de visualização ao longo do tempo) do conteúdo capturado pela rede de câmeras estáticas[1]. A abordagem é genérica no sentido de que a mesma pode integrar uma grande variedade de preferências de usuários que incluem recursos de exibição e transmissão, interesse semântico (como jogador preferido) ou preferências de narrativa (com relação à maneira preferida de visualizar a história, por exemplo, câmera preferida ou fator de ampliação).

[00050] Em um dado período de tempo, a presente invenção tem como objetivo selecionar a sequência de pontos de visualização que otimiza a produção de cena ao longo do tempo, com relação às pessoas/objetos de interesse detectados. Assim, um ponto de visualização se refere a um indicador de câmera e à janela que é enquadrada nessa visualização de câmera particular, para exibição real.

[00051] A otimização da sequência de pontos de visualização se baseia em inúmeras noções e princípios que podem ser descritos a seguir.

[00052] A cada instante, a otimização da produção precisa:

° maximizar a noção de totalidade, a qual mede em que extensão os (pixels dos) objetos de interesse são incluídos e visíveis dentro do ponto de visualização exibido. Opcionalmente, isto envolve minimizar o grau de oclusão de objeto, o qual mede a fração de um objeto que está presente na cena, mas é (por exemplo, pelo menos, parcialmente) escondido por outros objetos;

° maximizar a noção de proximidade, a qual se refere ao refinamento de detalhes, isto é, a densidade de pixels ou resolução, ao renderizar os objetos de interesse.

[00053] Aqueles dois objetivos são frequentemente contrários. Por essa razão, os métodos e sistemas de acordo com as modalidades da presente invenção propõem equilibrar totalidade e proximidade, opcionalmente como uma função das preferências individuais de usuários (em termos de resolução de ponto de visualização, ou câmera ou jogadores preferidos, por exemplo).

[00054] Finalmente, a suavidade de transições entre os parâmetros de renderização de quadros consecutivos do vídeo editado tem também que ser considerada na produção de um segmento temporal. Em outras palavras, é importante conservar a consistência entre a câmera e, por exemplo, os parâmetros de reenquadramento que são selecionados longo da linha do tempo, para evitar distrair o expectador da história por mudanças abruptas ou tremulação constante.

[00055] Com base nesses princípios guias, o processo de três etapas retratado na figura 14 foi desenvolvido. O mesmo pode ser descrito da seguinte maneira:

[00056] Etapa 1: A cada instante, e para cada visualização de

câmera, selecionar as variações em parâmetros como parâmetros de reenquadramento que otimizem a compensação entre totalidade e proximidade. Opcionalmente, a compensação totalidade/proximidade é medida como uma função das preferências do usuário. Por exemplo, dependendo da resolução na qual ele(a) acessa o conteúdo produzido, um usuário pode preferir um ponto de visualização pequena (aumento da aproximação) ou grande (diminuição a aproximação).

[00057] Etapa 2: Classificar o campo de visualização selecionado em cada visualização de câmera, de acordo com a qualidade (em termos de preferências do usuário) de sua compensação de totalidade/proximidade e seu grau de oclusões.

[00058] Etapa 3: Para os segmentos temporais disponíveis, computar os parâmetros de uma câmera virtual ideal que fotografa, aplica ampliação e troca entre câmeras para conservar altos graus de pontos de visualização selecionados enquanto minimiza a quantidade de movimentos da câmera virtual.

[00059] A primeira etapa consiste em selecionar o campo de visualização ideal para cada câmera, em um dado instante. Para simplificar as notações, a seguir, foi omitido o indicador do tempo t.

[00060] Um campo de visualização v_k na câmera estática k^n é definido pelo tamanho S_k e o centro c_k da janela que é enquadrada na visualização k^n para a exibição real.

[00061] O mesmo precisa ser selecionado de maneira a:

° incluir objetos de interesse;

° fornecer uma descrição refinada, isto é, alta resolução, desses objetos.

[00062] O campo de visualização ideal v_k^* é selecionado, preferencialmente, de acordo com as preferências do usuário, para maximizar uma soma ponderada de interesse de objeto conforme segue

$$\mathbf{v}_k^{*} = \arg\max_{\{S_k, \mathbf{c}_k\}} \sum_{n=1}^{N} I_n \cdot \alpha(S_k, \mathbf{u}) \cdot m\left(\frac{\left\|\mathbf{x}_{n,k} - \mathbf{c}_k\right\|}{S_k}\right)$$

[00063] Na equação acima:

[00064] ° I_n denota o nível de interesse designado o n-ésimo objeto reconhecido na cena. Essa designação pode ser feita por qualquer método adequado e a presente invenção pressupõe que essa designação tenha sido completada e os resultados podem ser usados pela presente invenção. Esses níveis de interesse podem ser definidos pelo usuário, por exemplo, uma vez para o evento inteiro, e disponibilizados para a presente invenção. Em cenários de aplicação nos quais os objetos são detectados, mas não rotulados, o peso é omitido, isto é, substituída por um valor unitário constante.

[00065] ° $x_{n,k}$ denota a posição do n-ésimo objeto na visualização de câmera *k*.

[00066] ° A função m(.) modula os pesos do n-ésimo objeto, de acordo com a distância do mesmo ao centro da janela de visualização, em comparação com o tamanho dessa janela. Intuitivamente, o peso deve ser maior e positivo quando o objeto de interesse se localiza no centro da janela de exibição, e deve ser negativo ou zero quando o objeto fica fora da área de visualização. Portanto, m(.) deve ser positivo entre 0 e 0,5, e menor ou igual a zero acima de 0,5. Muitas funções são apropriadas e a escolha de uma instância particular poderia, por exemplo, ser dirigida com base em processamentos computacionais. Exemplos de funções são as funções bastante conhecidas do Chapéu mexicano e Gaussiana. Outro exemplo é fornecido em detalhes em uma modalidade particular da invenção descrita no apêndice 1 desse pedido.

[00067] ° O vetor *u* reflete as restrições ou preferências do usuário em termos de resolução de janela de visualização e indicador de câmera. Em particular, seu componente u_{close} define a resolução do fluxo de saída, o qual é, em geral, restrito pela largura de banda de

transmissão ou resolução de dispositivo de usuário final. Seu componente u_{close} é ajustado para um valor maior do que 1 que aumenta para favorecer pontos de visualização próximos em comparação com visualizações grandes ampliadas. Os outros componentes de *u* se relacionam com as preferências de câmera e são definidos abaixo, enquanto da descrição da segunda etapa da invenção.

[00068] ° A função α(.) reflete a penalidade induzida pelo fato de que o sinal nativo capturado pela câmera k^n precisa ter sua resolução reduzida uma vez que o tamanho do ponto de visualização se torna maior do que a resolução máxima u_{res} permitida pelo usuário. Essa função, geralmente, diminui com S_k. Uma escolha apropriada consiste em ajustar a função para um quando $S_k < u_{res}$, e fazê-la diminuir posteriormente. Um exemplo de α(.) é definido por

$$\alpha(S, \boldsymbol{u}) = \left[\min\left(\frac{\mathrm{Ures}, 1}{S}\right)\right] u\ próximo$$

Substituir $u_{próxim}$ para u_{close}

[00069] Onde o exponente u_{close} é maior do que 1, e aumenta a favor dos pontos de visualização próximos em comparação aos campos de visualizações grandes reduzidos.

[00070] Vale notar que as compensações refletidas na equação acima podem ser formuladas de várias maneiras diferentes, porém equivalentes. Um exemplo de alternativa, mas equivalente, formulação foi implementada na modalidade da invenção definida no apêndice 1. Nessa formulação a soma de produto foi substituída um produto de somas, sem fundamentalmente afetar a ideia principal da invenção, a qual consiste em compensar proximidade e totalidade, de acordo com restrições de usuário (em relação à resolução de saída) e preferências (em relação ao ponto de visualização de aproximação aumentada ou diminuída).

[00071] A segunda etapa classifica o ponto de visualização associado a cada câmera de acordo com a qualidade de sua

compensação de totalidade/proximidade e seu grau de oclusões. A classificação mais alta deve corresponder a uma visualização que (1) faz a maioria dos objetos de interesse visíveis e (2) é próxima à ação, o que significa que apresenta objetos importantes com muitos detalhes, isto é, com uma alta resolução.

[00072] Formalmente, dado o interesse I_k de cada jogador, a classificação $I_k(v_k, u)$ associada à k^n visualização de câmera é definida da seguinte maneira:

$$I_k(\mathbf{v_k}, \mathbf{u}) = \sum_{n=1}^{N} I_n \cdot o_k(\mathbf{x}_n | \bar{\mathbf{x}}) \cdot h_k(\mathbf{x}_n) \cdot \beta_k(S_k, \mathbf{u}) m\left(\frac{\|\mathbf{x}_{n,k} - \mathbf{c}_k\|}{S_k}\right)$$

[00073] Na equação acima:

- I_n denota o nível de interesse designado para o n-ésimo objeto detectado na cena.
- x_n denota a posição do n-ésimo objeto no espaço 3D;
- $o_k(x_{nj}\ x)$ mede a razão de oclusão do n-ésimo objeto na visualização de câmera *k*, considerando a posição de todos os outros objetos. A razão de oclusão de um objeto é definida pela fração de pixels dos objetos que estão escondidos por outros objetos quando projetados no sensor da câmera;
- A altura $h_k(x_n)$ é definida pela altura em pixels da projeção na visualização *k* de 1,82 metros (seis pés) de altura vertical do objeto localizado em x_n, sendo que 1,8 metros (seis pés) é a altura média dos jogadores. O valor $h_k(x_n)$ é diretamente computado com base na calibração da câmera. Quando a calibração não está disponível, pode-se estimar com base na altura do objeto detectado na visualização k.
- A função $\beta_k(.)$ reflete o impacto das preferências de usuários em termos de visualização de câmera e resolução de exibição. Formalmente, β_k (S, u) = $u_k \cdot \alpha$ (S,u) pode ser definido como

$$\beta_k(S,\mathbf{u}) = u_k \cdot \alpha(S,\mathbf{u})$$

onde u_k denota o peso designado para a câmera k^n e $\alpha(S,u)$ é definido conforme indicado acima.

[00074] De modo similar ao que foi dito para a primeira etapa, vale mencionar que a formulação alternativa da mesma ideia básica pode ser imaginada. Por exemplo, a modalidade da invenção que descrita no apêndice 1 define a função para maximizar com base no produto de um fator de proximidade com um fator de totalidade, sendo que cada fator mede uma soma de peso de visibilidade e resolução de exibições de objetos individuais. Portanto, a mesma substitui a soma de produto por um produto de somas, mas ainda segue a mesma ideia básica levando em consideração as preferências de usuários enquanto compensa dois termos contrários, o que reflete no conceito de proximidade e totalidade, respectivamente.

[00075] De modo similar, uma formulação com base no peso de dois termos que refletem os conceitos de proximidade e totalidade descritos acima é também uma modalidade da presente invenção.

[00076] A terceira e última etapas consistem em suavizar os índices de câmera e parâmetro de ponto de visualização correspondentes.

[00077] Na modalidade proposta da invenção, o processo de suavização é implantado com base na definição de dois campos aleatórios de Markov (vide figura 5, e a descrição da modalidade abaixo). Outras modalidades podem também se basear em quaisquer mecanismos de filtro de passagem lenta linear ou não linear para suavizar a sequência de índices de câmera e parâmetros de pontos de visualização. A suavização poderia também ser feita através de um formalismo de modelo de gráfico, trabalhado com base em algoritmo de Viterbi convencional. Nesse caso, vértices de gráfico corresponderiam aos parâmetros de renderização candidatos para um dado quadro,

enquanto bordas conectariam estados de produção candidatos ao longo do tempo. O custo designado para cada borda refletiria no distúrbio induzido por uma mudança dos parâmetros de renderização entre dois quadros consecutivos.

[00078] O método e o sistema de produção de vídeo automatizado incluem também um diretor virtual, por exemplo, um módulo de diretor virtual para selecionar e determinar quais dos fluxos de câmera de vídeo múltiplos são um fluxo de câmera atual para ser visualizado. O diretor virtual, a cada instante, e para cada visualização de câmera, seleciona as variações em parâmetros, por exemplo, em parâmetros de reenquadramento que otimizem a compensação entre totalidade e proximidade. A compensação totalidade/proximidade é medida como uma função de preferências do usuário. Por exemplo, dependendo da resolução na qual um usuário acessa o conteúdo produzido, um usuário pode preferir um ponto de visualização pequeno (aumento da aproximação) ou grande (diminuição da aproximação). O módulo de diretor virtual classifica também o ponto de visualização selecionado em cada visualização de câmera de acordo com a qualidade (em termos de preferências do usuário) de sua compensação de totalidade/proximidade e seu grau de oclusões. Finalmente, o módulo de diretor virtual computa os parâmetros de uma câmera virtual ideal que fotografa, aplica ampliação e troca entre visualizações para o segmento temporal disponível, para conservar alta grau de pontos de visualização selecionados enquanto minimiza a quantidade de movimentos da câmera virtual.

[00079] Experimentos apontam que os pontos de visualização selecionados pelo diretor virtual, de acordo com modalidades da presente invenção, com base nas funções acima, correspondem às expectativas dos usuários finais. Mais que isso, testes subjetivos revelam que expectadores, em geral, preferem os pontos de

visualização selecionados com base no sistema automático em comparação com os selecionados por um produtor humano. Isso é parcialmente explicado pela carga severa imposta ao operador humano quando o número de câmeras aumenta. Portanto, a presente invenção também alivia a sobrecarga do operador humano, ao processar simultaneamente e em conjunto um grande número de câmeras fonte.

Breve Descrição das figuras

figura 1: fluido de trabalho hierárquico

figura 2: estrutura hierárquica

figura 3: função de ponderar

figura 4: comportamento de seleção de ponto de visualização

figura 5: modelo de duas etapas de estimativa de movimento de ponto de visualização

figura 6: planos de câmera

figura 7: amostras de visualizações de câmeras

figura 8: vídeo clipe curto

figura 9: sequências de ponto de visualização

figura 10: comportamento de sequência de câmera/ponto de visualização

figura 11: comparação de sequências de câmera e ponto de visualização

figura 12: quadros em sequências geradas

figura 13: comparação de sequências de câmera geradas

figura 14: modalidade de 3 etapas, da presente invenção

figura 15: dividir e concretizar modalidade da presente invenção

figura 16: uso de máscaras para a detecção

[00080] Desenhos adicionais são mostrados no apêndice 2. Esses desenhos se referem ao apêndice 2 e o texto do apêndice 2 deve ser lido em conjunto com esses desenhos e as referências específicas para

esse apêndice.

Descrição detalhada da presente invenção

[00081] A presente invenção fornece métodos e sistemas com base em computador para a geração de conteúdos de vídeo autônomos e com bom custo-benefício originário de dados de sensores múltiplos que incluem extração automática de conteúdos inteligentes de uma rede de sensores distribuída em volta da cena disponível. Assim, os conteúdos inteligentes se referem à identificação de segmentos salientes dentro do conteúdo audiovisual, pelo uso de algoritmos de analise de cena distribuídos. Esse conhecimento pode ser explorado para automatizar a produção e personalizar o sumário dos conteúdos de vídeo.

[00082] Sem perda de generalidade e sem limitar a presente invenção, somente câmeras estáticas serão escritas, principalmente, como uma modalidade ilustrativa.

[00083] Uma entrada são as posições de objetos de interesse. Para identificar segmentos salientes no conteúdo de vídeo bruto, uma análise de câmeras múltiplas é considerada, por meio da qual a detecção do objeto relevante, como métodos detecção de pessoas que se baseiam na fusão de informações prováveis do primeiro plano computadas em cada visualização podem ser usadas. Uma análise de visualizações múltiplas pode superar barreiras tradicionais como oclusões, sombras e iluminação em mudança. Isso contrasta com a análise através de um único sensor, a qual é frequentemente sujeita a interpretações ambíguas, devido à falta de precisão da cena, e configurações coincidentes de cenas adversas.

[00084] De acordo com algumas modalidades da presente invenção, assume-se que as posições dos objetos de interesse sejam conhecidas (pelo menos parcialmente) como uma função do tempo. Por exemplo, as modalidades da presente invenção inferem esse conhecimento a partir da análise dos campos de luz capturados por um conjunto de

câmeras estáticas distribuídas. Em tal modalidade uma máscara de ocupação de fundo pode ser computada por intercalar a probabilidade do primeiro plano medida em cada visualização. As posições reais dos jogadores podem, então, ser derivadas através de um processo iterativo ambicioso de percepção de oclusão. A análise de visualizações múltiplas pode ser usada para fornecer as entradas exigidas para o sistema e método para a produção autônoma de esporte de equipe da presente invenção e é descrita no artigo "Detection e Recognition de Sports(wo)men from Múltiplos Views", D. Delannay, N. Danhier, e C. De Vleeschouwer, Third ACM/IEEE International Conference on Distributed Smart Cameras, Como, Itália, setembro, 2009 é incorporado no presente documento por referência em sua totalidade como apêndice 2.

[00085] Modalidades da presente invenção procedem então em dois estágios.

[00086] Em um primeiro estágio, dadas as posições de cada objeto de interesse com tempo, os aspectos da invenção selecionam um conjunto chamado de parâmetros relevantes para renderizar a cena de interesse como uma função do tempo, pelo uso de uma câmera localizada em um ponto no qual pode ser qualquer ponto 3D arbitrário em volta da ação.

[00087] Aqui, os parâmetros de renderização definem um campo de visualização para a câmera e depende da infraestrutura da câmera que foi empregada para capturar as imagens da cena. Por exemplo, as modalidades da presente invenção usam uma câmera fixa e os parâmetros de renderização definem como enquadrar subimagens dentro da visualização de câmera. Em outras modalidades uma câmera motorizada e articulada pode ser usada e os parâmetros de renderização podem, então, se referir aos parâmetros de fotografia, inclinação e ampliação da câmera. A noção de parâmetros relevantes se relaciona com a definição de informativo, isto é, exibir as pessoas e

objetos de interesse, e imagens de percepção agradável.

[00088] Em um segundo estágio, as modalidades da presente invenção assumem que múltiplas câmeras (PTZ) são distribuídas na cena, e como selecionar a câmera certa para renderizar a ação em um dado tempo é então determinado. Isso é feito pela seleção e promoção de câmeras informativas e ao evitar troca entre câmeras inoportuna à percepção.

[00089] Juntos, o indicador de câmera e seu campo de visualização definem o ponto de visualização para renderizar a ação.

[00090] Para produzir sumários de vídeo confortáveis à percepção e com significado semântico com base na extração ou interpolação de imagens do conteúdo bruto, a presente invenção introduz três conceitos fundamentai, isto é, "totalidade", "suavidade" e proximidade (ou "refinamento"), para resumir a exigência narrativa e semântica dos conteúdos de vídeo. Com base nesses conceitos, a seleção de pontos de visualização de câmera de segmento temporais no sumário pode ser determinada, sendo que esses dois são problemas de otimização independente.

- **Totalidade** significa tanto integridade da renderização da visualização na seleção de câmera/ponto de visualização, e aquela do sumário da narração da história. Um ponto de visualização de alta totalidade inclui objetos mais salientes, enquanto uma história de alta totalidade consiste em mais ações principais.

- **Suavidade** refere-se ao deslocamento suave do ponto de visualização da câmera virtual, e à narração de história contínua que resulta da seleção de segmentos temporais contíguos. Ao conservar a suavidade é importante evitar distrair o expectador da história por mudanças abruptas de pontos de visualização ou pulos temporais constantes (Owen, 2007).

- **Proximidade ou Refinamento** refere-se à quantidade de

detalhes fornecidos sobre a ação renderizada. Espacialmente, favorece visualizações próximas. Temporalmente, se aplica à narração de história redundante, incluindo reproduções repetidas. Aumentar o refinamento de um vídeo não somente melhora a experiência de visualização, mas é também essencial em guiar o envolvimento emocional dos expectadores por capturas aproximadas.

[00091] De acordo com as modalidades da presente invenção, esses três conceitos são otimizados, por exemplo, maximizados para produzir um conteúdo visualmente agradável e significativo. Na prática, a maximização dos três conceitos pode resultar em decisões conflitantes, sobre algumas restrições de recuso limitadas, em geral expressas em termos da resolução espacial e duração temporal do conteúdo produzido. Por exemplo, na resolução fixa de vídeo de saída, ao aumentar a totalidade em geral se induz maiores pontos de visualização, os quais por sua vez diminuem o refinamento de objetos salientes. De modo similar, suavidade aumentada dos movimentos do ponto de visualização evita a busca precisa das ações de interesse ao longo do tempo. As mesmas observações são feitas em relação à seleção de segmentos e à organização das histórias ao longo do tempo, sobre algumas restrições de duração globais.

[00092] Consequentemente, as modalidades da presente invenção que se relacionam a métodos e sistemas baseados em computador fornecem um bom equilíbrio entre os três fatores principais. Por exemplo, métricas quantitativas são definidas para refletir a totalidade, refinamento/proximidade. A otimização restrita pode, então, ser usada para equilibrar esses conceitos.

[00093] Adicionalmente, para eficiência computacional aprimorada, tanto a produção quanto sumário são visados no paradigma dividir e realizar (vide fig. 15). Isso faz sentido, especificamente, já que os conteúdos de vídeo têm intrinsicamente uma estrutura hierárquica, que

inicia em cada quadro, capturas (conjunto de quadros consecutivos criados por trabalho de câmera similar), até segmentos semânticos (capturas consecutivas relacionadas logicamente à ação idêntica) e finalizando com a sequência geral.

[00094] Por exemplo, um evento de quadro de tempo pode ser primeiro cortado em segmentos temporais semanticamente significativos, como uma rodada de ofensa/defesa de esportes de equipe, ou uma entrada em novidades. Para cada segmento, várias opções de narrativa são consideradas. Cada opção define uma história local, a qual consiste em capturas múltiplas com diferentes coberturas de câmera. Uma história local não inclui somente capturas para renderizar a ação global disponível, mas também capturas para propósitos explanatórios e decorativos, por exemplo, reprodução repetida e aproximação de visualizações em esportes ou dados de gráfico em novidades. Dadas as marcações do tempo e a estratégia de produção (visualização a aproximada, reprodução repetida, etc.) das capturas que compõem uma opção de narrativa, o trabalho de câmera associado a cada captura é planejado automaticamente, levando em conta o conhecimento inferido sobre a cena por módulos de análise de vídeo.

[00095] Os benefícios e custos são, então, designados para cada história local. Por exemplo, o custo pode corresponder simplesmente à duração do sumário. O benefício reflete a satisfação do usuário (sobre algumas preferências individuais), e medidas como algumas exigências gerais, por exemplo, a continuidade e totalidade da história, são satisfeitas. Esses pares de benefícios e custos são, então, alimentados na máquina de sumário, a qual define um recurso de alocação de problema para encontrar a organização das histórias locais que alcançam os maiores benefícios no comprimento de sumário restrito.

[00096] O planejamento de trabalho de câmera será descrito com

referência a um exemplo, por exemplo, produção de vídeo de Vídeos de Esporte em Equipe de basquete. Apesar de abranger outros contextos (por exemplo, controle de câmera PTZ), o processo foi projetado para selecionar qual fração de qual visualização de câmera deve ser enquadrada em um conjunto de câmeras paradas distribuídas para renderizar a cena disponível em uma maneira prazerosa e semanticamente significativa ao assumir o conhecimento das posições dos jogadores.

Etapa 1: Seleção de Ponto de visualização pela câmera

[00097] A cada instante e em cada visualização, assume-se que os suportes dos jogadores são conhecidos e selecionam os parâmetros de reenquadramento que otimizam a compensação entre totalidade e refinamento.

[00098] Formalmente, um ponto de visualização v_{ki}, no qual a n-ésima visualização de câmera do i-ésimo quadro é definido pelo tamanho S_{ki}, e o centro c_{ki} da janela que é enquadrada na k-ésima visualização para exibição real. O mesmo precisa ser selecionado de forma a incluir objetos de interesse, e fornecer uma descrição refinada, isto é, alta resolução, desses objetos. Se não houver objetos salientes N nesse quadro, e a localização do n-ésimo objeto na k-ésima visualização for designada por x_{nki}, seleciona-se o ponto de visualização ideal v_{ki}* ao maximizar uma soma ponderada de objetos de interesse como segue:

$$\mathbf{v}_{ki}^{*} = \arg\max_{\{S_{ki},\mathbf{c}_{ki}\}} \sum_{n=1}^{N} I_n \cdot \beta(S_{ki},\mathbf{u}) \cdot \alpha\left(\frac{\|\mathbf{x}_{nki} - \mathbf{c}_{ki}\|}{S_{ki}}\right)$$

[00099] Na equação acima:

[000100] ° I_n determina o nível de interesse designado par o n-ésimo objeto detectado na cena. Observa-se que ao designar pesos para os jogadores de esporte de equipe permite focar em um jogador preferido, mas também implica no reconhecimento de cada jogador. Um peso de

unidade pode ser designado para todos os jogadores, produzindo assim um vídeo que renderiza a ação global do esporte de equipe.

[000101] ° O vetor **u** reflete as restrições e preferências do usuário em termos de resolução de ponto de visualização e visualização de câmera, u=[u^{close} u^{res} {u_k}]. Em particular, seu componente u^{res} define a resolução do fluxo de saída, o qual é em geral restringido pela largura de banda de transmissão ou resolução do dispositivo do usuário final. Seu componente u^{close} é configurado para um valor maior que 1, e aumenta para favorecer pontos de visualização próximos em comparação com visualizações grandes aumentadas por aproximação. Os outros componentes de **u** se relacionam com preferências de câmera e são definidos na segunda etapa abaixo.

[000102] °A função α(.) modula os pesos dos objetos de acordo com sua distância até o centro do ponto de visualização, em comparação ao tamanho dessa janela. Intuitivamente, o peso deve ser maior e positivo quando o objeto de interesse é localizado no centro da janela de exibição e dever ser negativo ou zero quando o objeto fica de fora da área de visualização. Muitas instâncias são apropriadas, por exemplo, a função bastante conhecida do Chapéu mexicano.

[000103] °A função *β*(.) reflete a penalidade induzida pelo fato do sinal nativo capturado pela câmera k^n tem que ter sua amostrada ampliada uma vez que o tamanho do ponto de visualização se torna maior do que a resolução máxima ^{res} permitida pelo usuário. Essa função diminui, em geral, com S_{ki}. Uma escolha apropriada consiste em ajustar a função para igual a um quando S_{ki}< u^{res}, e diminuir depois disso. Um exemplo de *β*(.) é definido por:

$$\alpha(\mathrm{Ski}, \boldsymbol{u}) = \left[\min\left(\frac{\mathrm{u\ res}, 1}{S}\right)\right] u\ póximo$$

[000104] Onde u^{close}> 1aumenta a favor de pontos de visualização próximos em comparação com visualizações grandes aumentadas por aproximação.

Etapa 2: Seleção de câmera em relação ao quadro

[000105] O ponto de visualização selecionado em cada visualização é classificado de acordo com a qualidade de sua compensação de totalidade/proximidade e seu grau de oclusões. A classificação mais alta desse corresponder à visualização que (1) faz a maioria dos objetos de interesse visíveis e (2) é próxima à ação, o que significa que apresente objetos importantes com muitos detalhes, isto é, uma alta resolução.

[000106] Formalmente, dado o interesse I_n de cada jogador, a classificação $I_{ki}(v_{ki}, \mathbf{u})$ associada a cada visualização de câmera é definida como segue:

$$I_{ki}(\mathbf{v}_{ki},\mathbf{u}) = u_k \cdot \sum_{n=1}^{N} I_n \cdot o_k\left(\mathbf{x}_{nki}\middle|\overline{\mathbf{x}}\right) \cdot h_k\left(\mathbf{x}_{nki}\right) \cdot \beta\left(S_{ki},\mathbf{u}\right) \cdot \alpha\left(\frac{\left\|\mathbf{x}_{nki}-\mathbf{c}_{ki}\right\|}{S_{ki}}\right)$$

[000107] Na equação acima:

- u_k determina o peso designado para a k-ésima câmera, enquanto α e β são definidos como na primeira etapa acima.
- $o_k(x_{nki}x)$ mede a razão de oclusão do n-ésimo objeto na visualização de câmera *k*, sabendo a posição de todos os outros objetos. A razão de oclusão de um objeto é definida como a fração de pixels dos objetos que estão escondidos por outros objetos quando projetado no sensor de câmera.
- A altura $h_k(x_{nki})$ definida como altura em pixels da projeção na visualização *k* de um objeto vertical de 1,8 m (seis pés) de altura localizado em x_{nki}. A altura de 1,8 m (seis pés) é a altura média dos jogadores. O valor de $h_k(x_{nki})$ é diretamente computado com base na calibragem da câmera. Quando a calibragem não está disponível, pode-se estimar com base na altura do objeto detectado na visualização *k*.

[000108] Etapa 3: Suavização de Sequências de Câmera/Ponto de visualização.

[000109] Para o segmento temporal disponível, os parâmetros de uma câmera virtual ideal que captura, aproxima e troca visualizações, são

computados para conservar elevados graus de pontos de visualização selecionados apesar de minimizar a quantidade de movimentos de câmera virtual.

[000110] O processo de suavização pode ser implementado com base na definição de dois Campos Aleatórios de Markov. No primeiro, $V^{\$}_k$ são capturados como dados observados na i-ésima imagem e é suposto que eles são as saídas distorcidas de alguns resultados suaves adjacentes v_{ki}. Dada a sequência de ponto de visualização suave recuperada para cada câmera, ganhos de câmera $l_{ki}(v_{ki}, u)$ daqueles pontos de visualização derivados são computados, e uma sequência de câmera suave do segundo campo de Markov é inferida por tornar as possibilidades $P(k|v_{ki}, u)$ de cada câmera proporcional aos ganhos $l_{ki}(v_{ki}, u)$.

[000111] Comparado aos simples filtros de suavização de Gaussian, é permitida a suavização adaptativa por ajustar diferente intensidade de suavização em cada quadro individual. Além disso, a suavização leve iterativa no método é capaz de alcançar resultados mais suaves do que suavização intensa de uma passagem.

[000112] Detecção e Reconhecimento de Jogador de Múltipla Visualização são obtidos em uma produção autônoma de conteúdo visual por contar com a detecção (e reconhecimento) de objeto de interesse na cena.

[000113] A probabilidade de primeiro plano é computada independentemente em cada visualização, usando técnicas padrões de modelagem de plano de fundo. Essas probabilidades são então unidas por serem projetadas no plano de fundo, e com isso, definem um conjunto das chamadas máscaras de ocupação de solo. A computação da máscara de ocupação de solo associada a cada visualização é eficiente, e essas máscaras são combinadas e processadas para inferir a posição atual de jogadores.

[000114] Formalmente, a computação da máscara de ocupação de solo G_k associada a k-ésima visualização é descrita como a seguir. Em um dado tempo, a k-ésima visualização é a fonte da probabilidade de imagem de primeiro plano $F_k \in [0,1]^{Mk}$, em que M_k é o número de pixels de câmera k, $0 < k < C$. Devido à suposição de verticalidade de jogador, segmentos de linha vertical ancorados em posições ocupadas no plano de solo sustentam uma parte do objeto detectado, e então, projetam de volta no primeiro plano, as silhuetas em cada visualização de câmera. Portanto, para refletir a ocupação de solo em x, o valor de G_K em x é definido para ser a integração da projeção (à frente) de F_k em um segmento vertical ancorado em x. Obviamente, essa integração pode, equivalentemente, ser computada em F_K ao longo da retroprojeção do segmento vertical ancorado em *x*. Isso acontece em métodos de contraste que computam a máscara por agregar as projeções da probabilidade de primeiro plano em um conjunto de planos que são paralelos ao solo.

[000115] Para acelerar as computações associadas à presente formulação, é observado que, através de transformação apropriada de F_K, é possível formatar o domínio de integração retroprojetado de forma que ele também corresponda a um segmento vertical na visualização transformada, com isso, torna a computação de integrais particularmente eficiente através do princípio de imagens integrais. A transformação foi projetada para endereçar um objetivo duplo. Primeiro, os pontos do espaço 3D localizados na mesma linha vertical precisam ser projetados na mesma coluna na visualização transformada (ponto de fuga vertical ao infinito). Segundo, objetos verticais que estão colocados sobre o solo e cujos pés são projetados na mesma linha horizontal da visualização transformada, precisa manter as mesmas razões de altura projetadas. Uma vez que a primeira propriedade é encontrada, os pontos 3D que pertencem à linha vertical que está

colocada acima de um dado ponto a partir do plano de solo, simplesmente projetam na coluna da visualização transformada que se coloca acima da projeção do plano 3D de solo ponto. Portanto, G_k,(x) é simplesmente computado como o integral da visualização transformada sobre seu segmento vertical retroprojetado. A preservação de altura ao longo das linhas da visualização transformada simplifica ainda mais as computações.

[000116] Para visualizações laterais, essas duas propriedades podem ser alcançadas por movimentar virtualmente (através de transformação homográfica) a direção de visualização de câmera (eixo geométrico principal) de forma a trazer o ponto de fuga vertical ao infinito e garantir que a linha do horizonte esteja horizontal. Para visualizações superiores, o eixo geométrico principal é ajustado de forma perpendicular ao solo e um mapeamento polar é desempenhado para alcançar as mesmas propriedades. Nota-se que em algumas configurações geométricas, estas transformações podem induzir distorção grave das visualizações.

[000117] Dada a máscara de ocupação de solos G_K para todas as visualizações, é explicado como inferir a posição das pessoas que se colocam sobre o solo. Por teoria, em um contexto de esporte de equipe, sabe-se que (i) cada jogador induz um conjunto denso na soma de máscaras de ocupação de solo e, (ii) no número de pessoas para detectar, sendo que é igual a um valor N conhecido, por exemplo, N = 12 para basquetebol (10 jogadores + 2 árbitros).

[000118] Por essa razão, em cada local de solo x, é considerada a soma de todas as projeções - normalizada pelo número de visualizações que realmente cobrem x -, e olhar os pontos de maior intensidade nessa máscara de ocupação de solo agregada. Para localizar esses pontos, foi considerada primeiramente, uma ávida abordagem simples que é equivalente a um procedimento de busca compatível iterativo. Em cada

etapa, o processo de busca compatível maximiza o produto interno entre um núcleo de Gaussian traduzido e uma máscara de ocupação de solo agregada. A posição do núcleo induz o produto interno maior a definir a posição do jogador. Antes de executar a próxima iteração, a contribuição do núcleo de Gaussian é subtraída da máscara agregada para produzir uma máscara residual. O processo reitera até que jogadores suficientes serem localizados.

[000119] Essa abordagem é simples, mas sofre a partir de muitas detecções falsas na interseção das projeções de distintas silhuetas de jogadores de visualizações diferentes. Isso acontece devido ao fato de que oclusões induzem não linearidades na definição da máscara de ocupação de solo. Em outras palavras, a máscara de ocupação de solo de um grupo de jogadores não é igual a soma de máscara de ocupação de solos projetada por cada jogador individual. Conhecimento sobre a presença de algumas pessoas no campo de solo afeta o valor informativo das máscaras de primeiro plano desses locais. Em particular, se a linha vertical associada a uma posição x é ocluída por/oclui outro jogador cuja presença é muito provável, essa visualização particular pode não ser explorada para decidir se existe um jogador x ou não.

[000120] Um refinamento envolve inicializar o processo por definir $G_k^1(x) = G_k(x)$ para ser a máscara de ocupação de solo associada a k-ésima visualização, e ajustar $W_k^1(x)$ para 1 quando x é coberto pela por k-ésima visualização, e para 0, caso contrário.

[000121] Cada iteração é então executada em duas etapas. Na iteração n, a primeira etapa pesquisa a posição mais provável do n-ésimo jogador, sabendo a posição dos (n-1) jogadores localizados em iterações anteriores. A segunda etapa atualiza a máscara de ocupação de solos de todas as visualizações para remover a contribuição do jogador recém-localizado. Formalmente, a primeira etapa de iteração n

agrega a máscara de ocupação de solo a partir de todas as visualizações, e então procura pelo conjunto mais denso nessa máscara. Portanto, computa a máscara agregada como:

$$G^n(x) = \frac{\sum_{k=1}^{C} w_k^n(x).G_k^n(x)}{\sum_{k=1}^{C} w_k^n(x)},$$

e então define a posição x_n mais provável para o n-ésimo jogador por:

$$x_n = \underset{y}{\operatorname{argmax}} < G^n, \phi(y) >$$

onde $\varphi(y)$ denota um núcleo Gaussiano centralizado em y, e cujo suporte espacial corresponde à largura típica de um jogador.

[000122] Na segunda etapa, a máscara de ocupação de solo de cada visualização é atualizada para se responsabilizar pela presença do n-ésimo jogador. Na posição de solo x, considera-se que um suporte típico de uma silhueta de jogador em visualização k: é uma caixa retangular de largura W e altura H, e observa-se que a parte da silhueta que oclui ou é ocluída pelo jogador recém-detectado não traz quaisquer informações cobre a presença potencial de um jogador em posição x. A fração $\varphi_k(x, X_n)$ da silhueta em posição de solo x que se torna não informativa na k-ésima visualização, é estimada como uma consequência da presença de um jogador em x_n. É então proposto atualizar a máscara de ocupação de solo e o peso de agregação da k-ésima câmera na posição x como a seguir:

$$G_k^{n+1}(x) = \max\left(0, G_k^n(x) - \varphi_k(x, x_n).G_k^1(x_n)\right),$$

$$w_k^{n+1}(x) = \max\left(0, w_k^n(x) - \varphi_k(x, x_n)\right).$$

[000123] Para eficiência computacional aprimorada, as posições x investigadas na abordagem refinada são limitadas à máxima local de 30 detectadas pela abordagem simples.

Para totalidade, é notado que o procedimento de atualização descrito acima omite a interferência potencial entre oclusões causadas por jogadores distintos na mesma visualização. Contudo, a consequência dessa aproximação está longe de ser dramática, visto que ela acaba omitindo parte das informações que foram significantes para acessar a ocupação em posições ocluídas, sem afetar as informações que são realmente exploradas. Para levar em conta tais interferências, exige-se retroprojeção das silhuetas de jogadores em cada visualização, o que tende então, a uma abordagem dispendiosa de computação e de memória. O método e sistema da presente invenção não sofrem das falhas comuns de algoritmos ávidos, como uma tendência de ficar preso em local ruim mínimo.

[000124] Os principais benefícios técnicos da presente invenção incluem pelo menos uma ou uma combinação de:

- A capacidade de cortar pixels apropriados na memória de imagem e/ou controlar um PTZ motorizado, de forma a renderizar uma ação de time, ou seja, uma ação que envolve movimentação múltipla de objetos/pessoas de interesse, a partir de um ponto 3D arbitrário.

- A capacidade para (i) controlar campo de seleção de visualização por câmera individual, e (ii) selecionar uma melhor câmera em um conjunto de câmeras. Tal habilidade torna possível tratar um número potencialmente muito grande de câmeras simultaneamente. Isso especialmente verdadeiro, visto que a renderização de seleção de parâmetros para uma câmera particular pode ser computada independentemente de outras câmeras.

- A possibilidade de reproduzir e então, tecnicamente personalizar o processo de seleção de ponto de visualização de acordo com preferências de usuário individual. Por exemplo, no contexto de um evento de esporte, treinadores (que preferem amplos pontos de visualização que mostram o jogo inteiro) têm diferentes expectativas em

relação à seleção de ponto de visualização do que o espectador comum (que prefere imagens mais próximas e emocionalmente mais ricas). Dessa forma, preferências são diretamente relacionadas a parâmetros técnicos de como as câmeras são controladas. Mecanizar o processo de produção proporciona uma solução técnica que equivale a responder as exigências individuais.

[000125] A presente invenção inclui em seu escopo aprimoramentos adicionais. A presente invenção inclui outros critérios para seleção resolvível analiticamente e/ou computacionalmente eficiente de pontos de visualização. Ela também inclui melhor representação para objetos salientes tais como, o uso de partículas em movimento ou modelos de corpo flexível ao invés de simples caixas confinadas. Além disso, a seleção e suavização de pontos de visualização e câmeras em quatro de subetapas na versão atual, simplificam a formulação. Contudo, elas podem ser resolvidas em uma estimativa unificada porque seus resultados afetam uns aos outros. A presente invenção também inclui outros critérios de seleção de pontos de visualização e câmeras independentes de avaliações pessoais.

[000126] A exploração de uma rede distribuída de câmeras para aproximar as imagens que seriam capturadas por um sensor virtual localizado em uma posição arbitrária, com cobertura de ponto de visualização arbitrária, pode ser usada com qualquer das modalidades da presente invenção. A presente invenção pode ser usada com esses mecanismos, porque de acordo com a presente invenção, uma seleção é feita do ponto de visualização mais apropriado em um ajuste/espaço de pontos de visualização candidatos. Portanto, a adição de algoritmos de renderização de ponto de visualização livre em modalidades da presente invenção contribui apenas para ampliar o ajuste de candidatos potenciais.

[000127] Os métodos e o sistema da presente invenção podem ser

implementados em um sistema de computação que pode ser utilizado com os métodos e em um sistema de acordo com a presente invenção que inclui programas de computador. Um computador pode incluir um terminal de exibição de vídeo, um meio de entrada de dados como um teclado e uma interface gráfica de usuário, que indica um meio como um mouse. O computador pode ser implementado como um computador de propósito geral, por exemplo, uma estação de trabalho UNIX ou um computador pessoal.

[000128] Tipicamente, o computador inclui uma Unidade Central de Processamento ("CPU"), como um microprocessador convencional do qual um processador Pentium fornecido por Intel Corp. USA é apenas um exemplo, e inúmeras de outras unidades interconectadas por sistema de barramento. O sistema de barramento pode ser qualquer sistema de barramento adequado. O computador inclui pelo menos uma memória. A memória pode incluir qualquer uma, de uma variedade de dispositivos de armazenamento de dados conhecidos por elementos versados na técnica tal como, memória de acesso aleatório ("RAM"), memória de apenas leitura ("ROM"), memória de leitura/escrita não volátil como um disco rígido, como conhecido elementos versados na técnica. Por exemplo, computador pode ainda incluir memória de acesso aleatório ("RAM"), memória de apenas leitura ("ROM"), bem como um adaptador de exibição para conectar barramento de sistema a um terminal de exibição de vídeo, e um adaptador opcional de entrada/saída (I/O) para conectar dispositivos periféricos (por exemplo, unidades de disco e fita) ao barramento de sistema. O terminal de exibição de vídeo pode ser a saída visual de computador, que pode ser qualquer dispositivo de exibição adequado como, um exibidor de vídeo com base CRT bem conhecido na técnica de hardware de computador. Contudo, com um computador de mesa, um computador com base portátil ou um notebook, o terminal de exibição de vídeo pode ser

substituído por exibidor de tela plana de plasma ou LCD. O computador ainda inclui usar um adaptador de interface para conectar um teclado, mouse e caixa de som opcional. O vídeo relevante exigido pode ser inserido diretamente no computador por uma interface gráfica ou de vídeo ou a partir de dispositivos de armazenamento, depois disso, um processador executa um método de acordo com a presente invenção. Os dados de vídeo relevantes podem ser proporcionados em um meio de armazenamento de sinal adequado como um disquete, um disco rígido substituível, um dispositivo de armazenamento ótico como um CD-ROM ou DVD-ROM, uma fina magnética ou similar. Os resultados do método podem ser transmitidos a um local próximo ou remoto. Um adaptador de comunicações pode conectar o computador a uma rede de dados com a Internet, uma Intranet, uma rede de área local (LAN ou WAN) ou uma CAN.

[000129] O computador também inclui uma interface gráfica de usuário que reside no meio legível por máquina para direcionar a operação do computador. Qualquer meio legível por máquina adequado pode manter a interface gráfica de usuário, tal como uma memória de acesso aleatório (RAM), uma memória de apenas leitura (ROM), um disquete magnético, fita magnética, ou disco ótico (os três últimos são localizados em unidades de disco e fita). Qualquer sistema de operação adequado e interface gráfica de usuário associado (por exemplo, Microsoft Windows, Linux) podem direcionar a CPU. Além disso, o computador inclui um programa de controle que reside no armazenamento de memória de computador. O programa de controle contém instruções que quando executadas na CPU permitem que o computador execute as operações descritas em relação a quaisquer dos métodos da presente invenção.

[000130] A presente invenção também proporciona um produto de programa de computador para executar o método da presente invenção

e ele pode residir em qualquer memória adequada. Contudo, é importante salientar que, apesar da presente invenção ter sido, e continuar sendo apreciada por versados na técnica, os mecanismos da presente invenção são capazes de serem distribuídos como um produto de programa de computador em uma variedade de formas, e a presente invenção se aplica igualmente ao tipo particular de mídia de produção de sinal usada atualmente para executar a distribuição. Exemplos de mídias de produção de sinal legível por computador incluem: mídias de tipo regravável como disquetes e CD ROMs e mídias de tipo de transmissão como ligações de comunicação analógica e digital. Consequentemente, a presente invenção também inclui um produto de software que quando executado em um dispositivo de computador adequado executa quaisquer dos métodos da presente invenção. Softwares adequados podem ser obtidos por programação em uma linguagem de alto nível adequada como C e compilação em um compilador adequado para o processador e computador alvo ou em uma linguagem interpretada como Java e então compilada em um compilador adequado for implementação com a Máquina Virtual Java.

[000131] A presente invenção proporciona software, por exemplo, um programa de computador que tem segmentos de código que proporcionam um programa que, quando executado em um mecanismo de processamento, proporciona um módulo diretor virtual. O software pode inclui segmentos de código que proporcionam, quando executados no mecanismo de processamento: quaisquer dos métodos da presente invenção ou implementam qualquer meio de sistema da presente invenção.

[000132] Outros aspectos e vantagens da presente invenção bem como um entendimento mais completo deles se tronarão aparentes a partir da seguinte descrição capturada em conjunto com as figuras embutidas e que as acompanham, que ilustram por forma de exemplo

os princípios da invenção. Além disso, é pretendido que o escopo da invenção seja determinado pelas reivindicações adjacentes e não pelo sumário precedente ou a seguinte descrição detalhada.

Apêndice 1

1. Introdução

[000133] Visando a produção de conteúdos semanticamente significativos e confortáveis de maneira perceptiva, a partir de dados crus de múltiplos sensores, é proposto um sistema de produção computacionalmente eficiente, com base no paradigma dividir e conquistar. Foram resumidos os principais fatores-alvo por três palavras-chave, que são "totalidade", "proximidade" e "suavidade". Totalidade representa a totalidade de renderizar a visualização. Proximidade define a delicadeza de descrição de detalhes, e suavidade é um termo que se refere à continuidade tanto do movimento de ponto de visualização quanto ao ato de contar história. Por trocar entre tais fatores, foram desenvolvidos métodos para selecionar pontos de visualização ideais e câmeras para ajustar a resolução de exibição e outras preferências de usuário, e suavizar essas sequências em um ato de contar histórias gracioso e contínuo. Existe uma longa lista de possíveis preferências de usuário, tais como perfil de usuário, histórico de navegação de usuário e capacidades de dispositivo. Resumiram-se preferências narrativas em quatros descritores, por exemplo, time preferido de usuário, jogador preferido de usuário, evento preferido de usuário e câmera preferida de usuário. Todas as limitações de dispositivo, como resolução de exibição, velocidade de rede, desempenho de decodificador, são simplificadas como resolução de exibição preferida. É então discutido, principalmente, preferências de usuário com esses cinco elementos no presente trabalho.

[000134] A capacidade de levar em conta tais preferências obviamente depende do conhecimento capturado sobre a cena através

de ferramentas de análise de vídeo, por exemplo, detectar qual time esta atacando ou defendendo. Contudo, de forma mais importante, é válido mencionar que o presente documento é genérico e pode incluir qualquer tipo de preferência de usuário.

[000135] Na seção 2, é explicado o trabalho de quadro para estimar tanto a seleção e suavização de pontos de visualização quanto as visualizações de câmera, e é dada sua formulação e implementação detalhada. Na seção 3, detalhes mais técnicos são dados e experimentos são feitos para verificar a eficiência do sistema. Finalmente, conclui-se que são criadas e listadas inúmeras trajetórias possíveis para pesquisa futura.

2. Produção autônoma de vídeos de basquetebol personalizados de dados de sensores múltiplos

[000136] Embora seja difícil definir uma regra absoluta para avaliar o desempenho de histórias organizadas e pontos de visualização determinados apresentando um cenário genérico, a produção de vídeos de esporte tem alguns princípios gerais. [11] Para jogos de basquetebol, resumiram-se essas regras em três compensações principais.

[000137] A primeira compensação surge da personalização da produção. Especificamente, é originada do conflito entre conservar regras de produção geral de vídeos de esportes e maximizar a satisfação de preferências de usuário. Algumas regras básicas de produção de vídeo para jogos de basquetebol não deve ser sacrificada para melhor satisfação de preferências de usuário, por exemplo, a cena deve sempre incluir a bola, e ponderação bem balanceada deve ser capturada entre o jogador dominante e o jogador preferido do usuário que renderizar um evento.

[000138] A segunda compensação é o balanço entre totalidade e proximidade da cena renderizada. O interesse intrínseco de jogos de basquetebol em parte, vem da complexidade do trabalho em equipe,

cuja descrição clara exige totalidade espacial em cobertura de câmera. Contudo, muitas atividades destacadas normalmente acontecem em uma área de jogo especifica e limitada. Uma visualização próxima que destaca tais áreas aumenta o envolvimento emocional do público com o jogo, por mover o público para mais perto da cena. A proximidade é também exigida para gerar uma visualização do jogo com resolução espacial suficiente sob uma situação com recursos limitados, como tamanho de exibição pequeno ou recursos de largura de banda limitados de dispositivos portáteis.

[000139] A compensação final equilibra a busca precisa de ações de interesse ao longo do tempo e suaviza a movimentação de ponto de visualização. A necessidade do publico saber a situação geral em relação ao jogo durante toda a competição é uma exigência primária e um principal propósito de troca de ponto de visualização. Quando os ângulos de diferentes câmeras são misturados para destacar ou outros efeitos especiais, a suavidade de troca de câmera deve ser mantida em mente para ajudar o publico a reorientar a situação do jogo rapidamente depois de movimentos de ponto de visualização. [11]

[000140] Dados os metadados reunidos a partir de dados de vídeo de sensores múltiplos, é planejada a cobertura de ponto de visualização e troca de câmera por considerar as três compensações acima. É dada uma visão geral do trabalho de quadro de produção na seção 2.1 e introdução de algumas anotações em metadados na seção 2.2. Na seção 2.3, é proposto critério para selecionar ponto de visualização e câmera em um quadro individual. Suavização de ponto de visualização e sequências de câmera são explicadas na seção 2.4.

2.1. Visão geral da estrutura de produção

[000141] É inevitável causar descontinuidade de conteúdos do ato de contar história conteúdos durante troca de visualizações de câmera. Para suprimir a influência dessa descontinuidade, normalmente é

localizado um ponto de visualização dramático ou troca de câmera durante o intervalo de dois eventos destacados, para evitar possível distração do publico da historia. Portanto, pode-se considerar a produção personalizada no paradigma dividir e conquistar, como mostrado na figura 1. A história inteira é primeiramente dividida em diversos segmentos. Pontos de visualização e câmeras ideais são determinados localmente em cada segmento por compensar entre benefícios e custos sob preferências de usuário especificas. Além disso, a estimativa de câmera ou pontos de visualização é desempenhada em uma estrutura hierárquica. A fase de estimativa adota etapas ativas de todos os quadros individuais da história inteira. Iniciando a partir de um quadro único, o ponto de visualização é otimizado em cada câmera de visualização individual, determinando a melhor visualização de câmera a partir de múltiplas câmeras candidatas sob os pontos de visualização selecionados, e finalmente, organizando a história inteira. Quando é necessário renderizar a história para o publico, um processamento de cima para baixo é capturado, sendo que, primeiramente divide o vídeo em segmentos não sobrepostos. Os quadros correspondentes para cada segmento são então, apanhados e exibidos no dispositivo alvo com câmeras e pontos de visualização específicos.

[000142] A estrutura hierárquica intrínseca de jogos de basquetebol proporciona solos razoáveis para a visão de cima, e também proporciona dicas em separação de segmento. Como mostrado na figura 2, um jogo é dividido por regras em uma sequência de períodos de posse de bola não sobrepostos. Um período de posse de bola é o período de jogo em que o mesmo time contém a bola e faz diversas tentativas de pontuação. Em cada período, diversos eventos devem ocorrer durante o processo de ataque/defesa. De acordo com o caso do evento ser relacionado a uma jogada de tempo de 24-segundos, eventos em um jogo de basquete podem ser classificados como eventos

de tempo e eventos sem tempo. Eventos de tempo não serão sobrepostos uns aos outros, contudo, eventos sem tempo devem ser sobrepostos com eventos de tempo/sem tempo. Em geral, um período de posse de bola é particularmente um período fluente e exige uma continuidade de nível de período de movimento de ponto de visualização.

[000143] Neste artigo, primeiramente, define-se os critérios para avaliar os pontos de visualização e câmeras em cada quadro individual. A suavidade de modo de câmera de pontos de visualização é então aplicada a todos os quadros em cada período de posse de bola. Com base em determinados pontos de visualização, uma sequência de câmera é selecionada e suavizada.

2.2. Metadados e Preferências de Usuário

[000144] Os dados de entrada alimentados no sistema do presente documento incluirão dados de vídeo, metadados associados e preferências de usuário. Suponha-se que um banco de dados de sequências de vídeo de basquete, que são captados simultaneamente por *K* câmeras diferentes. Todas as câmeras são vagamente sincronizadas e produzem o mesmo número de quadros, isto é, *N* quadros, para cada câmera. No i-ésimo quadro capturado no momento t_i, M_i objetos salientes diferentes, denotados por $\{o_{im} | m = 1, ..., M_i\}$, são detectados no total de todas as visualizações da câmera. Há dois tipos de objetos salientes definidos. A primeira classe inclui as regiões para os jogadores, árbitros, e a bola, que são utilizados para o entendimento da cena. A segunda classe inclui a cesta, banco do treinador e alguns pontos de referência da quadra, que são usados no entendimento da cena e na calibração da câmera. Objetos da primeira classe são automaticamente extraídos da cena tipicamente baseada no algoritmo de subtração de segundo plano, enquanto que os de segunda classe são manualmente identificados, pois suas posições são constantes em

câmeras fixas. Define-se o objeto m-ésimo saliente como o_{im} = [o_{kim} = 1 ... *K*], em que o_{kim} é o m-ésimo objeto saliente na k-ésimésima câmera.

[000145] Todos os objetos salientes são representados por regiões de interesse. A região r é um conjunto de coordenadas de pixel que são pertencentes a esta região. Se o_{im} não aparece na visualização da k-ésimésima câmera, definimos o_{kim} para o conjunto vazio *ø*. Com r1 e r2 sendo duas regiões arbitrárias, definem-se primeiro várias funções elementares em uma ou duas regiões como

$$\text{Área}: \mathcal{A}(\mathbf{r}_1) = \sum_{\mathbf{x}\in\mathbf{r}_1} 1; \quad (1)$$

$$\text{Centro}: \mathcal{C}(\mathbf{r}_1) = \frac{1}{\mathcal{A}(\mathbf{r}_1)} \sum_{\mathbf{x}\in\mathbf{r}_1} \mathbf{x}; \quad (2)$$

$$\text{Visibilidade}: \mathcal{V}(\mathbf{r}_1|\mathbf{r}_2) = \begin{cases} 1, & \mathbf{r}_1 \subseteq \mathbf{r}_2 \\ -1, & \text{caso contrário;} \end{cases} \quad (3)$$

$$\text{Distância}: \mathcal{D}(\mathbf{r}_1, \mathbf{r}_2) = \|\mathcal{C}(\mathbf{r}_1) - \mathcal{C}(\mathbf{r}_2)\|; \quad (4)$$

que serão usadas em seções do presente documento adiante.

[000146] Além disso, definem-se as preferências de usuário por um parâmetro estabelecido u, que inclui preferências narrativas e restritivas, como favoritos e capacidades de dispositivo.

2.3. Seleção de Câmera e Pontos de Visualização em Quadros Individuais

[000147] Para simplificar, deixa-se de lado o problema de suavização na primeira etapa, e começa-se por considerar a seleção de um ponto de visualização apropriado em cada quadro independente. Usam-se as duas subseções a seguir para explicar a solução do presente documento para este problema sob dois aspectos, isto é, a avaliação de vários pontos de visualização na mesma visualização de câmera e avaliação de visualizações de câmera diferentes.

2.3.1. Computação de Pontos de Visualização Ideais em Cada Câmera Individual

[000148] Embora a avaliação do ponto de visualização seja uma tarefa

altamente subjetiva, que ainda carece de uma regra objetiva, têm-se alguns requisitos básicos na seleção de ponto de visualização do presente documento. Deve ser eficiente computacionalmente, e deve ser adaptável sob resoluções de dispositivo diferentes. Para um dispositivo com alta resolução de exibição, geralmente é preferível uma visualização completa da cena toda. Quando a resolução é limitada devido a restrições de dispositivo ou canal, deve-se sacrificar parte da cena para uma representação melhorada de detalhes locais. Para um objeto bem próximo do limite ponto de visualização, o mesmo deve ser incluído para melhorar a abrangência global da narração de história, se mostrar alta relevância para o evento atual em quadros posteriores, e deve ser excluído para evitar que a sequência de ponto de visualização de oscile, se sempre aparecer próximo do limite. A fim de manter uma área segura para lidar com esse tipo de objeto, nós preferimos que os objetos salientes visíveis dentro do ponto de visualização determinado estejam mais próximos do centro, enquanto objetos invisíveis devem ser retirados do limite do ponto de visualização, na medida do possível.

[000149] Deixa-se o ponto de visualização para a construção de cena no i-ésimo quadro da k-ésimésima câmera ser v_{ki}. O ponto de visualização v_{ki} é definido como uma região retangular. Para a representação natural da cena, limita-se a razão de aspecto de todos os pontos de visualização como sendo a mesma razão de aspecto do dispositivo de exibição. Portanto, para cada v_{ki}, têm-se apenas três parâmetros livres, isto é, o centro horizontal v_{kix}, o centro horizontal v_{ki} e a largura v_{kiw}, para regular. O ponto de visualização ideal individual é obtido maximizando-se o ganho de interesse da aplicação do ponto de visualização v_{ki} ao i-ésimo quadro da k-ésimésima câmera, o que é definido como uma soma ponderada de interesses de atenção de todos os objetos salientes visíveis naquele quadro, isto é,

$$\mathcal{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u}) = \sum_{m} w_{kim}(\mathbf{v}_{ki}, \mathbf{u})\mathcal{I}(\mathbf{o}_{kim}|\mathbf{u}), \qquad (5)$$

em que $\mathcal{I}(\mathbf{o}_{kim}|\mathbf{u})$ é o interesse de um objeto saliente o_{kim} sob a preferência de usuário u. No presente documento, a função de interesse predefinida $\mathcal{I}(\mathbf{o}_{kim}|\mathbf{u})$ resultará em ponderação diferente, de acordo com diferentes valores de u, que reflete as preferências de usuário de narrativas. Por exemplo, é atribuído maior interesse a um jogador especificado pela audiência do que a um jogador não especificado, e é dado o máximo interesse à bola, de modo que a mesma está sempre incluída na cena. Será explicado um ajuste de $\mathcal{I}(\mathbf{o}_{kim}|\mathbf{u})$ com mais detalhes na próxima seção.

[000150] Define-se $w_{kim}(v_{ki},u)$ para ponderar a importância de atenção de um único objeto dentro de um ponto de visualização. Matematicamente, toma-se $w_{kim}(v_{ki},u)$ da forma como a seguir

$$w_{kim}(\mathbf{v}_{ki},\mathbf{u}) = \frac{\mathcal{V}(\mathbf{o}_{kim}|\mathbf{v}_{ki})}{\ln \mathcal{A}(\mathbf{v}_{ki})} \exp\left[-\frac{\mathcal{D}(\mathbf{o}_{kim},\mathbf{v}_{ki})^2}{2[u^{\mathrm{DEV}}]^2}\right], \qquad (6)$$

em que se usa u^{DEV} para denotar a limitação da resolução de exibição atual na preferência de usuário u. A presente definição de $w_{kim}(v_{ki},u)$ consiste em três partes principais: a parte exponencial que controla a intensidade de concentração de objetos salientes ao redor do centro, de acordo com a resolução de pixels da exibição do dispositivo; a parte de passagem por zero $\mathcal{V}(\mathbf{o}_{kim}|\mathbf{v}_{ki})$ que separa interesses positivos de interesses negativos no limite do ponto de visualização; e a parte fracionária anexada $\ln \mathcal{A}(\mathbf{v}_{ki})$ que calcula a densidade de interesses para avalias a proximidade e é ajustada como uma função de logaritmo. Deve-se observar que $\mathcal{V}(\mathbf{o}_{kim}|\mathbf{v}_{ki})$ é positiva apenas quando o objeto saliente o_{kim} está totalmente contido dentro do ponto de visualização v_{ki}, que mostra a tendência de se manter um objeto saliente intacto na seleção de ponto de visualização. Conforme é mostrado na figura 3, a ideia básica da presente definição é alterar a importância relativa de totalidade e proximidade regulando-se a acuidade do pico central e modificando-se o comprimento das abas. Quando u^{DEV} é

pequena, a parte exponencial decai muito rapidamente, o que tende a enfatizar os objetos mais próximos do centro e ignorar objetos fora do ponto de visualização. Quando u^{DEV} aumenta, as penalidades para objetos invisíveis são aumentadas, o que é um incentivo para ser completa e exibir todos os objetos salientes. Portanto, $\mathcal{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u})$ descreve a compensação entre a totalidade (exibição de tantos objetos quanto possível) e o refinamento (renderização dos objetos com maior resolução) de descrição de cena em quadros individuais.

[000151] Um ponto de visualização que maximiza $\mathcal{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u})$ aciona objetos visíveis mais próximo sdo centro e leva a separações maiores de objetos visíveis a partir do centro. Deixa-se $\widehat{\mathbf{v}}_{ki}$ ser o ponto de visualização ideal computado individualmente para cada quadro, isto é,

$$\widehat{\mathbf{v}}_{ki} = \arg\max_{\mathbf{v}_{ki}} \quad \mathcal{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u}). \qquad (7)$$

Alguns exemplos de $\widehat{\mathbf{v}}_{ki}$ sob diferentes resoluções de exibição são dados na figura 4.

2.3.2. Seleção de Visualizações de Câmera para um Dado Quadro

[000152] Apesar de serem usados dados de vários sensores, o que realmente importa não é o número de sensores ou o posicionamento dos mesmos, mas a presente maneira de utilizar esses pontos de visualização para produzir um ponto de visualização unificado virtual que tem um bom equilíbrio entre a ênfase de locais de detalhes e visão global de cenários. Uma vez que é difícil gerar vídeos de alta qualidade óptica livre com os métodos estado da técnica, consideramos apenas a seleção de uma visualização de câmera a partir de todas as câmeras apresentadas no presente caso para tornar esse sistema mais genérico. Define-se c = {c_i} como uma sequência de câmera, em que c_i denota o índice da câmera para o quadro i-ésimo. Um entendimento trivial na avaliação de uma visualização de câmera é que todos os objetos salientes devem ser claramente renderizados com poucas oclusões e

alta resolução. Para o i-ésimo quadro na k-ésimésima câmera , define-se a taxa de oclusão de objetos importantes como a razão normalizada da área unida de objetos salientes em relação à soma da área individual das mesmas, isto é,

$$\mathcal{R}_{ki}^{\mathrm{OC}}(\mathbf{v}_{ki}) = \frac{\mathcal{N}_{ki}(\mathbf{v}_{ki})}{\mathcal{N}_{ki}(\mathbf{v}_{ki}) - 1} \left\{ 1 - \frac{\mathcal{A}\left[\bigcup_m (\mathbf{o}_{kim} \cap \mathbf{v}_{ki})\right]}{\sum_m \mathcal{A}\left[\mathbf{o}_{kim} \cap \mathbf{v}_{ki}\right]} \right\}$$

em que $\bigcup_m x_m$ calcula a união de todas os blocos limite $\{x_m\}$. Usa-se $\mathcal{N}_{ki}(\mathbf{v}_{ki}) = \sum_{m, \mathbf{o}_{kim} \cap \mathbf{v}_{ki} \neq \phi} 1$ para representar o número de objetos visíveis dentro do ponto de visualização v_{ci}. Para normalizar a taxa de oclusão contra vários números de objetos salientes em diferentes quadros, reescalou-se $\mathcal{R}_{ki}^{\mathrm{OC}}(\mathbf{v}_{ki})$ para a faixa de 0 a 1, aplicando-se $\mathcal{N}_{ki}(\mathbf{v}_{ki})/(\mathcal{N}_{ki}(\mathbf{v}_{ki}) - 1)$. Define-se a proximidade de objetos salientes como áreas de pixel médio usadas para renderizar objetos, isto é,

$$\mathcal{R}_{ki}^{\mathrm{CL}}(\mathbf{v}_{ki}) = \log \frac{1}{\mathcal{N}_{ki}(\mathbf{v}_{ki})} \sum_m \mathcal{A}\left[\mathbf{o}_{kim} \cap \mathbf{v}_{ki}\right]. \qquad (8)$$

[000153] Também se define a totalidade desta visualização de câmera como a porcentagem de objetos salientes incluídos, isto é,

$$\mathcal{R}_{ki}^{\mathrm{CP}}(\mathbf{v}_{ki}, \mathbf{u}) = \frac{1}{\sum_m \mathcal{I}(\mathbf{o}_{kim}|\mathbf{u})} \sum_{\substack{m \\ \mathbf{o}_{cim} \cap \mathbf{v}_{ci} \neq \phi}} \mathcal{I}(\mathbf{o}_{kim}|\mathbf{u}). \qquad (9)$$

[000154] Portanto, o ganho de interesse da escolha da k-ésima câmera para o i-ésimo quadro é avaliado por $\mathcal{I}_i(k|\mathbf{v}_{ki}, \mathbf{u})$, que lê,

$$\mathcal{I}_i(k|\mathbf{v}_{ki}, \mathbf{u}) = w_k(\mathbf{u}) \mathcal{R}_{ki}^{\mathrm{CL}}(\mathbf{v}_{ki}) \mathcal{R}_{ki}^{\mathrm{CP}}(\mathbf{v}_{ki}, \mathbf{u}) \exp[-\frac{\mathcal{R}_{ki}^{\mathrm{OC}}(\mathbf{v}_{ki})^2}{2}]. \qquad (10)$$

[000155] Pondera-se o suporte da preferência de usuário atual para a câmera k por $w_k(\mathbf{u})$, que atribui um valor mais alto para a câmera k se isso é especificado pelo usuário e atribui um valor mais baixo de não é especificado. Define-se, então, a probabilidade de tomar-se a k-ésima

câmera para o i-ésimo quadro sob $\{v_{ki}\}$ como

$$P(c_i = k|\mathbf{v}_{ki}, \mathbf{u}) \equiv \frac{\mathcal{I}_i(k|\mathbf{v}_{ki}, \mathbf{u})}{\sum_j \mathcal{I}_i(j|\mathbf{v}_{ki}, \mathbf{u})} \qquad (11)$$

2.4. Geração de Ponto de Visualização Suave/Sequências de Câmera

[000156] Uma sequência de vídeo com pontos de visualização otimizados individualmente terá flutuações óbvias, o que leva a artefatos visuais desconfortáveis. Resolve-se esse problema gerando-se uma sequência de movimentação suave das câmeras e pontos de visualização com base na otimização individual dos mesmos. Usa-se um gráfico na figura 5 para explicar este procedimento de estimativa, que cobre dois passos de todo o sistema, isto é, a suavização dos movimentos orientada pela câmera e a geração de ponto de visualização de uma sequência de câmera suave com base em pontos de visualização determinados. Num primeiro momento, toma-se $\widehat{\mathbf{v}}_{ki}$ como dados observados e supõe-se que eles são saídas de ruído distorcido de alguns resultados subjacentes suaves v_{ki}. Usa-se inferência estatística para recuperar uma sequência ponto de visualização suave para cada câmera. Levando-se os ganhos de câmera dos pontos de visualização derivados em consideração, gera-se, então, uma sequência de câmera suave.

2.4.1. Suavização de Movimento de Ponto de Visualização Orientada pela Câmera

[000157] Parte-se da suavidade de movimento de ponto de visualização em um vídeo da mesma câmera. Existem duas forças contraditórias que acionam a otimização do movimento de ponto de visualização: por um lado, pontos de visualização otimizados devem estar mais próximos do ponto de visualização ideal de cada quadro individual; por outro lado, a suavidade entre os quadros de pontos de visualização impede que uma mudança drástica ocorra. Portanto, modela-se o movimento de ponto de visualização suave como um

campo aleatório de Markov (MRF) gaussiano, em que a suavidade orientada pela câmera é modelada como precedente do ponto de visualização de configuração, isto é,

$$P(\{\mathbf{v}_{ki}\}|\mathbf{u}) = \frac{\exp(-\frac{1}{2}\sum_i \sum_{j\in\mathcal{N}_i} \mathcal{H}^{\mathrm{PRI}})}{\sum_{\{\mathbf{v}_{ki}\}} \exp(-\frac{1}{2}\sum_i \sum_{j\in\mathcal{N}_i} \mathcal{H}^{\mathrm{PRI}})}, \tag{12}$$

$$\mathcal{H}_{ij}^{\mathrm{PRI}} = \frac{(v_{kix} - v_{kjx})^2}{2\sigma_{1x}^2} + \frac{(v_{kiy} - v_{kjy})^2}{2\sigma_{1y}^2} + \frac{(v_{kiw} - v_{kjw})^2}{2\sigma_{1w}^2}, \tag{13}$$

em que N_i são os arredores do i-ésimo quadro, enquanto uma distribuição condicional

$$P(\{\widehat{\mathbf{v}}_{ki}\}|\mathbf{u},\{\mathbf{v}_{ki}\}) = \prod_i \frac{\exp(-\mathcal{H}_i^{\mathrm{LL}})}{\sum_{\mathbf{v}_{ki}} \exp(-\mathcal{H}_i^{\mathrm{LL}})} \tag{14}$$

$$\mathcal{H}_i^{\mathrm{LL}} = \frac{(v_{kix} - \widehat{v}_{kix})^2}{2\beta_{ki}\sigma_{2x}^2} + \frac{(v_{kiy} - \widehat{v}_{kiy})^2}{2\beta_{ki}\sigma_{2y}^2} + \frac{(v_{kiw} - \widehat{v}_{kiw})^2}{2\beta_{ki}\sigma_{2w}^2} \tag{15}$$

descreve o ruído que produz os resultados finais. Adiciona-se um parâmetro β_{ki} para controlar a flexibilidade do quadro atual em suavização. Um β_{ki} menor pode ser ajustado para aumentar a tendência do quadro atual em aproximação do ponto de visualização ideal localmente do mesmo. A estimação de pontos de visualização ideais $\{v_{ki}\}$ é feita maximizando-se a probabilidade posterior de $\{v_{ki}\}$ sobre $\{\widehat{\mathbf{v}}_{ki}\}$ observado, isto é, $P(\{v_{ki}\}|u, \{\widehat{\mathbf{v}}_{ki}\})$, que é expresso por uma distribuição canônica de Gibbs [12], isto é,

$$P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\widehat{\mathbf{v}}_{ki}\}) = \frac{\exp\{-\mathcal{H}^{\mathrm{V}}\}}{\sum_{\{\mathbf{v}_{ki}\}} \exp\{-\mathcal{H}^{\mathrm{V}}\}} \tag{16}$$

com

$$\mathcal{H}^{\mathrm{V}} = \frac{1}{2}\sum_i \sum_{j\in\mathcal{N}_i} \mathcal{H}^{\mathrm{PRI}} + \sum_i \mathcal{H}_i^{\mathrm{LL}}. \tag{17}$$

[000158] Em física estatística, a configuração ideal da probabilidade posterior mais alta é determinada maximizando-se a energia livre [13] a seguir:

$$\mathcal{F}^{\mathrm{V}} = \langle \mathcal{H}^{\mathrm{V}} \rangle - \langle \ln P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\widehat{\mathbf{v}}_{ki}\}) \rangle \tag{18}$$

em que $\langle x \rangle = \sum_{\{\mathbf{v}_{ki}\}} xP(\{\mathbf{v}_{ki}\}|\mathbf{u}, \{\widehat{\mathbf{v}}_{ki}\})$ é o valor de expectativa de uma quantidade x. Formam-se os critérios a seguir considerando-se a restrição de normalização de $P(\{v_{ki}\}|u, \{\widehat{\mathbf{v}}_{ki}\})$ como

$$\mathcal{L}^{\mathrm{V}} = \mathcal{F}^{\mathrm{V}} + \eta(1 - \sum_{\{\mathbf{v}_{ki}\}} P(\{\mathbf{v}_{ki}\}|\mathbf{u}, \{\widehat{\mathbf{v}}_{ki}\})) \tag{19}$$

em que η é um multiplicador lagrangeano. Usa-se aproximação de campo médio [13] que supõe que $P(\{\mathbf{v}_{ki}\}|\mathbf{u}, \{\widehat{\mathbf{v}}_{ki}\}) \approx \prod_i P(\mathbf{v}_{ki}|\mathbf{u}, \{\widehat{\mathbf{v}}_{ki}\})$ dissocia correlações de dois corpos. Tomando-se o diferencial de $\mathcal{L}^{\mathrm{VP}}$ em relação a $P(v_{kix}|u, \{\widehat{\mathbf{v}}_{ki}\})$ e ajustando-o para zero, obtém-se a estimativa ideal para $P(v_{kix}|u, \{\widehat{\mathbf{v}}_{ki}\})$ como

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}^{\mathrm{VP}}}{\partial P(v_{kix}|\mathbf{u}, \{\widehat{\mathbf{v}}_{ki}\})} \\ &= \sum_{j \in \mathcal{N}_i} \frac{v_{kix}^2 - 2v_{kix}\langle v_{kjx}\rangle^T}{2\sigma_{1x}^2} + \frac{(v_{kix} - \widehat{v}_{kix})^2}{2\beta_{ki}\sigma_{2x}^2} + \ln P(v_{kix}|\mathbf{u}, \{\widehat{\mathbf{v}}_{ki}\}) - \eta. \end{aligned} \tag{20}$$

[000159] Portanto, tem-se a probabilidade posterior

$$P(v_{kix}|\mathbf{u}, \{\widehat{\mathbf{v}}_{ki}\}) \propto \exp\left\{ -\sum_{j \in \mathcal{N}_i} \frac{(v_{kix} - \langle v_{kjx}\rangle)^2}{2\sigma_{1x}^2} - \frac{(v_{kix} - \widehat{v}_{kix})^2}{2\beta_{ki}\sigma_{2x}^2} \right\}. \tag{21}$$

[000160] Já que uma distribuição gaussiana cujo valor médio tem a probabilidade maximizada, o ponto de visualização ideal para v_{kix} é resolvido como

$$v_{kix}^* = \langle v_{kix}\rangle = \frac{\sum_{j \in \mathcal{N}_i} \sigma_{2x}^2 \beta_{ki} \langle v_{kjx}\rangle + \widehat{v}_{kix}\sigma_{1x}^2}{\sum_{j \in \mathcal{N}_i} \sigma_{2x}^2 \beta_{ki} + \sigma_{1x}^2}, \tag{22}$$

$$v_{kiy}^* = \langle v_{kiy}\rangle = \frac{\sum_{j \in \mathcal{N}_i} \sigma_{2y}^2 \beta_{ki} \langle v_{kjy}\rangle + \widehat{v}_{kiy}\sigma_{1y}^2}{\sum_{j \in \mathcal{N}_i} \sigma_{2y}^2 \beta_{ki} + \sigma_{1y}^2}, \tag{23}$$

$$v_{kiw}^* = \langle v_{kiw}\rangle = \frac{\sum_{j \in \mathcal{N}_i} \sigma_{2w}^2 \beta_{ki} \langle v_{kjw}\rangle + \widehat{v}_{kiw}\sigma_{1w}^2}{\sum_{j \in \mathcal{N}_i} \sigma_{2w}^2 \beta_{ki} + \sigma_{1w}^2} \tag{24}$$

com os resultados ideais para v_{kiy} e v_{kiw} também dados por

derivação semelhante. Usa-se $\mathbf{v}_{ki}^*$ nas seções a seguir para denotar o ponto ideal representado por v_{kix}^*, v_{kiy}^* e v_{kiw}^*.

2.4.2. Suavização de Sequência de Câmera

[000161] Uma sequência de câmera suave será gerada a partir pontos de visualização determinados. Para simplificar, usa-se $\rho_{ki} \equiv \log P(c_i = k|\mathbf{v}_{ki}^*, \mathbf{u})$ para encurtar a formulação, que é computada usando-se a Eq.11. Temos de compensar entre minimizar a troca de câmera e maximizar o ganho global das câmeras. Usa-se outro MRF para modelar esses dois tipos de forças. A suavidade da sequência de câmera é modelada por uma distribuição canônica de Gibbs, em que se lê:

$$P(\{c_i\}|\{\mathbf{v}_{ki}^*\}, \mathbf{u}) = \frac{\exp(-\mathcal{H}^{\mathrm{C}})}{\sum_{\{c_i\}} \exp(-\mathcal{H}^{\mathrm{C}})}, \tag{25}$$

com

$$\mathcal{H}^{\mathrm{C}} = -\gamma \sum_{i,k} \delta_{c_i,k} \rho_{ki} - \frac{(1-\gamma)}{2} \sum_i \sum_{j \in \mathcal{N}_i} \alpha_{ij} \delta_{c_i,c_j}. \tag{26}$$

em que α_{ij} é um parâmetro para normalizar a força relativa de suavização em relação ao tamanho dos arredores, que lê

$$\alpha_{ij} = \frac{K}{|j-i| \sum_{l \in \mathcal{N}_i} \frac{1}{|l-i|}}. \tag{27}$$

γ é um hiperparâmetro para controlar a força de suavização. Usa-se a aproximação de campo médio que supõe que $P(\{c_i\}|\{\mathbf{v}_{ki}^*\}, \mathbf{u}) \approx \prod_i P(c_i|\{\mathbf{v}_{ki}^*\}, \mathbf{u})$ novamente para alcançar a estimativa ideal. Omite-se a derivação detalhada e apenas se mostra o resultado final, que deriva que a probabilidade marginal de se tomar a câmera k para o i-ésimo quadro é

$$P(c_i = k|\{\mathbf{v}_{ki}^*\}, \mathbf{u}) = \langle \delta_{c_i,k} \rangle^{\mathrm{C}} = \frac{\exp\left\{(1-\gamma) \sum_{j \in \mathcal{N}_i} \alpha_{ij} \langle \delta_{c_j,k} \rangle^{\mathrm{C}} + \gamma \rho_{ki}\right\}}{\sum_k \exp\left\{(1-\gamma) \sum_{j \in \mathcal{N}_i} \alpha_{ij} \langle \delta_{c_j,k} \rangle^{\mathrm{C}} + \gamma \rho_{ki}\right\}}, \tag{28}$$

em que $\langle x \rangle^{C} = \sum_{\{c_i\}} xP(\{c_i\}|\{\mathbf{v}_{ki}^{*}\}, \mathbf{u})$ é o valor de expectativa de uma quantidade x. O processo de suavização é realizado iterando-se a regra de ponto fixo a seguir até que se alcance a convergência,

$$\langle \delta_{c_i,k} \rangle^{C} = \frac{\exp\left\{(1-\gamma)\sum_{j\in\mathcal{N}_i}\alpha_{ij}\langle \delta_{c_j,k}\rangle^{C} + \gamma\rho_{ki}\right\}}{\sum_{k}\exp\left\{(1-\gamma)\sum_{j\in\mathcal{N}_i}\alpha_{ij}\langle \delta_{c_j,k}\rangle^{C} + \gamma\rho_{ki}\right\}}. \qquad (29)$$

Após a convergência, seleciona-se a câmera que maximiza $\langle \delta_{ci,k} \rangle^{C}$, isto é,

$$c_i^{*} = \arg\max_{c_i} \langle \delta_{c_i,k} \rangle^{C}. \qquad (30)$$

3. Resultados experimentais e Discussões

[000162] Organizou-se uma aquisição de dados na cidade de Namur, Bélgica, em ambiente de jogo real, onde sete câmeras foram utilizadas para gravar quatro jogos. Todos os vídeos são distribuídos publicamente no site do projeto APIDIS [1] e explicação mais detalhada sobre as configurações de aquisição pode ser encontrada na Ref. [14]. Resumidamente, as câmeras são todas câmeras Arecont Vision IP AV2100M, cujas posições na quadra de basquete são mostradas na figura 6. As lentes olho de peixe usadas para as câmeras visualização superior são lentes Fujinon FE185C086HA-1. Os quadros a partir de sete câmeras foram todos enviados para um servidor, em que o tempo de chegada de cada quadro foi usado na sincronização de câmeras diferentes. Na figura 7, são dadas amostras de imagens de todas as sete câmeras. Devido ao número limitado de câmeras, ajusta-se a maioria das câmeras para cobrir a quadra esquerda. Como resultado, será focada principalmente a quadra esquerda para investigar o desempenho do nosso sistema de produção personalizada de vídeos de esportes.

[000163] Já que a produção de vídeo ainda carece de uma regra objetiva

para a avaliação de desempenho, muitos parâmetros são determinados heuristicamente, com base em avaliação subjetiva. Definem-se vários objetos salientes e a relação entre o tipo de objeto e o interesse é dada na Tabela 1. Se o usuário mostra interesse especial por um objeto saliente, a ponderação será multiplicada por um fator 1,2. Para a suavização de ponto de visualização, ajusta-se β_{ki} para 1 para a suavização de ponto de visualização orientado pela câmera nos experimentos a seguir. Deixa-se $\sigma_{1x} = \sigma_{1y} = \sigma_{1w} = \sigma_1$ e $\sigma_{2x} = \sigma_{2y} = \sigma_{2w} = \sigma_2$.

[000164] Um videoclipe curto, com cerca de 1.200 quadros é usado para demonstrar as características comportamentais do presente sistema, especialmente sua adaptabilidade em resolução de exibição limitada. Tabela 1: Ponderação de Diferentes Objetos Salientes

Tipo de objeto	Bola	Jogador	Juiz	Cesta	Banco do treinador	Outros
$\mathcal{I}(\mathbf{o}_{kim}\|\mathbf{u})$	2	1	0,8	0,6	0,4	0,2

[000165] Este clipe abrange três períodos de posse de bola e inclui cinco eventos no total. Na figura 8, mostram-se intervalos de tempo de todos os eventos, cuja maioria dos momentos de destaque também são marcados por linhas vermelhas sólidas. Na versão final deste projeto, devem ser gerados metadados através da compreensão automática da cena. No presente caso, que foca na produção personalizada, primeiro, avaliam-se os métodos em metadados coletados manualmente. Será explorada a eficiência de cada etapa de processamento individual do presente método, e então será feita uma avaliação global com base em saídas geradas no fim. Devido à limitação de página, os resultados numéricos são dados e representados por gráficos no presente caso, enquanto os vídeos correspondentes só estão disponíveis no site do projeto APIDIS. [1] Os revisores são convidados a fazer *download* das amostras de vídeo produzido com base nas preferências de usuário diferentes para avaliar subjetivamente a eficiência e relevância da abordagem proposta.

[000166] Começa-se a investigar o desempenho do presente método de seleção individual de pontos de visualização. As sequências orientadas pela câmera de pontos de visualização determinados automaticamente pelo presente método são colocadas em uma tabela na figura 9, em que larguras de pontos de visualização ideal em três resoluções de exibição diferentes, isto é, 160x120, 320x240 e 640x480 são exibidos para todas as sete câmeras. A suavização ponto de visualização fraco tem sido aplicada para melhorar a legibilidade dos vídeos gerados, em que a força de suavização é ajustada como $\sigma_2/\sigma_1 = 4$. A partir da comparação dos resultados em três resoluções de tela diferentes, a constatação mais óbvia é que uma resolução maior de exibição leva a uma maior largura de ponto de visualização enquanto uma resolução inferior de exibição prefere um tamanho ponto de visualização menor, assim como esperado de presente critério de seleção. Uma vez que as câmeras 1, 6 e 7 apenas cobrem metade da quadra, seus tamanhos de pontos de visualização serão fixados quando todos os jogadores estiverem na outra metade da quadra, o que explica os segmentos lisos em suas respectivas subgráficos. A partir dos dados de vídeo, podemos confirmar mesmo quando a resolução de exibição é muito baixa, o sistema extrairá um ponto de visualização de um tamanho razoável, onde a bola é dimensionada para um tamanho visível. Embora em alguns quadros apenas a bola seja exibida para a menor resolução de tela, isso não será um problema porque os quadros serão filtrados pela seleção pós-câmera.

[000167] Os tamanhos de ponto de visualização de sequências suavizadas sob suavização de forças diferentes são comparados na figura 10(a). Com todos os outros parâmetros são os mesmos, a razão de σ_2 para σ_1 é regulada para todos os cinco casos. Uma maior proporção de σ_2 para σ_1 corresponde a um processo de suavização mais forte, enquanto uma menor razão significa suavização mais fraca. Quando $\sigma_2/\sigma_1 = 1$ em que uma suavização muito fraca é aplicada, obtemos uma sequência

bastante acidentada, o que resulta em um vídeo tremulante com muitos movimentos do ponto de visualização drástico. Com o aumento da razão σ_2/σ_1, a curva do movimento de ponto de visualização começa a ter picos menos nítidos, o que fornece conteúdo perceptualmente mais confortável. Outra observação importante é que as sequências geradas serão bastante diferentes da seleção inicial com base em informações de saliência, se uma suavização muito forte tiver sido realizada com uma σ_2/σ_1 muito grande. Isso causará problemas como o jogador favorito ou a bola fora do ponto de visualização suavizado. A relação σ_2/σ_1 deve ser determinada considerando-se a compensação entre pontos de visualização localmente otimizados e sequências de ponto de visualização globalmente suavizado. Constatando-se visualmente os vídeos gerados, considera-se que os resultados com um nivelamento fraco tais como $\sigma_2/\sigma_1 = 4$ já estão perceptivelmente aceitável, visualizando-se o vídeo de demonstração.

[000168] Em seguida, verifica-se o algoritmo de suavização para a sequência de câmera. As sequências de câmeras suavizadas sob várias forças de suavização γ são representados na figura 10(b). O processo de suavização leva a probabilidade definida na Eq.11 como valores iniciais, e repete a regra de atualização de ponto fixo com arredores do tamanho e meio até a convergência. A sequência de câmera sem suavização corresponde ao mais alto subgráfico na figura 10(b), enquanto a sequência com a suavização mais forte é plotada no subgráfico do fundo. É claro que existem muitas opções de câmera drásticas em uma sequência não suavizada, o que leva a artefatos ainda mais desagradáveis do que as posições ponto de visualização flutuantes, como se pode ver nos vídeos gerados. Portanto, prefere-se suavização forte em sequências de câmeras e será usado $\gamma = 0,8$ nos experimentos a seguir.

[000169] Na figura 11(a) e (b), comparamos pontos de visualização e câmeras em sequências geradas em relação a resoluções de exibição diferentes, respectivamente. De cima para baixo, mostram-se

resultados para resolução de exibição u^{DEV}= 160, 320 e 640 em três subgráficos. Quando a mesma câmera é selecionada, observamos que um ponto de visualização maior é preferido por uma resolução de exibição mais alta. Quando câmeras diferentes são selecionadas, é preciso considerar a posição da câmera selecionada e a posição do ponto de visualização determinado ao avaliar a cobertura da cena de saída. Mais uma vez, confirma-se que os tamanhos dos pontos de visualização aumentam quando a resolução da exibição se torna maior. Antes do 400º quadro, o evento ocorre na quadra direito. Verificou-se que a 3ª câmera, isto é, a de visualização superior com lente ângulo grande, aparece mais vezes na sequência de u^{DEV} = 640 do que na de u^{DEV} = 160 e os pontos de visualização da mesma também são mais amplos, o que prova que uma maior resolução prefere uma visão mais ampla. Embora a 2ª câmera apareça muitas vezes em u^{DEV} = 160, pontos de visualização correspondentes da mesma são muito menores em largura. Esta câmera é selecionada porque fornece uma visão lateral da quadra direita com os objetos salientes reunidos mais próximo do que outros pontos de visualização da câmera devido à geometria projetiva. Pela mesma razão, a 3ª câmera aparece mais frequentemente em u^{DEV} = 160 quando o jogo se move para a quadra à esquerda a partir do 450º quadro para o 950º quadro. Esta conclusão é confirmada por miniaturas na figura 12, em que frames de índice de 100 a 900 estão dispostos em uma tabela para as três resoluções de exibição acima.

[000170] Devido ao fato de que diferentes câmeras foram selecionadas, os pontos de visualização determinados abaixo de u^{DEV} = 640 parecem estar mais próximo do que os abaixo de u^{DEV} = 320 nas cinco últimas colunas da figura 12. Isso reflete a inconsistência de importância relativa entre a totalidade e a proximidade na seleção do ponto de visualização.

[000171] Uma vez que apenas pontos centrais de objetos salientes

são calculados nos critérios para a seleção ponto de visualização, pontos de visualização de resultados não são contínuos em resolução diferente. Apesar de a câmera 7 ser semelhante à câmera 1 com ampliação linear, pontos de visualização ideais das mesmas podem ter ênfases diferentes em totalidade e proximidade. Esta consistência também existe na seleção de câmeras e pontos de visualização separados. Se a seleção de ponto de visualização se concentra mais na proximidade e seleção de câmera se concentra mais na totalidade, uma pequena área de corte na câmera 7 será selecionada primeiro na seleção de ponto de visualização para u^{DEV} = 320, e depois será rejeitada na seleção da câmera seguinte devido à insuficiência de totalidade. Um teste subjetivo nos ajudará a regular a ponderação relativa de totalidade e de proximidade. É mais importante implantar a seleção simultânea de pontos de visualização e câmeras, o que requer inclusão de informações de posição das câmeras como o uso de homografia, e um critério analiticamente solucionável para a seleção de ponto de visualização. Estas questões são o principal trabalho do presente documento em um futuro próximo.

[000172] Em todos os experimentos acima, nenhuma narrativa de preferências do usuário está incluída. Se o usuário tiver interesse especial em uma determinada visualização da câmera, poderíamos atribuir uma ponderação de $w_k(u)$ maior para a câmera especificada. No presente caso, define-se $w_k(u) = 1.0$ para câmeras não especificadas e $w_k(u) = 1.2$ para uma câmera especificada pelo usuário. Comparam-se sequencias de câmera sob diferentes preferências na figura 13. Como pode-se facilmente ver no gráfico, uma câmera aparece mais vezes quando é especificada, o que reflete a preferência do usuário em visualizações da câmera. Quanto à preferência do usuário em equipes ou jogadores, a diferença entre pontos de visualização com e sem as preferências do usuário é difícil dizer sem uma regra de avaliação bem

definida, porque todos os jogadores estão sempre juntos durante o jogo. Na verdade, há maior interesse em refletir as preferências do usuário sobre os jogadores ou equipes, extraindo-se os quadros relativos. Portanto, omitem-se os resultados na seleção do jogador ou da equipe, mas os mesmos são explorados mais tarde, juntamente com os resultados do trabalho futuro na sumarização de vídeo.

4. Comentários de Conclusão

[000173] Um sistema autônomo para produzir vídeos personalizados a partir de múltiplas visualizações de câmeras foi proposto. Discutiu-se a adaptação automática de pontos de visualização com respeito à resolução de exibição e conteúdos de cenário, a fusão de dados dentro de múltiplas visualizações de câmeras, e a suavidade do ponto de visualização e as sequências de câmera para contar a história de forma fluente. Há quatro principais vantagens dos métodos: 1) Baseada em semântica. Ao invés de usar recursos baixos tal como bordas ou aparência de quadros, a produção é baseada no entendimento semântico do cenário, que pode lidar com a preferência de usuário semântica mais complexa. 2) Eficaz computacionalmente. Toma-se uma estratégia de dividir e conquistar e considera-se um processamento hierárquico, que é eficaz em lidar com conteúdos de vídeo longos porque seu tempo total é quase linearmente proporcional ao número de eventos incluídos. 3) Genericidade. Como os submétodos em cada etapa individual são todos independentes da definição dos interesses e objetos salientes, essa moldura não é limitada a vídeos de basquetebol, mas capaz de ser aplicada a outros cenários controlados. 4) Não supervisionado. Embora haja alguns parâmetros sobrando para definição pelos usuários, o sistema é não supervisionado.

Apêndice 2

[000174] Os métodos apresentados neste documento têm como objetivo detectar e reconhecer jogadores em um campo de esporte,

baseando-se em um conjunto distribuído de câmeras sincronizadas livremente. A detecção supõe a verticalidade de jogador e soma a projeção cumulativa das máscaras de atividade de primeiro plano das visualizações múltiplas em um conjunto de planos que são paralelos ao plano de fundo. Depois do resumo, os valores de projeção grande indicam a posição do jogador no plano básico. Esta posição é usada como uma âncora para a caixa delimitadora de jogador projetada em cada uma das visualizações. Dentro desta caixa delimitadora, as regiões fornecidas pela segmentação de desvio de meio são classificadas baseando-se em atributos contextuais, por exemplo, posição e tamanho relativos, para selecionar os que são prováveis de corresponder a um dígito. A normalização e a classificação das regiões selecionadas, então, fornecem o número e identidade do jogador. Como o número de jogador pode ser lido somente quando se volta em direção à câmera, o rastreamento baseado em gráfico é considerado para propagar a identidade de um jogador ao longo de sua trajetória.

I. Introdução

[000175] Na sociedade atual, a produção de conteúdo e o consumo de conteúdo são confrontados com uma mutação fundamental. Duas tendências complementares são observadas. Pro um lado, os indivíduos tornam-se mais e mais heterogêneos no modo em que acessam o conteúdo. Estes querem acessar o conteúdo dedicado através de um serviço personalizado capaz de fornecer o que é de interesse quando quiserem e através do canal de comunicação de sua escolha. Por outro lado, os indivíduos e organizações conseguem acesso mais fácil às facilidades técnicas exigidas para estar envolvido na criação de conteúdo e processo de difusão.

[000176] Neste documento, descrevem-se as ferramentas de análise de vídeo que participam das evoluções futuras da indústria de produção de conteúdo em direção a infraestruturas automatizadas permitindo que

o conteúdo seja produzido, armazenado e acessado a baixo custo e de uma maneira dedicada e personalizada. Mais especificamente, a aplicação rastreada considera o resumo personalizado e autônomo de eventos de esporte, sem a necessidade de processos feitos a Mao custosos. No cenário de aplicação suportado pelo conjunto de dados fornecido, os sensores de aquisição cobrem uma quadra de basquetebol. A interpretação e análise distribuída da cena são, então, exploradas para decidir o que mostrar sobre o evento e como mostrar isto, de modo a produzir um vídeo composto de um subconjunto valioso dos fluxos fornecidos por cada câmera individual. Em particular, a posição dos jogadores fornece a entrada exigida para acionar a seleção autônoma dos parâmetros de ponto de visualização [5], enquanto que a identificação e rastreamento dos jogadores detectados suportam a personalização do resumo, por exemplo, através de realce e/ou repetição das ações do jogador preferenciais.

Parte deste trabalho foi financiada pelo projeto europeu FP7 da APIDIS e pelo NSF belga.

II. Visão geral do sistema

[000177] Para demonstrar o conceito da produção personalizada e autônoma, o projeto de pesquisa FP7 APIDIS Europeu (www.aipidis.org) distribuiu um sistema de aquisição de câmeras múltiplas em volta de uma quadra de basquetebol. A definição de aquisição consiste em um conjunto de 7 câmeras IP calibradas, cada uma coletando 2 quadros de Mpixels a uma taxa maior que 20 quadros/segundo. Depois de uma sincronização temporária aproximada dos fluxos de vídeo, este documento investiga como aumentar o conjunto de dados baseando-se na detecção, rastreamento e reconhecimento dos jogadores.

[000178] A figura 1 avalia a abordagem proposta para computar e marcar rastreamentos de jogadores. Depois da detecção de multivisualização em

conjunto de pessoas que estão no campo básico em cada instante, um algoritmo de rastreamento baseado em gráfico iguala as posições que estão suficientemente próximas – em posição e aparência – entre os quadros sucessivos, através disso definindo um conjunto de rastreamentos disjuntos interrompidos, também chamados de rastreamentos parciais. Em paralelo, conforme retratado na figura 5, a classificação e análise de imagem são consideradas para cada quadro de cada visualização, para reconhecer os dígitos que aparecem potencialmente nas camisas dos alvos detectados. Esta informação é, então, agregada através do tempo para marcar os rastreamentos parciais.

[000179] As principais contribuições deste documento têm que ser achadas na solução de detecção de pessoas proposta, que é retratada na figura 2. Em resumo, o processo de detecção segue uma abordagem de baixo para cima para extrair grupos mais densos em um mapa de ocupação de plano básico que é computado baseando-se na projeção de máscaras de atividade de primeiro plano. Duas melhoras fundamentais são propostas em comparação ao estado da técnica. Primeiro, a máscara de atividade de primeiro plano não é somente projetada no plano básico, como recomendada em [9], mas em um conjunto de planos que são paralelos ao fundo. Segundo, uma heurística original é implantada para lidar com oclusões e aliviar as detecções falsas que ocorrem na interseção das mascaras projetadas a partir de silhuetas dos jogadores distintas pelas visualizações distintas. As simulações demonstram que essas duas contribuições melhoram bem significantemente o desempenho de detecção.

[000180] O resto do documento é organizado conforme segue. As Seções III, V e IV, respectivamente, focam na detecção, rastreamento e reconhecimento dos problemas. Os resultados experimentais são apresentados na Seção VI para validar a abordagem. A Seção VII conclui.

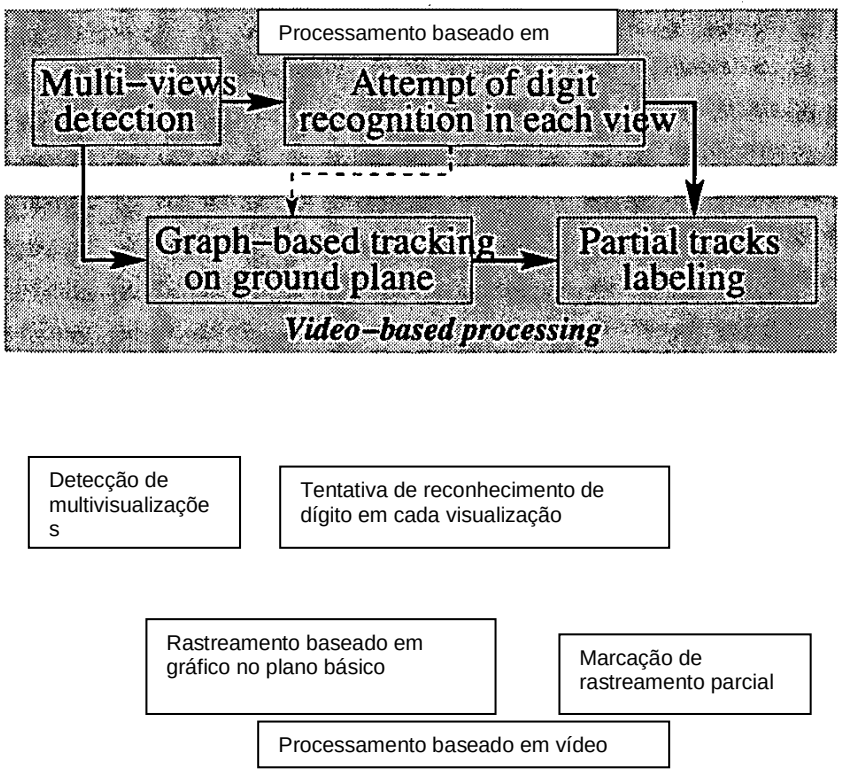


figura 1. Computação de rastreamentos de jogadores e encadeamento de marcação. A seta pontilhada reflete a inclusão opcional dos resultados de reconhecimento de dígito dentro do modelo de aparência considerado para rastreamento.

III. Detecção de pessoas de multivisualizações

[000181] Manter o rastreamento das pessoas que obstruem umas às outras com uso de um conjunto de câmeras sincronizadas livremente, estacionárias, calibradas e amplamente espaçadas C é uma questão importante porque esse tipo de organização é comum para aplicações que variam de evento (de esporte) que relatam à vigilância em espaço publico. Nessa seção, considera-se uma abordagem de detecção de mudança para inferir a posição dos jogadores no campo básico, em cada instante.

A. Trabalho relacionado

[000182] A detecção de pessoas a partir das máscaras de atividade de primeiro plano em visualizações múltiplas tem sido investigada em detalhes nos últimos anos. Diferenciam-se duas classes de abordagens.

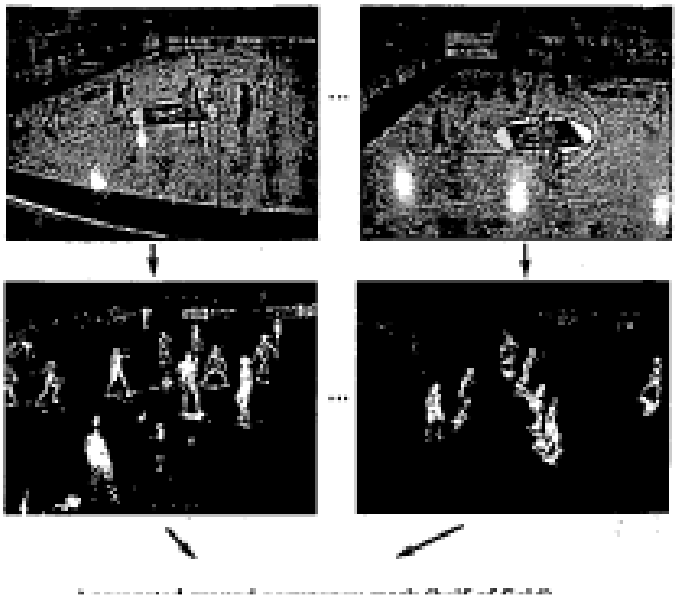
[000183] Por outro lado, os autores em [9] e [10] adotam uma

abordagem de baixo pra cima e projetam os pontos dos arredores de primeiro plano (silhuetas subtraídas de segundo plano) de cada visualização para o plano básico. Especificamente, os mapas de probabilidade de mudança computados em cada visualização são deturpados para o plano básico nas homografias que têm sido inferidas offline. Os mapas projetados são, então, multiplicados juntos e limiarizadas para definir os remendos do plano básico para qual a aparência mudou em comparação ao modelo de segundo plano e de acordo com o algoritmo de detecção de mudança.

[000184] Por outro lado, os trabalhos em [2], [7] e [1] adotam uma abordagem de cima para baixo. Estes consideram uma rede de pontos no plano básico e estimam as probabilidades de ocupação de cada ponto na rede baseando-se na retroprojeção de algum tipo de modelo gerativo em cada uma das múltiplas visualizações calibradas. Portanto, esses mesmos começam a partir do plano básico e validam a hipótese de ocupação com base no modelo de aparência associado em cada uma das visualizações. As abordagens propostas nessa segunda categoria diferem principalmente com base no tipo de modelo gerativo que se considera (retângulo ou dicionário aprendido) e no modo que se decide sobre a ocupação em cada ponto da rede (combinação de classificadores baseados em visualizações múltiplas em [2], rede de ocupação probabilística inferida das máscaras de subtração de segundo plano em [7] e mapa de ocupação binária limitada de escassez [1]).

[000185] A primeira categoria de métodos tinha a vantagem de ser eficaz computacionalmente, como a decisão sobre a ocupação de plano básico é diretamente tomada a partir da observação da projeção das máscaras de detecção de mudança das diferentes visualizações. Em contraste, a complexidade da segunda categoria dos algoritmos depende do número pontos de plano básico a ser investigado (escolhido para limitar a área a ser monitorada) e da carga computacional associada à validação de cada

hipótese de ocupação. Este processo de validação geralmente envolve a retroprojeção de um padrão de mundo 3D em cada uma das visualizações. Com esse respeito, nota-se que, devido distorções de projeção e lente, mesmo o entortamento de padrões retangulares 3D simples geralmente resulta em padrões não retangulares em cada uma das visualizações, através disso prevenindo o uso de técnicas de imagens integrais eficazes. Portanto, na maioria dos casos práticos, o segundo tipo de abordagem é significantemente mais complexo que o primeiro. Em retorno, oferece-se desempenho aumentado desde que não somente os pés, mas a silhueta do objeto inteiro seja considerada para fazer uma decisão.



Máscara de ocupação básica agregada (metade do campo)

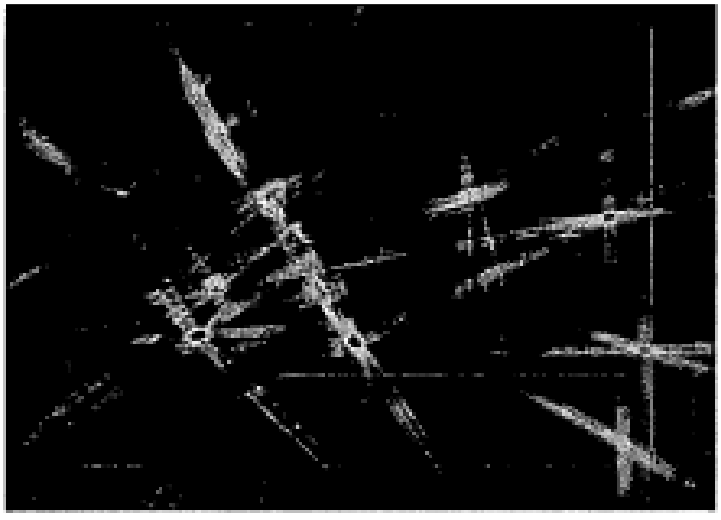


figura 2. Detecção de pessoas de multivisualizações. As máscaras de primeiro plano dão projetadas em um conjunto de planos que são paralelos ao plano básico para definir um mapa de ocupação de plano básico, a partir do qual a posição dos jogadores é inferida diretamente.

[000186] A abordagem é uma tentativa de tirar o melhor proveito de ambas as categorias. Propõe-se uma abordagem de baixo para cima computacionalmente eficaz que é capaz de explorar o conhecimento a priori inteiro que se tem sobre a silhueta do objeto. Especificamente, a computação de baixo para cima da máscara de ocupação básica descrita na Seção III-B explora o fato de que a base da silhueta está situada no plano básico (similarmente às soluções de cima para baixo anteriores), mas também que a silhueta é uma forma vertical grosseiramente retangular (que foi previamente reservada para as abordagens de cima para baixo). Como uma segunda contribuição, a Seção III-C propõe uma heurística ávida simples para resolver a interferência que ocorre entre as silhuetas projetadas a partir de visualizaçãos distintas por objetos distintos. Os resultados experimentais revelam que essa interferência foi a fonte de muitas falsas identificações enquanto interferem as posições reais dos objetos da máscara de ocupação básica. Até agora, esse fenômeno foi somente levado em consideração pela abordagem de cima para baixo descrita em [7], através de uma aproximação repetitiva complexa das probabilidades posteriores em conjunto da ocupação. Em contraste, embora próximo, a abordagem parece ser tanto eficiente e eficaz.

B. Abordagem proposta: computação de máscara de ocupação de plano básico

[000187] Similar a [9], [10], [7], [1], a abordagem realiza a detecção de mudança de visualização independentemente de cada visualização para computar um mapa de probabilidade de mudança. Para este

propósito, um algoritmo de subtração de segundo plano convencional com base na mistura de modelagens gaussianas é implantado. Para a fusão das silhuetas de primeiro plano binário resultantes, o método projeta estas para construir uma máscara de ocupação básica. Entretanto, em contraste com as abordagens de baixo para cima anteriores [9], [10], não se considera a projeção no plano básico somente, mas em um conjunto de planos que são paralelos ao plano básico e corta-se o objeto para detectar as alturas diferentes. Sob a suposição de que o objeto de interesse mantém-se aproximadamente verticalmente, a projeção cumulativa de todas essas projeções em um plano de visualização de topo virtual reflete realmente a ocupação de plano básico. Essa seção explica como a máscara associada a cada visualização é computada. A próxima seção investiga como consolidar a informação fornecida pelas visualizações múltiplas para detectar pessoas.

[000188] Formalmente, a computação da máscara de ocupação básica G_i associadas à i-ésima visualização é descrita conforme segue. Em um tempo dado, a i-ésima visualização é a fonte de uma imagem de silhueta subtraída de segundo plano binário $B_i \in \{0,1\}^{M_i} \cdot\cdot$, onde M_i é o número de pixels da câmera i, $1 \leq i \leq C$. Conforme explicado acima, B_i é projetado em um conjunto fr planos de referência L que são definidos a serem paralelos ao plano básico, em intervalos de altura regulares, e até a altura típica de um jogador. Portanto, para cada visualização i, define-se G_i^j a ser a projeção da i-ésima máscara binária no j-ésimo plano. G_i^j é computado aplicando-se o entortamento homográfico de cada pixel da comera i a sua posição correspondente no j-ésimo plano de referência, com $0 \leq j < L$. Pela construção, os pontos de B_i que são marcados a 2 por causa da presença de um jogador no j-ésimo plano de referência projetam-se para a posição de visualização de topo correspondente em G_i^j. Portanto, o somatório G_i das projeções obtidas

em diferentes alturas e a partir de diferentes visualizações é esperado para posições de visualização de topo de destaque de jogadores que permanecem verticalmente.

[000189] Conforme L aumenta, a computação de G_i em uma posição básica x tende em direção da integração da projeção de B_i em um segmento vertical ancorado em x. Essa integração pode ser equivalentemente computada em B_i, ao longo da retroprojeção do segmento vertical. Para acelerar adicionalmente as computações, observa-se que, através da transformação apropriada de B_i, é possível formar os domínios de integração retroprojetada de modo que estes correspondam aos segmentos das linhas verticais na visualização transformada, através disso fazendo a computação dos integrais particularmente eficazes através do princípio de imagens integrais. A figura 3 ilustra essa transformação específica para uma visualização em particular. A transformação foi designada para abordar um objeto duplo. Primeiro, os pontos do espaço 3D localizado na mesma linha vertical têm que ser projetados na mesma coluna na visualização transformada (ponto de desaparecimento vertical na infinidade). Segundo, os objetos verticais que permanecem no fundo e cujos pés são projetados na mesma linha horizontal da visualização transformada têm que manter as mesmas razões de alturas projetadas.

[000190] Uma vez que a primeira propriedade é conhecida, os pontos 3D que pertencem à linha vertical que permanecem acima de um ponto dado a partir do plano básico simplesmente projetam-se na coluna da visualização transformada que permanece acima da projeção do ponto de plano básico 3D. Portanto, $G_i(x)$ é simplesmente computado como o integral da visualização transformada sobre seu segmento retroprojetado. A preservação da altura o longo das linhas da visualização transformada ainda simplifica adicionalmente as computações.

[000191] Para visualizações laterais, essas duas propriedades podem ser alcançadas movendo-se virtualmente – através transformações homográficas – a direção de visualização de câmera (eixo geométrico principal) de modo a trazer o ponto de desaparecimento vertical na infinidade e asseguram que a linha de horizonte é horizontal. Para visualizações de topo, o eixo geométrico principal é definido perpendicular ao fundo e um mapeamento polar é realizado para alcançar as mesmas propriedades. Note que em algumas configurações geométricas, essas transformações podem induzir diversas inclinações das visualizações.

C. Abordagem proposta: detecção de pessoas a partir da ocupação básica

[000192] Dadas as máscaras de ocupação básica G_i para todas as visualizações, explica-se agora como inferir a posição das pessoas que permanecem no fundo. A priori, sabe-se que (i) cada jogador induz um grupo denso na soma das máscaras de ocupação básica e (ii) o número de pessoas para detectar é igual a um valor conhecido K, por exemplo, K = 12 para basquetebol (jogadores + juízes).

[000193] Por essa razão, em cada localização básica x, considera-se a soma de todas as projeções – normalizadas pelo número de visualizações que realmente cobrem x – e procura-se pelos locais de maior intensidade nessa máscara de ocupação básica agregada (veja a figura 2 para um exemplo de máscara de ocupação básica agregada). Para localizar esses locais, considera-se primeiro uma abordagem ávida simplista que é equivalente a um procedimento de busca de igualação repetitivo. Em cada etapa o procedimento de busca de igualação maximiza o produto interno entre um núcleo gaussiano traduzido e a máscara de ocupação básica agregada. A posição do núcleo que induz o produto interno maior define a posição do jogador, Antes de executar a próxima repetição, a contribuição do núcleo gaussiano é subtraída da máscara agregada para produzir uma

máscara residual. O processo repete até que jogadores suficientes tenham sido encontrados.

Apêndice 2

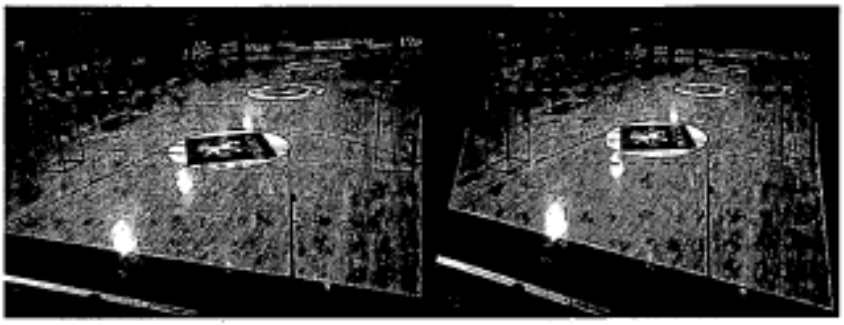


figura 3. A computação eficiente da máscara de ocupação básica: a visualização original (na esquerda) é mapeada para um plano através de uma combinação de homografias que são escolhidas de modo que (1) a verticalidade é preservada durante a projeção a partir da cena 3D para a visualização transformada e (2) a razão das alturas entre a cena 3D e a visualização projetada é preservada para objetos que estão situados na mesma linha na visualização transformada.

[000194] Essa abordagem é simples, mas sofre de muitas detecções falsas na interseção das projeções de silhuetas de jogadores distintos de diferentes visualizações. Isso é devido ao fato de que as oclusões induzem não linearidades[1] na definição da máscara de ocupação básica. Portanto, uma vez que algumas pessoas são conhecidas a estarem presentes no campo básico afetam a informação que pode ser retirada das máscaras de mudança binária em cada visualização. Em particular, se a linha vertical associada a uma posição x for obstruída por/ obstruir outro jogador cuja presença é muito parecida, essa visualização em particular não deve ser explorada para decidir se há um jogador em x ou não.

[000195] Por essa razão, propõe-se refinar a abordagem simplista conforme segue.

[000196] Para inicializar o processo, define-se $G_i^0(x)$ para ser a máscara de ocupação básica G_i associada à i-ésima visualização (veja a Seção III-B) e definir $w_i^0(x)$ para 1 quando x é coberto pela i-ésima visualização e para 0 caso contrário. Cada repetição é, então, executada em duas etapas. Na repetição n, a primeira etapa procura pela posição mais provável do n-ésimo jogador, sabendo a posição dos (n + 1) jogadores localizados nas repetições anteriores. A segunda etapa atualiza as máscaras de ocupação básica de todas as visualizações para remover a contribuição do jogador recentemente localizado.

[000197] Formalmente, a primeira etapa de repetição n agrega a máscara de ocupação básica de todas as visualizações e, então, procura pelo grupo mais denso nessa máscara. Portanto computa-se a máscara agregada G^n na repetição n como

$$G^n(\mathbf{x}) = \frac{\sum_{i=1}^{C} w_i^n(\mathbf{x}) G_i^n(\mathbf{x})}{\sum_{i=1}^{C} w_i^n(\mathbf{x})}, \quad (1)$$

e, então, define a posição mais provável x_n para n-ésimo jogador por

$$x_n = \underset{y}{argmax} < G^n(x), k(y) >, \quad (2)$$

onde k(y) denota um núcleo gaussiano centralizado em y e cujo suporte espacial corresponde à largura típica de um jogador.

[1] Em outras palavras, a máscara de ocupação básica de um grupo de jogadores não é igual à soma das máscaras de ocupação básica projetadas por cada jogador individual.

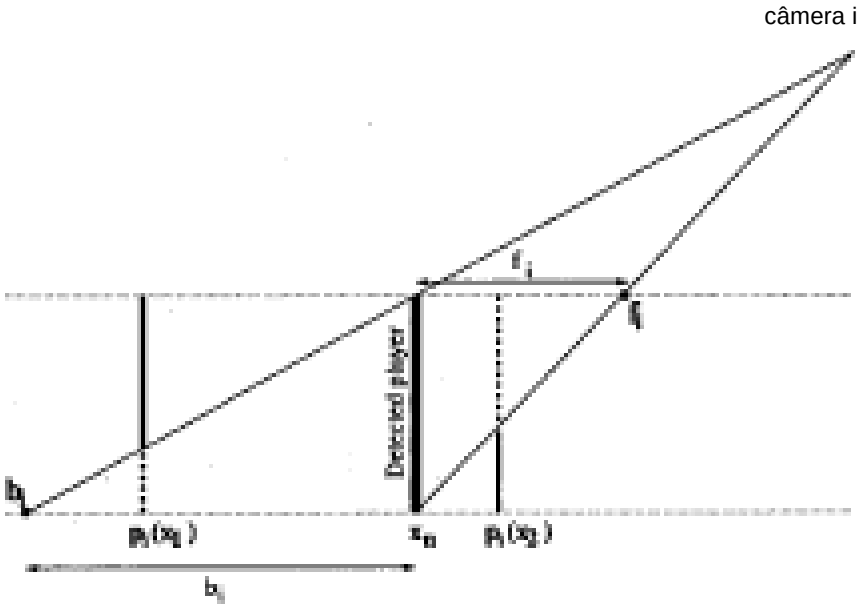


figura 4. O impacto das oclusões na atualização da máscara de ocupação básica associada com a câmera i. A parte tracejada da silhueta vertical que permanece em $p_i(x_i)$ e $p_2(x_2)$ é conhecida para ser marcada como primeiro plano desde que um jogador seja conhecido para permanecer em x_n. Portanto estas se tornam inúteis para inferir se um jogador está localizado em x_i e x_2, respectivamente.

[000198] Na segunda etapa, a máscara de ocupação básica de cada visualização é atualizada para considerar a presença do n-ésimo jogador. Na posição básica x, considera-se que o suporte típico de uma silhueta de jogador na visualização i é uma caixa retangular de largura W e altura H e observa-se que a parte da silhueta que obstrui ou é obstruída pelo jogador recentemente detectado não traz qualquer informação sobre a presença potencial de um jogador na posição x. Deixar $\alpha_i(x, x_n)$ denotar a fração da silhueta na posição básica x que se torna não informativa na visualização i como uma consequência da presença de um jogador em x_n. Para estimar essa razão, considera-se a geometria do problema. A figura 4 retrata um plano P_i que é ortogonal ao fundo, enquanto passa através da i-ésima câmera e da posição de jogador x_n. Em P_i, consideram-se dois pontos de interesse, nomeados

b_i e f_i, que correspondem ais pontos em que os raios, originados na i-ésima câmera e passando através da cabeça e pés do jogador, interceptam o plano básico e o plano paralelo ao fundo na altura H, respectivamente. Denota-se f_i (b_i) para ser a distancia entre f_i (b_i) e a linha vertical que suporta o jogador n em P_i. Também se considera $p_i(x)$ para denotar a projeção ortogonal de x em P_i e deixar $d_i(x)$ medir a distância entre x e P_i. Baseando-se nessas definições, a razão $\alpha_i(x, x_n)$ é estimada por

$$\alpha i(x, xn) = \begin{matrix} \delta - \min(\|p_i(x) - x_n\|, \delta))/\delta \\ .[1 - \min(d_i(x)/W, 1)] \end{matrix} \quad (3)$$

com δ sendo igual a f_i ou b_i, dependendo de se $p_i(x)$ estiver posicionado na frente ou atrás de x_n, com respeito à câmera. Em (3), os primeiro e segundo fatores refletem o desalinhamento de x e x_n em P_i e ortogonalmente a P_i, respectivamente.

[000199] Dado $\alpha_i(x, x_n)$, a máscara de ocupação básica e o peso de agregação da i-ésima câmera na posição x são atualizados conforme segue:

$$G_i^{n+1}(x) = max(G_i^n(x) - \propto_i (x, x_n) G_i^0(x_n), 0) \quad (4)$$

$$w_i^{n+1}(x) = max(w_i^n(x) - \propto_i (x, x_n) \quad (5)$$

[000200] Para a eficácia computacional melhorada, limitam-se as posições x investigadas na abordagem refinada para os 30 valores máximos locais que têm sido detectados pela abordagem simplista.

[000201] Para a totalidade, nota-se que o procedimento de atualização descrito acima omite a interferência potencial entre as oclusões causada pelos jogadores distintos na mesma visualização. Entretanto, a consequência dessa aproximação está longe de ser dramática, uma vez que esta termina dentro, sem afetar a informação que é realmente explorada. Levar essas interferências em consideração iria requerer retroprojetar as silhuetas de jogador em cada visualização, através disso tendendo em direção a uma abordagem de cima para baixo custosa de memória e computacionalmente tal como a

apresentada em [7].

[000202] Além disso, vale a pena mencionar que, em um contexto de cima para baixo, os autores em [1] ou em [7] propõem formulações que buscam simultaneamente pelas posições K que explicam melhor as observações de máscaras de primeiro plano múltiplas. Entretanto, considerando juntamente todas as posições aumenta-se a dimensionalidade do problema e impacta dramaticamente a carga computacional. Como os resultados experimentais mostram que o método proposto não sofre das fraquezas usuais dos algoritmos ávidos, tal como uma tendência a ser pego em valores mínimos locais, acredita-se que este se compara de modo muito favorável a qualquer formulação em conjunto do problema, tipicamente resolvido baseando-se em técnicas de otimização próximas repetitivas.

IV. Reconhecimento de dígito de jogadores

[000203] Esta seção considera o reconhecimento dos caracteres impressos nas camisas de esporte dos atletas. A abordagem proposta é retratada na figura 2. Para cada posição detectada no plano básico, uma caixa delimitadora conservativa de 0,8 m x 2 m é projetada em cada uma das visualizações. Cada caixa é, então, processada de acordo com uma abordagem que é similar ao método grosso para fino introduzido em [12]. Na etapa inicial, a imagem de caixa delimitadora é segmentada em regiões. As regiões candidatas de dígito são, então, filtradas com base em dígitos ‘0 a 9’ ou classes binárias e a identidade do jogador é definida pelo voto da maioria, baseando-se nos resultados obtidos em diferentes visualizações. A abordagem proposta difere-se da [12] no modo que cada uma dessas etapas é implantada.

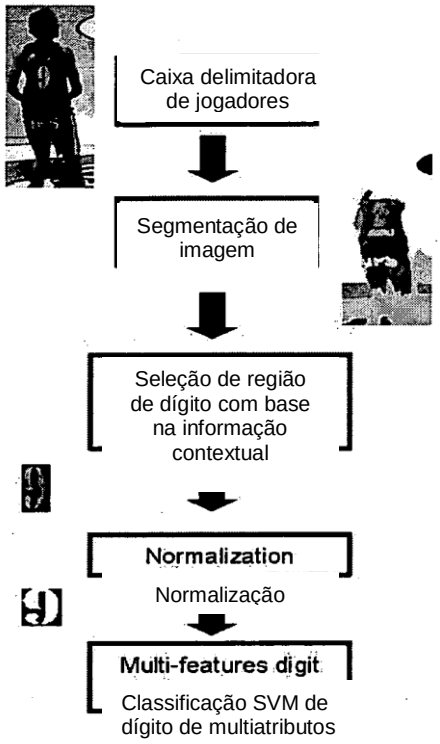


figura 5. Reconhecimento dos dígitos impressos nas camisas dos jogadores através da segmentação, seleção e classificação das regiões que são prováveis de representar dígitos.

[000204] A etapa de segmentação é baseada no algoritmo de desvio de meio [6], que é uma técnica de reconhecimento padrão que é particularmente bem adequada para delinear regiões mais densas em algum espaço de atributo estruturado arbitrariamente. Na segmentação de imagem de desvio de meio, a imagem é tipicamente representada como uma treliça de bidimensional de pixels L*u*v tridimensionais. O espaço da treliça é conhecido como o domínio espacial, enquanto que a informação de cor corresponde ao domínio de faixa. Os vetores de localização e de faixa são concatenados em um domino de faixa espacial em conjunto e um núcleo multivariado é definido como o produto de dois núcleos radialmente simétricos em cada domínio, que permite a definição independente dos parâmetros de largura de banda h_8 e h_r para os domínios de faixa e espacial, respectivamente [6]. Os

valores máximos locais da densidade de domínio em conjunto são, então, computados e os modos que estão mais próximos que h_8 no domínio espacial e h_r no domínio de faixa são cortados em modos significantes. Cada pixel é, então, associado com um modo significante da densidade de domínio em conjunto localizada em seus arredores. Eventualmente, as regiões espaciais que contêm menos que M pixels são eliminadas. No caso, como há um forte contraste entre o dígito e a camisa, pode-se arcar com um valor alto para h_r, que é definido como 8 nas simulações. O parâmetro h_8 troca o tempo de execução da segmentação e filtragem subsequente e estágios de classificação. De fato, um valor h_r pequeno define um núcleo menor, que torna a segmentação mais rápida, mas também resulta em um número maior de regiões para processar em estágios subsequentes. Nas simulações, h_r foi definido como 4, enquanto que M foi fixado como 20.

[000205] Para filtrar as regiões que obviamente não correspondem aos dígitos, confia-se nas seguintes observações:

• As regiões de dígitos válidas nunca tocam a borda da caixa delimitadora (conservativa);

• As regiões de dígitos válidas são cercadas por uma única região colorida homogeneamente. Em prática, o algoritmo seleciona as regiões para quais arredores dos 4 pontos extremos (em cima/embaixo, esquerda/direita) da região pertencem a mesma região;

• A altura e a largura das regiões válidas variam entre dois valores que são definidos relativamente para o tamanho da caixa delimitadora. Como o tamanho da delimitação é definido de acordo com as medidas do mundo real, o critério de tamanho adapta implicitamente a faixa de valores de altura e largura para os efeitos perspectivos que resultam da distancia entre a câmera e o objeto detectado.

[000206] Para totalidade, vale a pena mencionar que algumas fontes particulares dividem alguns dígitos em duas regiões distintas. Por esta

razão, regiões de dígito candidato são compostas de ou um único ou um par de regiões que atendem os critérios acima.

[000207] As (o par de) regiões que foram selecionadas como elegíveis para processamento subsequente são então normalizadas e classificadas. A normalização implica alinhamento horizontal do maior eixo geométrico principal, conforme derivado através de computação de momentos de inercia e conversão em uma máscara binária 24 x 24. A classificação é baseada na estratégia SVM de múltiplas classes "um contra um" [8], conforme recomendado e implantado pela biblioteca LIBSVM [3]. Um SVM de duas classes é treinado para cada par de classes e uma estratégia de maioria de votos é explorada para inferir a classe (dígito 0 a 9 ou classe binária) a partir do conjunto de decisões de classificação binária. Na prática, para alimentar o classificador, cada amostra região é descrita por um vetor de característica de 30 dimensões, ou seja:

- 1 valor para definir o número de orifícios na região;
- 3 valores que correspondem aos segundos momentos de ordem m02, m20 e m22;
- 2 valores para definir o centro de massa da região;
- 2 x 12 valores para definis o histograma da região ao longo do eixo geométrico vertical e horizontal.

[000208] Os números com dois dígitos são reconstruídos com base na detecção de dois dígitos adjacentes. Para executar as presentes simulações, treinou-se o classificados SVM com base em mais de 200 amostra manualmente segmentadas de cada dígito e em 1200 amostras da classe binária. As amostras de classe binária correspondem a regiões de não dígito que são automaticamente segmentadas em uma das visualizações e cujo tamanho é consistente com o de um dígito.

V. Rastreamento de jogadores detectados

[000209] Para rastrear os jogadores detectados, implantou-se um

algoritmo rudimentar embora efetivo. A propagação dos rastreamentos é atualmente feita por um horizonte de um quadro, com base no algoritmo de atribuição geral de Munkres [11]. O chaveamento é usado para prevenir combinações indesejadas e um módulo de análise de alto nível é usado para ligar juntos rastreamentos parciais com uso de estimação de cor da blusa. No futuro, técnicas de combinação de gráfico devem ser usadas para avaliar hipóteses de combinação de horizonte mais longo. Análise de alto nível mais sofisticada deve ser também implantada, por exemplo, para explorar as informações de reconhecimento de jogador disponíveis ou para duplicar os rastreamentos parciais que seguem dois jogadores que estão muito próximos um ao outro.

VI. Validação experimental

A. Detecção de jogador e rastreamento

[000210] Para avaliar o algoritmo de detecção de jogador, mediram-se as taxas de detecção perdida média e de detecção falsa por 180 instantes diferentes e regularmente espaçados no intervalo de 18:47:00 a 18:50:00, que corresponde a um segmento temporal para o qual uma verdade de fundo manual está disponível. Essas informações de verdade de fundo consistem nas posições dos jogadores e árbitros no sistema de referência de coordenada do corte. Considera-se que dois objetos não podem ser combinados se a distância medida no fundo for maior que 30 cm. A figura 6 apresenta diversas curvas de ROC, cada curva sendo obtida por variação do limiar de detecção para um dado método de detecção. Três métodos são comparados e para cada um deles se avalia o algoritmo proposto para mitigar detecções falsas. Como um primeiro método de referência, considerou-se a abordagem seguida por [9], [10], que projeta as máscaras de primeiro plano de todas as visualizações somente no plano básico. O desempenho pobre desta última abordagem é principalmente devido às sombras dos

jogadores e à pequena contribuição dos pés dos jogadores para as máscaras de primeiro plano. Para validar esta interpretação, no segundo método, projetaram-se as máscaras de primeiro plano em um único plano localizado um metro acima do plano básico. Fazendo isto, a influência das sombras é drasticamente atenuada, enquanto a contribuição principal agora se origina das partes centrais do corpo, que são normalmente bem representadas nas máscaras de primeiro plano. Observa-se que melhoras significantes comparadas a [9] e [10]. O terceiro e último método de detecção apresentado na figura 6 é o método proposto. Observa-se que o benefício obtido da integração de ocupação de fundo é impressionante. A melhora trazida pelo detector de alarme falso é também bem óbvia. Adicionalmente, o xis na figura 6 representa um ponto de operação alcançado após rastreamento rudimentar de posições detectadas. Observa-se que o rastreamento em consistência temporal de conta pode melhorar ainda mais os resultados de detecção.

[000211] Na configuração APIDIS, todas as áreas do corte de cesta não são cobertas pelo mesmo número de câmeras. A figura 7 mostra a influência da cobertura de câmera nas taxas de detecções falsas ou perdidas. Isso também mostra que nas áreas com alta cobertura, a maioria das detecções perdidas de devem aos jogadores estarem muito próximos um do outro.

B. Reconhecimento do jogador

[000212] Para validar o encadeamento de reconhecimento de jogador, selecionou-se 190 blocos de ligação de câmeras de visualização lateral. Em cada bloco de ligação selecionado, o dígito foi visível e poderia ser lido por um observador humano a despeito das distorções de aparência possivelmente significantes.

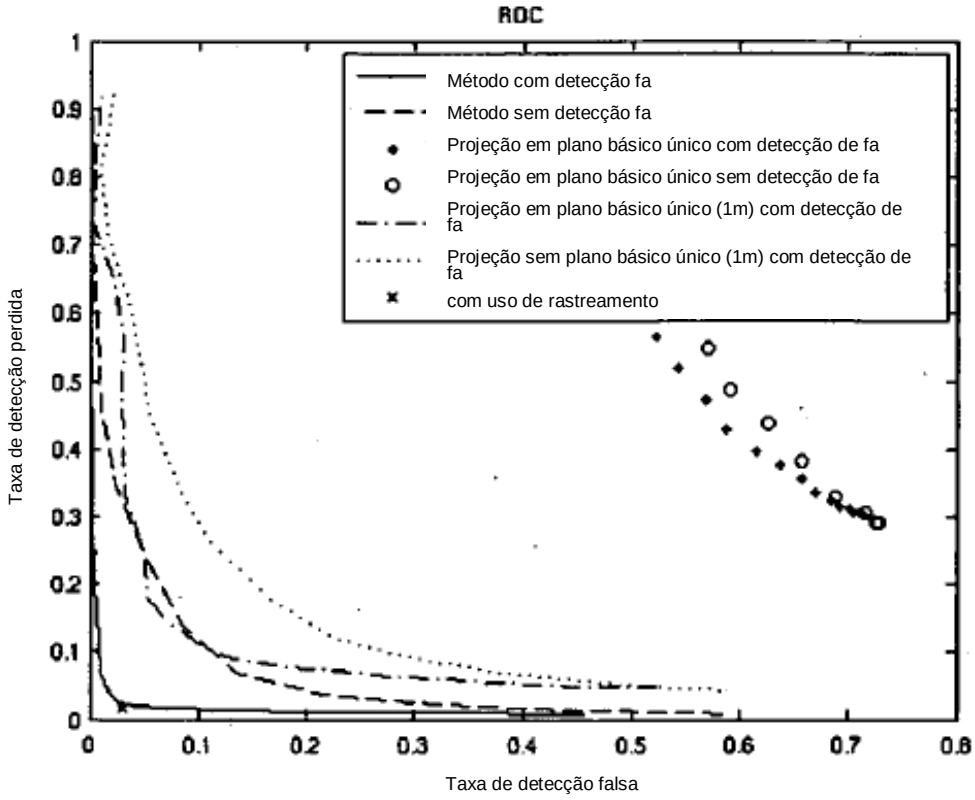


Figura 6. Análise de ROC de desempenho de detecção de jogador

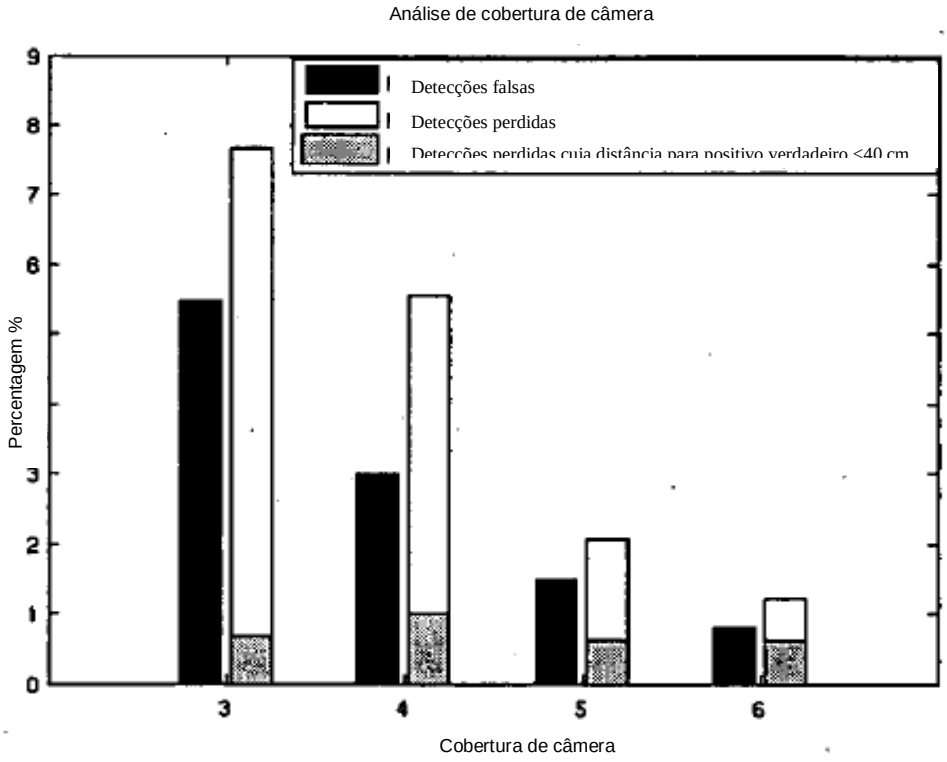


Figura 7. Cobertura da câmera wrt relacionada ao

desempenho de detecção de jogador

[000213] A Tabela I sumariza os resultados de reconhecimento. A taxa de reconhecimento é acima de 73%. De forma mais interessante, observa-se que quando o dígito não foi reconhecido, isto foi em sua maioria atribuído à classe binária, ou não passou a análise contextual devido a erro de segmentação. Além disso, os remanescentes 4% de positivo falso não incluem qualquer erro de combinação real entre dois dígitos. Na verdade, 75% de positivos falsos foram devidos à perda de um dígito em um número de dois dígitos. Em outros casos, dois dígitos foram reconhecidos, o correto e um detectado falso.

[000214] Além disso, uma análise mais detalhada revelou que a maioria dos jogadores não reconhecidos estavam no lado oposto do campo, comparados à visualização de câmera a partir da qual o bloco de ligação foi extraído. Neste caso, a altura do dígito diminui a menos de 15 pixels, o que explica o desempenho de reconhecimento pobre, abaixo de 50%. Em contraste, uma câmera localizada no mesmo lado do campo que o jogador alcança próximo a 90% de taxa de reco reconhecimento correto.

[000215] Com base nessas observações, está-se sensatamente consciente que de o desempenho de reconhecimento do sistema será suficientemente bom para atribuir um marcador correto a segmentos curtos de trajetórias de jogador, assim fornecendo uma ferramenta valiosa tanto para elevar ambiguidades de rastreamento ou para favorecer um jogador preferencial durante a produção sumária do vídeo.

Reconhecimento	73%
Erro de segmentação	11%
Negativo falso	12%
Positivo falso	4%

TABELA I

Desempenho de reconhecimento de jogador

VII. Conclusão

[000216] Apresentaram-se algoritmos de processamento de vídeo para definir a posição e identidade de atletas que jogam em um campo de esporte cercado por um conjunto de câmeras livremente sincronizadas. As transmissões de detecção na definição de um mapa de ocupação de fundo, enquanto o reconhecimento de jogador constrói em pré-filtragem de regiões segmentadas e em classificação de SVM de múltiplas classes. Os experimentos no conjunto de dados de vida real de APIDIS demostra a relevância das abordagens propostas.

REFERÊNCIAS

[1] A. Alahi, Y. Boursier, L. Jacques, e P. Vandergheynst, "A sparsity constrained inverse problem to locate people in a network of cameras," em *Proceedings of the 16th International Conference on Digital Signal Processing (DSP),* Santorini, Grécia, Julho de 2006.

[2] J. Berclaz, F. Fleuret, e P. Fua, "Principled detection-by-classification from multiple views," em *Proceedings of the International Conference on Computer Vision Theory and Application (VISAPP)*, volume 2, Funchal, Madeira, Portugal, Janeiro de 2008, páginas 375 a 382.

[3] C-C. Chang e C-J. Lin, "LIBSVM: A library for support vector machines," em *http://www.csie.ntu.edu.tw/ cilin/papersflibsvm.pdf*.

[4] F. Chen e C. De Vleeschouwer, "A resource allocation framework for summarizing team sport videos," em *IEEE International Conference on Image Processing*, Cairo, Egito, Novembro de 2009.

[5] "Autonomous production of basket-ball videos from multi- sensored data with personalized viewpoints," em *Proceedings of the 10th International Workshop on Image Analysis for Multimedia Interactive Services*, Londres, Reino Unido, Maio de 2009.

[6] D. Comaniciu e P. Meer, "Mean shift: a robust approach

toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 24, n°. 5, páginas 603 a 619, Maio de 2002.

[7] F. Fleuret, J. Berclaz, R. Lengagne, e P. Fua, "Multi-camera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 30, n°. 2, páginas 267 a 282, Fevereiro de 2008.

[8] C-W. Hsu e C-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, volume 13, n°. 2, páginas. 415 a 425, Março de 2002.

[9] S. Khan e M. Shah, "A multiview approach to tracing people in crowded scenes using a planar homography constraint," em *Proceedings of the 9th European Conference on Computer Vision (ECCV*), volume 4, Graz, Áustria, Maio de 2006, páginas 133 a 146.

[10] A. Lanza, L. Di Stefano, J. Berclaz, F. Fleuret, e P. Fua, "Robust multiview change detection," em *British Machine Vision Conference (BMVC)*, Warwick, Reino Unido, Setembro de 2007.

[11] J. Munkres, "Algorithms for the assignment and transportation problems," em *SIAM J. Control*, volume 5, 1957, páginas 32 a 38.

[12] Q. Ye, Q. Huang, S. Jiang, Y. Liu, e W. Gao, "Jersey number detection in sports video for athlete identification," em *Proceedings of the SPIE, Visual Communications and Image Processing*, volume 5960, Beijing, China, Julho de 2005, páginas 1599 a 1606.

REIVINDICAÇÕES

1. Sistema baseado em computador para selecionar pontos de visualização ótimos, como uma função do tempo, para adequar a resolução de tela e outras preferências de usuário, e para suavizar essas sequências para contar história de modo contínuo e gracioso compreendendo um motor e memória de processamento para produção autônoma de um vídeo editado a partir de múltiplos fluxos de vídeo capturados por uma pluralidade de câmeras distribuídas ao redor de uma cena de interesse, o sistema **caracterizado pelo fato de** que compreende:

um detector para detectar objetos em imagens dos fluxos de vídeo;

um primeiro meio para selecionar um ou mais pontos de visualização de câmera obtidos ou pelo corte da imagem capturada por uma câmera fixa ou pela seleção de parâmetros de movimento-inclinação-zoom ("pan-tilt-zoom") de uma câmera virtual ou motorizada, a dita câmera virtual aproximando uma imagem em uma posição arbitrária com um campo de visão arbitrário pela exploração de uma rede distribuída das ditas câmeras, o ponto de visualização da câmera sendo selecionado para incluir os objetos de interesse e com base no processamento de união das posições dos múltiplos objetos que foram detectados, em que a seleção ocorre de modo que balanceia as métricas de integridade e proximidade como uma função das preferências de usuário individuais, em que a integridade conta o número de objetos de interesse que são incluídos e visíveis dentro do ponto de visualização exibido, e a proximidade mede o número de pixels que estão disponíveis para descrever os objetos de interesse;

um segundo meio para selecionar parâmetros de renderização que maximizam e suavizam as métricas de proximidade e integridade pelos segmentos de concatenização nos fluxos de vídeo

proporcionados por uma ou mais câmeras individuais, em que a construção é feita de um modo que balanceia as métricas de proximidade e integridade ao longo do tempo, enquanto suaviza a sequência dos ditos parâmetros de corte e/ou de movimento-inclinação-zoom associados aos segmentos concatenados em que o processo de suavização é implementado com base no mecanismo de filtragem temporal passa baixa linear ou não linear.

2. Sistema, de acordo com a reivindicação 1, **caracterizado pelo fato de** que ainda compreende um quinto meio para selecionar o ponto de visualização ótimo em cada visualização de câmera, em um dado instante, em que o quinto meio para selecionar o ponto de visualização ótimo é adaptado, para um ponto de visualização v_k na k-ésima visualização de câmera é definido pelo tamanho S_k e o centro $\mathbf{c}_k$ da janela que é cortada na k-ésima visualização para exibição real e é selecionado para incluir os objetos de interesse e para fornecer uma alta resolução, é adaptado para selecionar uma descrição dos objetos e um ponto de visualização ótimo v_k* para maximizar uma soma ponderada de interesses de objeto como segue

$$\mathbf{v}_k^* = \arg\max_{\{S_k, \mathbf{c}_k\}} \sum_{n=1}^{N} I_n \cdot \alpha(S_k, \mathbf{u}) \cdot m\left(\frac{\left\|\mathbf{x}_{n,k} - \mathbf{c}_k\right\|}{S_k}\right)$$

em que I_n indica o nível de interesse atribuído ao n-ésimo objeto detectado na cena, $x_{n,k}$ denota a posição do n-ésimo objeto na visualização de câmera k,

a função m(...) modula os pesos do n-ésimo objeto de acordo com sua distância para o centro da janela do ponto de visualização, em comparação com o tamanho desta janela, a função sendo positiva no intervalo [0,0.5] e negativo ou zero mais além,

o vetor *u* reflete as preferências do usuário, em particular, o componente u_{res} define a resolução do fluxo de saída que é geralmente limitado pela largura de banda de transmissão ou pela resolução do

dispositivo de usuário final, e

a função α (.) reflete a penalidade induzida pelo fato de que os sinais nativos capturados pela k-ésima câmera tem que ser subamostrados uma vez que o tamanho do ponto de visualização torna-se maior do que a máxima resolução u_{res} permitida pelo usuário.

3. Sistema, de acordo com a reivindicação 2, **caracterizado pelo fato de** que α(....) diminui com Sk e a função α(....) é igual a 1 quando $S_k<u_{res}$, e diminui depois disso, em que α (....) é opcionalmente definido por:

$$\alpha(S,\mathbf{u})=\left[\min\left(\frac{u_{res}}{S},1\right)\right]^{u_{close}},$$

em que o expoente u_{close} é maior do que 1, e aumenta à medida que o usuário prefere a renderização de resolução completa de área de aumento de zoom, em comparação com pontos de visualização grandes mas subamostrados.

4. Sistema, de acordo com qualquer uma das reivindicações 1 a 3, **caracterizado pelo fato de** que ainda compreende um quarto meio para selecionar a câmera em um dado instante que torna a maioria dos objetos de interesse visível, e está próximo da ação, em que um índice de câmera ótimo k* é selecionado de acordo com uma equação que é similar ou equivalente a:

$$k^{*}=\arg\max_{\{k\}}\sum_{n=1}^{N} I_n\cdot o_k\left(\mathbf{x}_n\middle|\overline{\mathbf{x}}\right)\cdot h_k\left(\mathbf{x}_n\right)\cdot \beta_k\left(S_k{*},\mathbf{u}\right)$$

em que I_n indica o nível de interesse atribuído ao n-ésimo objeto detectado na cena,

x_n denota a posição do n-ésimo objeto no espaço 3D;

o_k ($x_{n|}$ x) mede a relação oclusão do n-ésimo objeto em câmera de visualização k, sabendo a posição de todos os outros objetos, a relação de oclusão de um objeto sendo definida para ser a fração de pixels do objeto que está escondida por outros objetos quando

projetado sobre o sensor da câmera;

a altura $h_k(x_n)$ é definida como sendo a altura em pixels da projeção em visualização k de uma altura de referência de um objeto de referência localizado em x_n, o valor de $h_k(x_n)$ sendo diretamente computado com base na calibração da câmara ou podendo ser estimado com base na altura do objeto detectado em visualização k quando a calibração não estiver disponível, e

a função $\beta_k(.)$ reflete o impacto das preferências do usuário em termos de visualização da câmera e resolução de tela.

5. Sistema, de acordo com a reivindicação 4, **caracterizado pelo fato de** que $\beta_k(.)$ é definido como

$$\beta_k(S, \mathbf{u}) = u_k \cdot \alpha(S, \mathbf{u}),$$

em que u_k representa o peso atribuído à k-ésima câmera e um $\alpha(S, u)$ é definido como na reivindicação 3.

6. Meio de armazenamento de sinal legível por máquina não transitória, **caracterizado pelo fato de** que armazena um produto de programa de computador que compreende segmentos de código os quais, quando executados em um motor de processamento, implementam o sistema conforme definido em qualquer uma das reivindicações 1 a 5.

Figura 1: Fluxo de trabalho hierárquico em produção personalizada

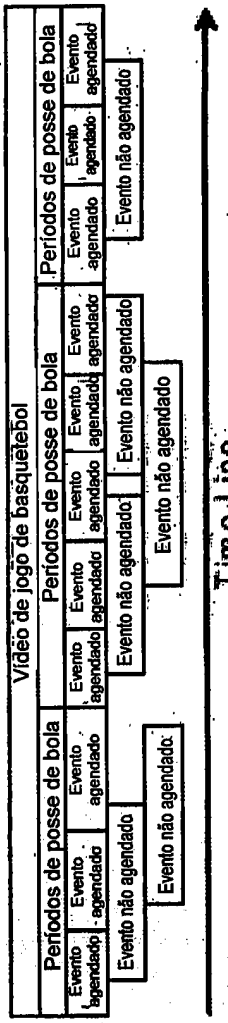
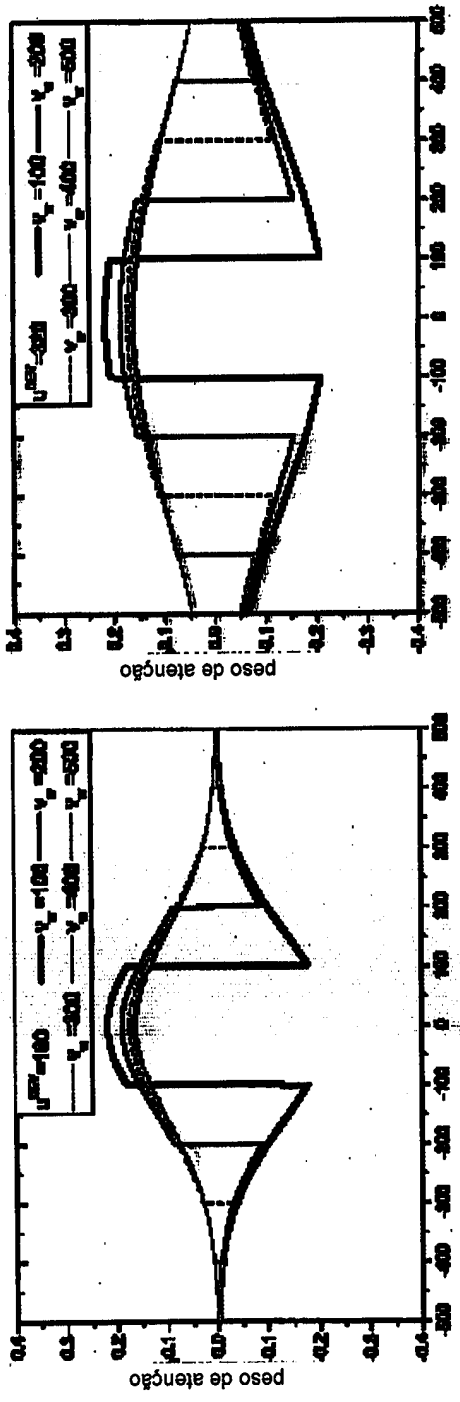


Figura 2: Estrutura hierárquica de vídeos de basquetebol.



(a) $u^{\mathrm{DEV}} = 160$ (b) $u^{\mathrm{DEV}} = 320$

Figura 3: Função ponderal de atenção proposta no presente trabalho. Compensa-se entre proximidade e totalidade por controle da agudeza de aba e pico da função de ponderação.

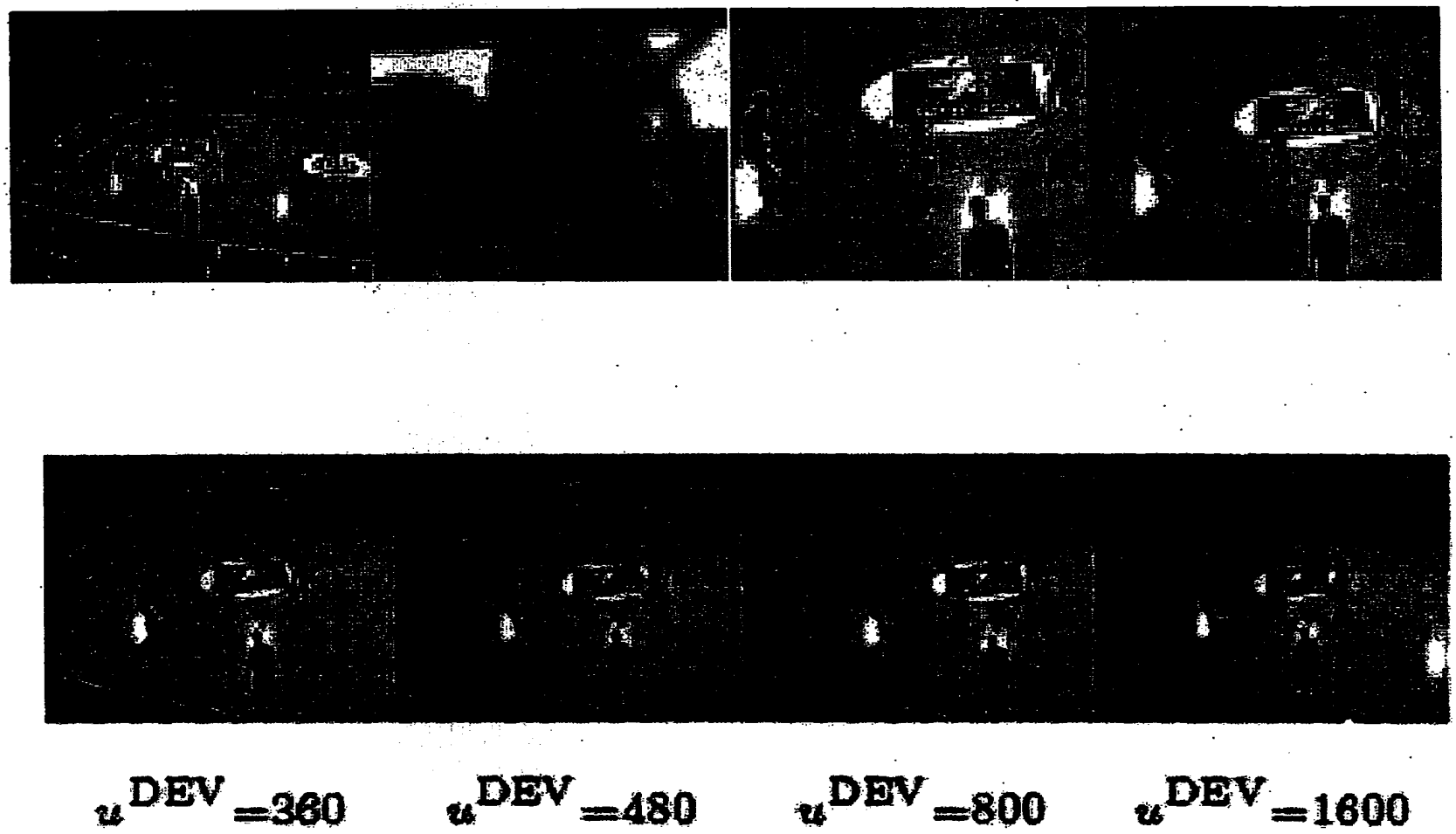


Figura 4: Comportamento de seleção de ponto de vista sob diferentes tamanhos de exibição

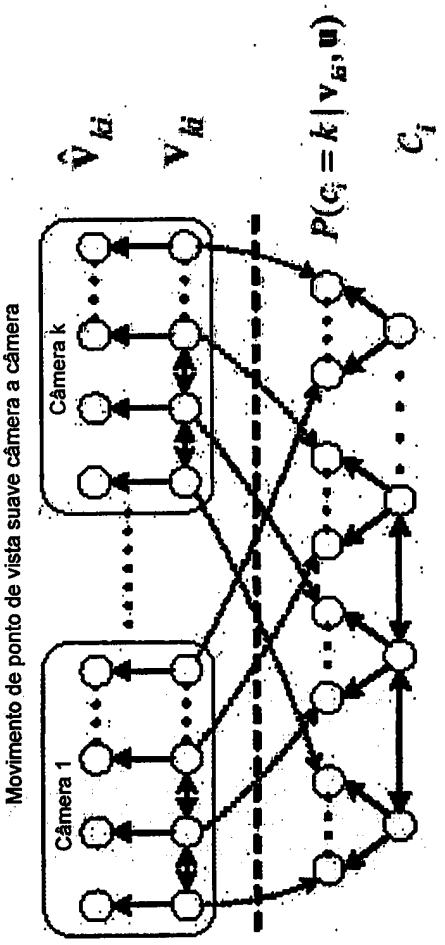


Figura 5: Modelo de gráfico para estimativa de duas etapas de movimento de ponto de vista suave

Câmera 1: câmeras de ação (corte esquerdo)
Câmera 2: câmera herói (corte esquerdo)
Câmera 3: câmera olho de peixe (foco de corte direito)
Câmera 4: câmera herói
Câmera 5: câmera olho de peixe (foco de corte esquerdo)
Câmera 6: câmeras de ação (corte direito)
Câmera 7: câmera herói (corte esquerdo)
Figura 7: Visualizações de amostra reunidas por câmeras diferentes

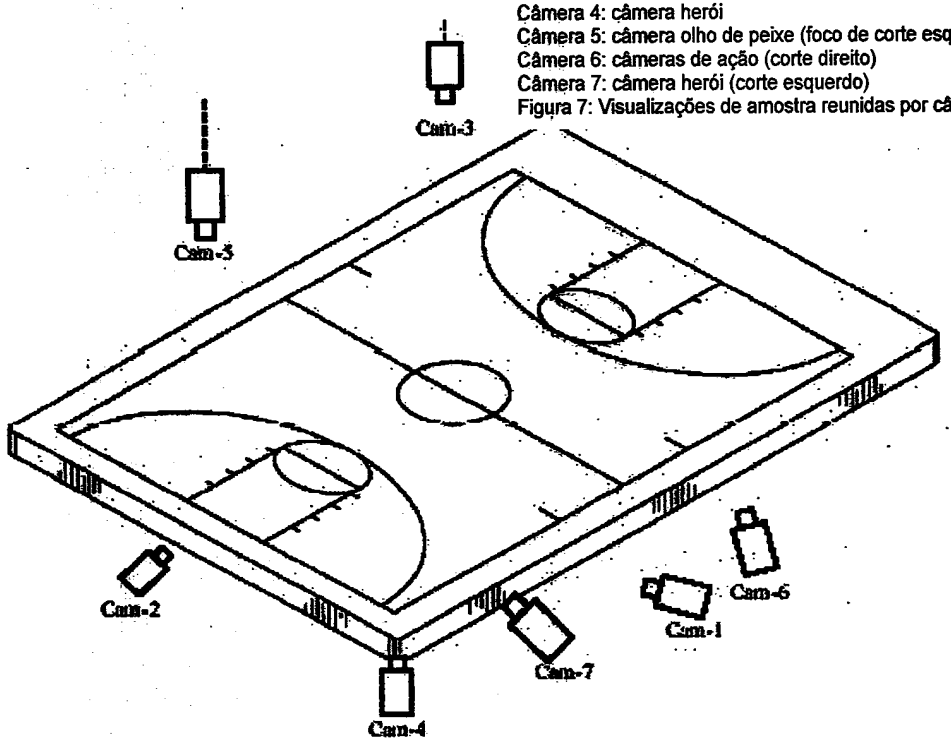


Figura 6: Planos de câmera em reunião de dados de vídeo experimental.

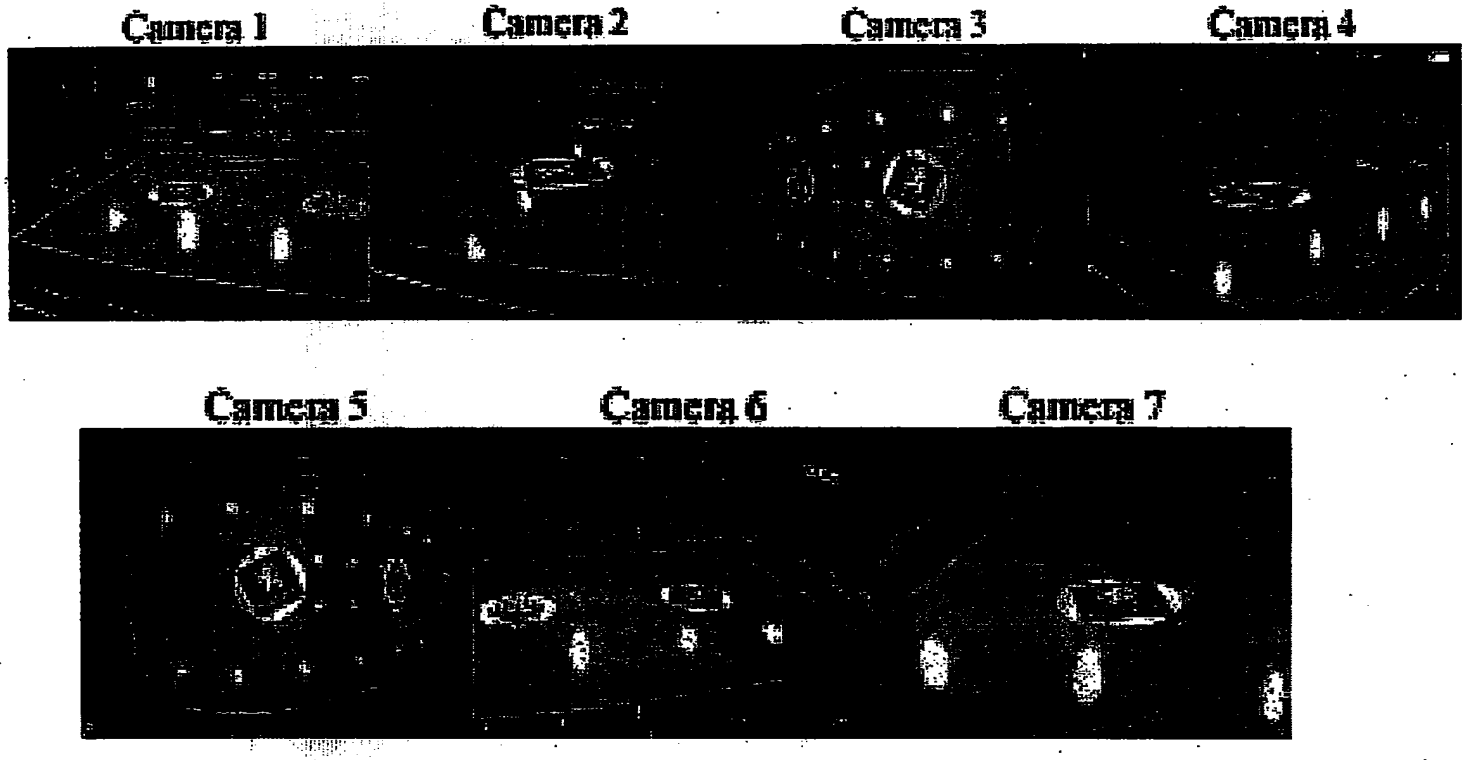


Figura 7: Visualizações de amostra reunidas por câmeras diferentes

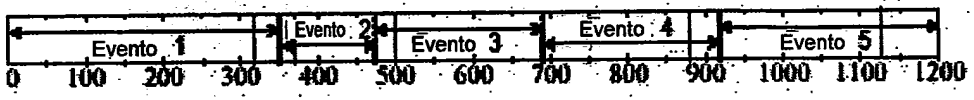


Figura 8: Um vídeo clip curto com 1.200 quadros é usado para demonstrar o sistema. Há cinco eventos agendados dentro deste clip. Para cada evento, o momento mais iluminado é marcado em uma barra sólida vermelha.

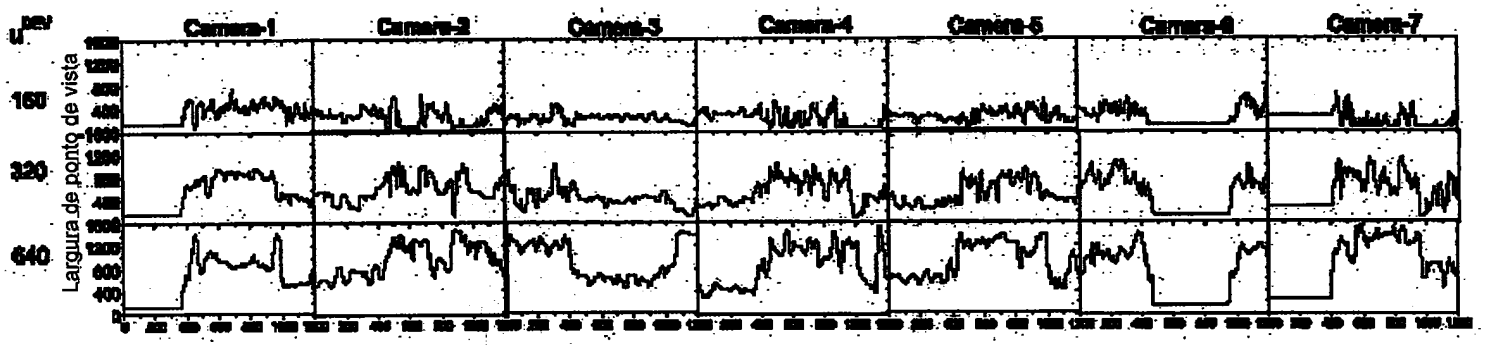


Figura 9: As sequências de ponto de vista câmera a câmera automaticamente selecionadas sob três diferentes resoluções de exibição. A suavização de ponto de vista fraca foi aplicada e a força de suavização é definida em

$$\sigma_2/\sigma_1 = 4.$$

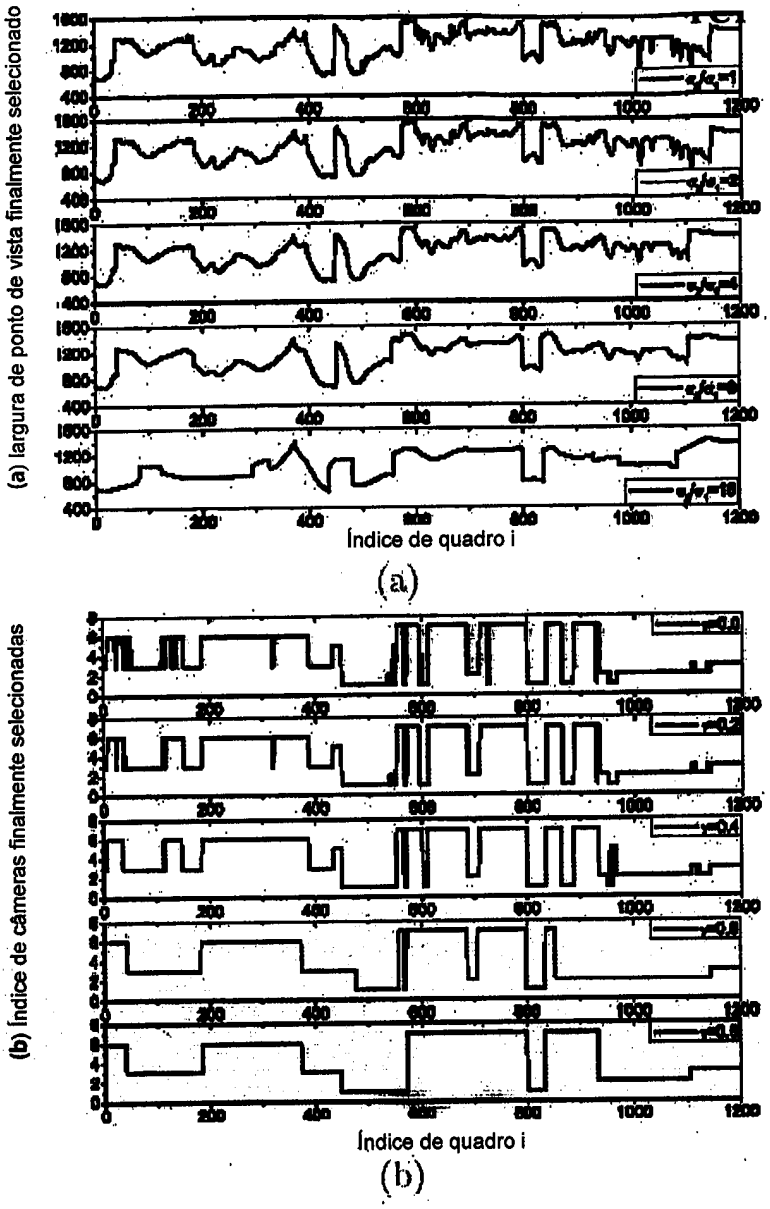


Figura 10: Comportamento de sequência de câmera/ponto de vista otimizado diferentes forças de suavização. Resolução de exibição é definida em

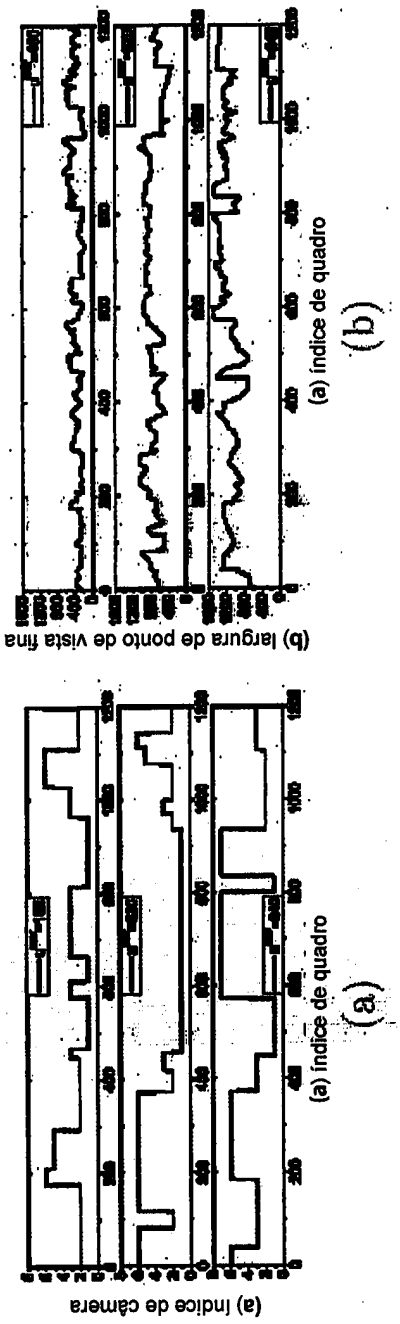


Figura 11: Comparação de sequências de câmera e ponto de vista geradas sob três diferentes resoluções de exibição 160, 320 e 640. A força de suavização de ponto de vista é σ2/σ1 = 4 e a força de suavização de câmera γ é definida como 0,8.

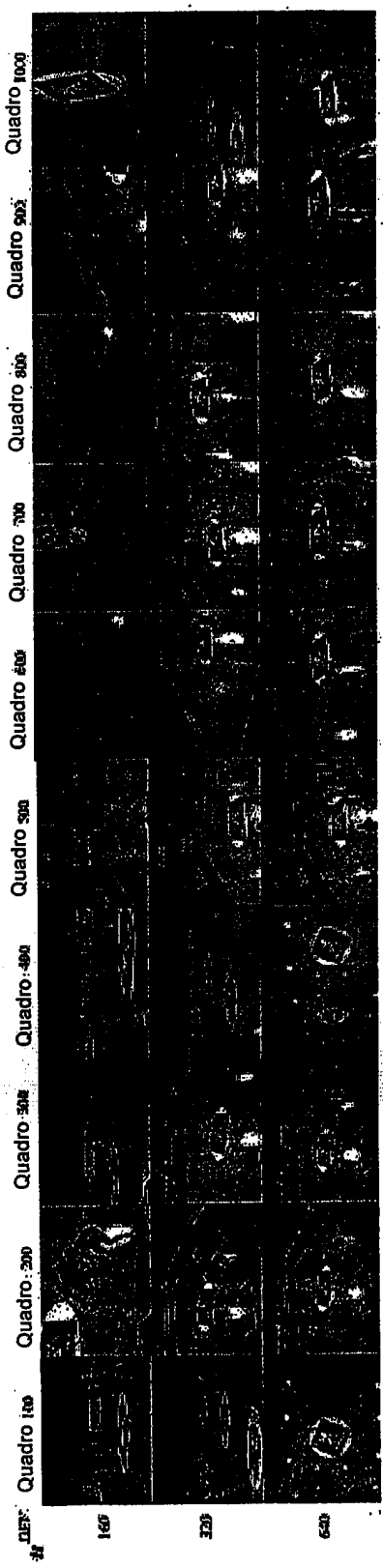


Figura 12: Miniaturas de quadros em sequências geradas sob três diferentes resoluções 160, 320 e 640. Força de suavização de ponto de vista é σ2/σ1 = 4 e a força de suavização de câmera γ é definida como 0,8.

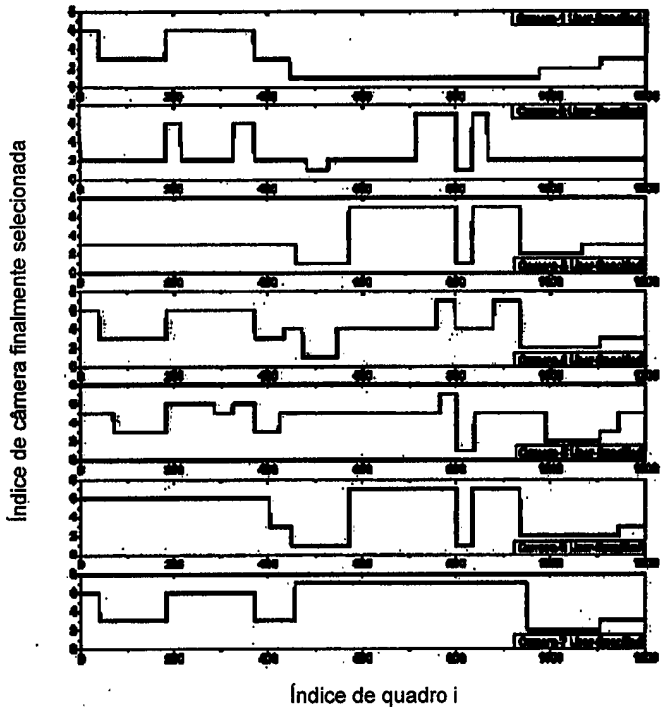


Figura 13: Comparação de sequências de câmera e ponto de vista geradas sob diferentes preferências do usuário para uDEV= 640. A força de suavização de ponto de vista é σ2/σ1 = 4 e a força de suavização de câmera γ é definida como 0,8.

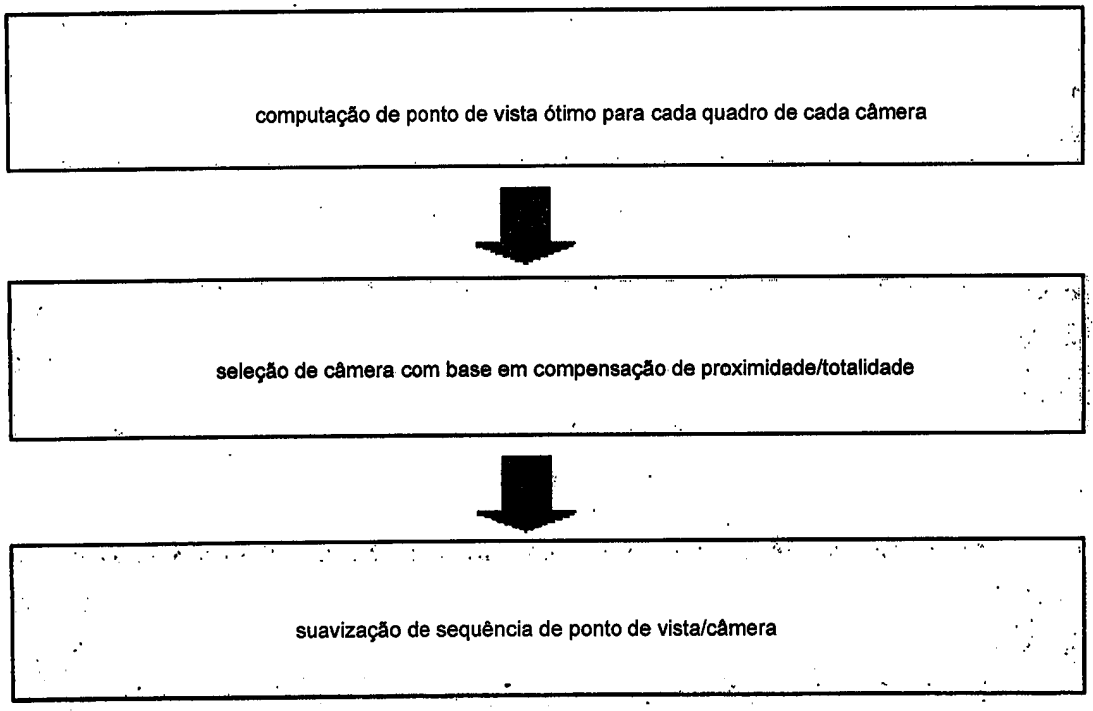


Figura 14: A solução proposta é implantada como um processo de 3 etapas. Vale notar que a suavização de ponto de vista poderia também ser implantada dentro da visualização de câmera correspondente, antes da seleção de câmera real.

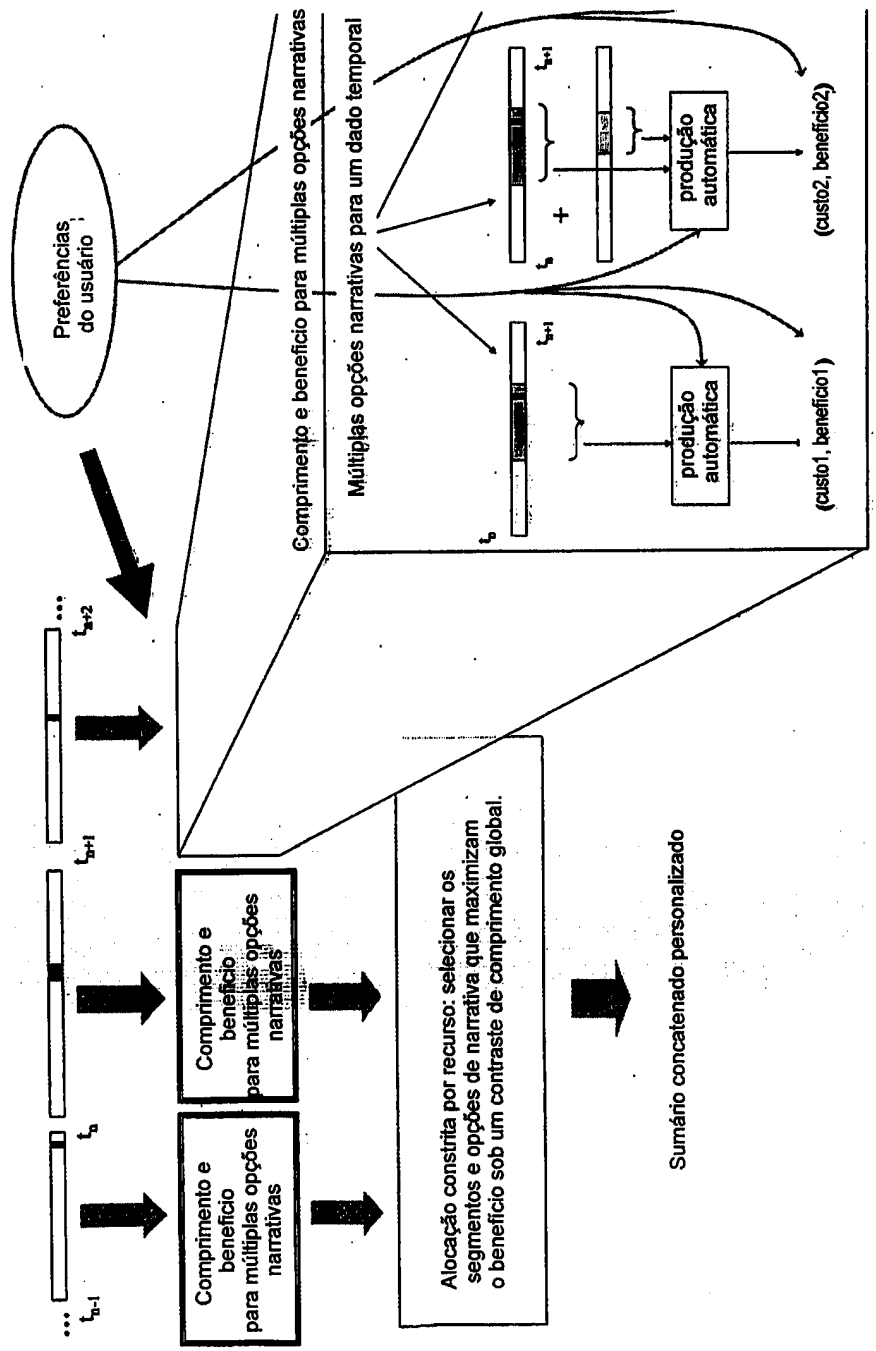

Preferências do usuário
Comprimento e benefício para múltiplas opções narrativas
Comprimento e benefício para múltiplas opções narrativas
Alocação constrita por recurso: selecionar os segmentos e opções de narrativa que maximizam o benefício sob um contraste de comprimento global.
Sumário concatenado personalizado
Comprimento e benefício para múltiplas opções narrativas
Múltiplas opções narrativas para um dado temporal
produção automática
produção automática
(custo1, benefício1)
(custo2, benefício2)

FIG. 15

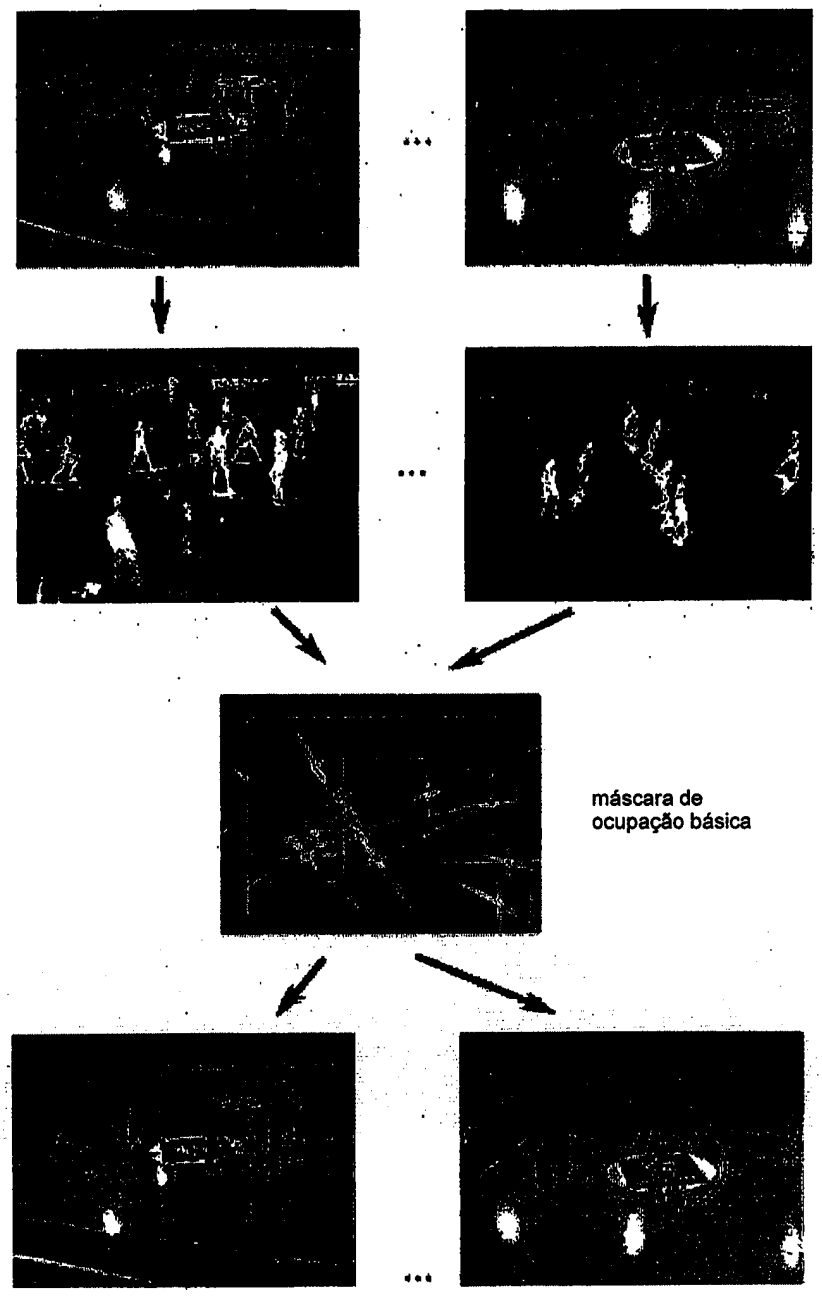


Figura 16: Máscaras de primeiro plano são mescladas para criar uma máscara de ocupação básica