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

(11) BR 102015005929-9 B1

(22) Data do Depósito: 17/03/2015

(45) Data de Concessão: 30/11/2021

(54) Título: SISTEMA E MÉTODO PARA COMPOSIÇÃO E COMPRESSÃO DE VÍDEO COM BASE EM CONTEXTO A PARTIR DE OBJETOS COM RESOLUÇÃO ESPACIAL NORMALIZADA

(51) Int.Cl.: G06F 3/0484; G06T 1/00.

(52) CPC: G06F 3/04847; G06F 3/0484; G06T 1/0007.

(73) Titular(es): SAMSUNG ELETRÔNICA DA AMAZÔNIA LTDA..

(72) Inventor(es): FERNANDA A. ANDALÓ; OTÁVIO A. B. PENATTI; VANESSA TESTONI; FERNANDO KOCH.

(57) Resumo: SISTEMA E MÉTODO PARA COMPOSIÇÃO E COMPRESSÃO DE VÍDEO COM BASE EM CONTEXTO A PARTIR DE OBJETOS COM RESOLUÇÃO ESPACIAL NORMALIZADA. A presente invenção refere-se a um sistema e método para gerar imagens e vídeos com eficiência como uma disposição de objetos de interesse (por exemplo, rostos e mãos, placas, etc.), em uma resolução desejada, para executar tarefas de visão, tais como reconhecimento de face, análise de expressão facial, detecção de gestos manuais, entre outras. A composição de tais imagens e vídeos leva em consideração a similaridade de objetos na mesma categoria para codificá-los de forma mais eficaz, fornecendo uma economia em termos de tempo de transmissão e armazenamento. Menos tempo de transmissão representa vantagens para tal sistema em termos de eficiência, enquanto menos armazenamento significa menor custo para o armazenamento de dados.

Relatório Descritivo da Patente de Invenção para: **“SISTEMA E MÉTODO PARA COMPOSIÇÃO E COMPRESSÃO DE VÍDEO COM BASE EM CONTEXTO A PARTIR DE OBJETOS COM RESOLUÇÃO ESPACIAL NORMALIZADA”.**

Campo da Invenção

[0001] A presente invenção refere-se a um método e sistema para gerar imagens e vídeos comprimidos que contenham objetos de interesse, inicialmente em diferentes resoluções, utilizando uma resolução espacial normalizada. O método da presente invenção pode gerar imagens e vídeos eficientes utilizando-se de uma disposição de objetos de interesse (por exemplo, faces e mãos, placas de carro, etc.), com uma resolução desejada, para executar tarefas de visão computacional, tais como reconhecimento de face, análise de expressão facial, detecção de gestos manuais, reconhecimento óptico de caracteres (OCR) de placas de carro, entre outras. A composição de tais imagens e vídeos leva em consideração a similaridade entre os objetos de uma mesma categoria para codificá-los de forma mais eficaz, fornecendo uma economia em termos de tempo de transmissão e armazenamento. Menos tempo de transmissão representa vantagens para tal sistema em termos de eficiência, enquanto menos armazenamento significa menor custo para o armazenamento de dados.

Antecedentes da Invenção

[0002] Sistemas baseados em visão computacional estão se tornando mais populares hoje em dia, principalmente por causa do aumento do poder dos dispositivos e dos novos recursos para o armazenamento de informações. Tais sistemas são frequentemente utilizados para extrair e analisar automaticamente as informações úteis contidas em imagens e vídeos.

[0003] Considerando a elevada resolução de câmeras digitais recentes e limitações de largura de banda, é muito importante o desenvolvimento de soluções que podem fornecer possibilidades de reduzir a quantidade de dados que devem ser transferidos através da rede. Além disso, ter menos dados também reduz o impacto dos requisitos de armazenamento em qualquer sistema. A redução da resolução espacial de imagens não é uma opção neste cenário pelo fato de que imagens de baixa resolução tornam a maioria das técnicas de visão computacional muito menos precisa. Uma resolução mínima específica é necessária a fim de realizar cada tarefa de visão computacional como, por exemplo, o reconhecimento de expressão facial em imagens de rostos.

[0004] Muitos cenários têm limitações de infraestrutura, incluindo conexão de Internet/largura de banda deficiente e também pouco espaço para armazenar

arquivos. Mas mesmo quando não existem preocupações sobre as limitações de infraestrutura e de largura de banda, a transmissão e armazenamento de vídeos brutos é um desafio, possivelmente tornando alguns sistemas inviáveis na prática por causa da grande quantidade de dados a serem transmitidos e armazenados. Como exemplo, considerando-se cenário escolar em que faces de alunos precisam ser extraídas de imagens para identificação posterior, cada face deve ser representada por 30 a 40 pixels na horizontal, segundo recomendação de softwares atuais de reconhecimento de face. Faces gravadas entre 5 e 10 metros de distância da câmera, com resolução de vídeo de 1920x1080, são representadas no quadro final com 65 a 30 pixels na horizontal, ou seja, criticamente perto da resolução mais baixa necessária para tarefas de identificação. Portanto, a resolução de vídeo de 1920x1080 seria a mínima necessária e, nesse cenário de aplicação, uma aula de 30 minutos precisaria de pelo menos 4 GB de espaço de armazenamento. Considerando-se que várias aulas devem ser registradas em vídeo diariamente e simultaneamente, isso representa uma quantidade considerável de informação a ser transmitida e armazenada. É evidente que esta enorme quantidade de informação de vídeo gerada não é um problema apenas no cenário escolar.

[0005] As soluções atuais não abordam todo o processo de criação otimizada e de compressão de imagens/vídeos, dependendo do contexto desejado. Técnicas de *tiled streaming* e codificação de vídeo baseada em região de interesse (*Region-of-Interest* – ROI) são duas soluções relacionadas. De modo a reduzir a largura de banda, técnicas de *tiled streaming* podem codificar uma sequência de vídeo dividindo os seus quadros em uma grade de grupos (*tiles*) independentes. Uma imagem ou quadros de um vídeo podem ser divididos inicialmente em grupos (*tiles*) e, em seguida, codificados e armazenados de forma escalar. Este conteúdo pode ser então transmitido com uma resolução de qualidade ou espacial compatível com a largura de banda disponível. Por exemplo, uma versão da sequência em resolução inferior pode ser inicialmente transmitida até que um usuário dê um *zoom* e, depois disso, apenas os grupos (*tiles*) que cobrem a ROI selecionada pelo usuário podem ser transferidos em maior resolução. O codec de imagem JPEG-XR, que é o estado da técnica, é um exemplo de um codec escalável que permite agrupamento (*tiling*). Nos métodos de codificação de vídeo baseados em ROI, a identificação do primeiro plano-plano de fundo (*foreground-background*) é conduzida de modo que as regiões do plano de fundo são mais

comprimidas na etapa de codificação, reduzindo o consumo de largura de banda.

[0006] Como a maioria dos sistemas baseados em visão computacional pode exigir imagens/vídeos de alta resolução para funcionar corretamente, apenas realizar compressão não é aceitável. Uma alternativa interessante para economizar armazenamento e ainda manter resolução suficiente para tarefas de visão computacional é criar imagens/vídeos contendo apenas os objetos de interesse e, em seguida, codificá-los corretamente. Gerando inicialmente tais imagens/vídeos, a etapa de codificação seguinte aproveita a similaridade e proximidade dos objetos de interesse para realizar uma compressão ainda mais eficaz. Portanto, há um ganho duplo: um relacionado com a geração de conteúdo e outro relacionado com a compressão otimizada.

[0007] Na presente invenção, como será detalhado a seguir, as imagens/vídeos são criados a partir de objetos de interesse, codificados com resolução espacial normalizada e resolução específica de qualidade, dependendo do contexto. A resolução espacial normalizada é conseguida através de técnicas de aumento (*up-sampling*) e redução (*down-sampling*) de taxa de amostragem e as diferentes resoluções de qualidade são alcançadas por parâmetros de codificação adequados (por exemplo, parâmetros de

quantização diferentes) selecionados durante o processo de compressão. Portanto, a utilização da presente invenção é uma solução interessante para a compressão, mantendo resolução suficiente para sistemas baseados em visão computacional.

[0008] O trabalho intitulado: "*Region of Interest Encoding in Video Conference Systems*", publicado por C Bulla et al., em: *The Fifth International Conferences on Advances in Multimedia* (MMedia), de 2013, apresenta um sistema de codificação baseado em região de interesse para aplicações de conferência de vídeo. O sistema é dividido em dois módulos: o emissor e o receptor. O emissor compreende um detector facial para detectar faces em vídeos como regiões de interesse (ROI), um método de rastreamento para rastrear cada RoI nos quadros, e um esquema de codificação de RoI que codifica as regiões de interesse com boa qualidade e o fundo com qualidade inferior. O fluxo de vídeo codificado é transmitido para todos os clientes que recebem, ou receptores, que podem decodificá-lo, cortar as regiões de interesse, e exibi-las. A última etapa de exibição é chamada "Composição de Cenário" e é obtida mostrando apenas as pessoas detectadas. Cada pessoa é escalada e colocada lado a lado no cliente receptor. Diferentemente do trabalho de C Bulla et al., a presente

invenção realiza a “Composição de Cenário” localmente, ou seja, agrupa as regiões de interesse em um quadro antes de transmitir o vídeo, o que permite economias na transmissão de dados. No trabalho de C Bulla et al., a composição de cena é feita no receptor, o que significa que os quadros completos são transmitidos através da rede. A segunda diferença é que a composição de cena no trabalho de C Bulla et al. depende de parâmetros de visualização, enquanto que a presente invenção depende dos parâmetros definidos pelo usuário influenciados pela aplicação alvo, tornando-a mais ampla. A terceira diferença é relacionada com a aplicação alvo. No trabalho de C Bulla et al., o vídeo final é visto pelos usuários e, para este fim, a composição de cena deve ser visualmente agradável, com alinhamento espacial, espaços entre os rostos, etc. Na presente invenção, os objetos de interesse podem ser organizados em uma grade quadrada, por exemplo, para melhor explorar similaridades e, consequentemente, obter uma melhor compressão. Além disso, o método apresentado no trabalho de C Bulla et al. é aplicável somente para conferência de vídeo. Todos os detalhes foram discutidos para alcançar melhores resultados neste cenário. O sistema no trabalho de C Bulla et al. só funciona para faces de pessoas, enquanto a presente invenção pode trabalhar com qualquer objeto de interesse. A

presente invenção é muito mais genérica no sentido de que pode ser aplicada a vários outros cenários.

[0009] O documento de patente US 2013/107948 A1, intitulado: "*Context-Based Encoding and Decoding*", publicado em 02 de maio de 2013, descreve um codec que leva em consideração as regiões de interesse similares entre quadros para produzir melhores previsões do que a estimativa e compensação de movimento com base em bloco. Instâncias de objetos similares são associadas através dos quadros do vídeo para formar sequências relacionadas a blocos específicos de dados de vídeo a serem codificados. Diferentemente do documento US 2013/107948 A1, a presente invenção não propõe um novo codec, mas sim apresenta um método de organização de dados que permite que os codecs atuais produzam resultados mais eficientes.

[0010] O documento de patente WO 2014/025319 A1, intitulado: "*System and Method for Enabling User Control of Live Video Stream(s)*", publicado em 13 de fevereiro de 2014, descreve um sistema que permite que vários usuários controlem fluxos de vídeo ao vivo de forma independente, por exemplo, ao solicitar ampliação (*zoom*) independente de áreas de interesse. Considera que um fluxo atual é obtido e armazenado num número de segmentos de vídeo em diferentes resoluções. Cada quadro dos segmentos de vídeo é codificado

em uma técnica de agrupamento (*tiling*) virtual, em que cada quadro dos segmentos de vídeo codificados é dividido em uma disposição de grupos (*tiles*), e cada grupo (*tile*) compreende uma disposição de fatias. Mediante solicitação do usuário para ampliar uma área de interesse específica, os grupos (*tiles*) correspondentes a essa área, em um segmento de vídeo adequado com maior resolução, são transferidos para serem exibidos ao usuário. As fatias fora da área de interesse são removidas antes da exibição. A presente invenção difere do documento WO 2014/025319 A1 em muitos aspectos. Em primeiro lugar, a presente invenção cria uma imagem ou vídeo único contendo apenas objetos de interesse representados com uma resolução espacial normalizada a ser transmitida e armazenada, e não para armazenar várias imagens/vídeos com resoluções diferentes. No documento WO 2014/025319 A1, a região de interesse, ou seja, a zona que vai ter uma maior resolução, é definida em tempo real pelo usuário e a resolução daquela área também é escolhida com base na solicitação do usuário. No método da presente invenção, os objetos de interesse podem ser detectados através da aplicação de um algoritmo de detecção de objetos de acordo com a especificação do usuário. A criação da imagem/vídeo final contendo objetos com resolução normalizada será feita apenas uma vez e, em

seguida, vai ser transmitida e armazenada. Outra diferença é a aplicação final. A solução apresentada no documento WO 2014/025319 A1 tem uma aplicação específica que se relaciona com a exibição de uma área de interesse com uma resolução específica. O método da presente invenção cria uma imagem/vídeo final com objetos representados com resolução normalizada a serem analisados por um sistema baseado em visão. Logo, torna-se evidente que o método da presente invenção tem aplicação mais ampla, uma vez que os seus parâmetros não estão limitados a solicitações específicas de usuários para controlar fluxos de vídeo.

[0011] O trabalho intitulado: "*Supporting Zoomable Video Streams with Dynamic Region-of-Interest Cropping*", publicado por NQM Khiem et al., em ACM Conference on Multimedia Systems (MMSys), de 2010, apresenta dois métodos para streaming de uma região arbitrária de interesse (ROI) a partir de um vídeo de alta resolução para suportar ampliação e visão panorâmica: streaming em blocos (*tiled*) e streaming monolítico. O primeiro método relaciona-se com a presente invenção pelo fato de que divide cada quadro (*frame*) de um vídeo em uma disposição de blocos. Mas de forma diferente, os blocos são codificados e armazenados como um fluxo independente, na sua maior resolução. Na presente invenção, todos os blocos são representados com a

mesma resolução espacial. No trabalho de NQM Khiem et al., um usuário recebe do servidor uma versão reduzida de um vídeo e solicita uma ampliação em uma área específica. Os fluxos de blocos que se sobrepõem com a região de interesse são enviados para o usuário em uma resolução maior. Na abordagem da presente invenção, a imagem/vídeo final é transmitida para o servidor para ser ainda armazenada e analisada por um sistema baseado em visão computacional.

[0012] O trabalho intitulado: "*Adaptive Encoding of Zoomable Video Streams Based on User Access Pattern*", publicado por Khiem Quang Minh Ngo, Ravindra Guntur e Wei Tsang Ooi, *ACM Conference on Multimedia Systems* (MMSys), de 2011, apresenta um método para criar vídeos passíveis de ampliação, permitindo aos usuários seletivamente ampliar e ter o panorama de regiões de interesses dentro do vídeo para visualização em resoluções mais altas. A ideia é a mesma que a do trabalho de NQM Khiem et al., mas em vez de dividir cada quadro em uma grade fixa de blocos, os padrões de acesso de usuário são levados em consideração. Considerando os históricos de padrões de acesso de usuários para regiões de um vídeo, o método cria um mapa de calor (*heatmap*) com a probabilidade de uma região ser acessada (ampliada) pelos usuários. O trabalho de Khiem et al. fornece um algoritmo guloso para criar um mapa de blocos de

modo que cada bloco contenha uma região de interesse provável. Cada bloco de vídeo de alta resolução na mesma posição, considerando todos os quadros, é então codificado em um fluxo independente. Quando um usuário solicita uma região de interesse, os blocos sobrepostos são enviados para serem exibidos com largura de banda mínima porque a região de interesse provavelmente está inteiramente dentro de um bloco. As diferenças em relação à presente invenção, além daquelas discutidas no trabalho de NQM Khiem et al., são que: no trabalho de Khiem et al. os blocos são adaptáveis; os blocos da presente invenção não são codificados como fluxos diferentes e estão relacionados aos objetos alvo extraídos dos quadros do vídeo de entrada.

[0013] O trabalho intitulado: "*Adaptive Resolution Image Acquisition Using Image Mosaicing Technique from Video Sequence*", publicado por S Takeuchi et al., em: *Proceedings of International Conference on Image Processing*, 2000, descreve um método de mosaico de imagem em camadas a partir de uma sequência de vídeo para criar uma imagem com resolução adaptativa. O método considera como entrada uma sequência de vídeo capturada com uma câmera que amplia certas regiões onde texturas finas estão presentes. Cada quadro é classificado em uma camada, de acordo com o nível de ampliação. As imagens em cada camada

são, então, registradas para criar uma imagem única. Ao fazer isso, o método cria uma imagem em camadas em que cada camada representa uma imagem com uma resolução diferente. Diferentemente, o método da presente invenção compõe uma imagem final usando uma disposição que contém os objetos de interesse em uma resolução desejada.

[0014] O documento de patente US 8,184,069 B1, intitulado: "*Systems and Methods for Adaptive Transmission of Data*", publicado em 22 de abril de 2012, descreve um sistema e método para transmitir, receber e exibir dados. Ele fornece uma taxa de transmissão de dados constante para o dispositivo e controla a largura de banda pela apresentação de informação direcionada a uma área de interesse do usuário. Por exemplo, a largura de banda pode ser reduzida ao apresentar informação em alta resolução direcionada a uma área de interesse (por exemplo, uma área para a qual o usuário está olhando), e dados em baixa resolução direcionados a outras áreas. Para determinar a área de interesse, o método utiliza um display de cabeça utilizado pelo usuário e prioriza a transmissão de dados com base nesta informação. Diferentemente, a presente invenção não necessita de qualquer dispositivo de usuário para detectar áreas de interesse. Além disso, o documento

US 8,184,069 B1 não inclui qualquer método específico para compor os quadros finais.

[0015] O documento de patente US 8,665,958 B2, intitulado: "*Method and Apparatus for Encoding Video Signal Using Motion Compensation Based on Affine Transformation*", publicado em 04 de março de 2014, apresenta um método de codificação de vídeo que pode determinar se um bloco inclui um objeto com uma transformação afim. Em caso positivo, o método gera um bloco de predição realizando uma compensação de movimento com base em transformação afim no bloco atual, alcançando alta eficiência na codificação/decodificação de vídeo. A presente invenção extrai objetos a partir dos quadros do vídeo de entrada e cria blocos a partir deles, sem considerar qualquer transformação, apenas ajustando a sua resolução. A invenção propriamente dita do documento US 8,665,958 B2 não pode alcançar os mesmos resultados obtidos pelo método da presente invenção, mas poderia ser aplicado como um módulo adicional/complementar (ainda que opcional).

Sumário da Invenção

[0016] A presente invenção apresenta um método e sistema para gerar imagens e vídeos nos quais objetos de interesse são codificados em uma resolução desejada, dependendo dos parâmetros informados pelo usuário.

[0017] O método e sistema da presente invenção são destinados para utilização em sistemas que precisam analisar imagens e vídeos digitais, e extrair informações relevantes deles, mantendo baixa largura de banda e armazenamento de dados.

[0018] Uma concretização da invenção é composta por uma câmera que pode capturar os objetos de interesse em uma cena e um dispositivo com capacidade de processamento suficiente para executar o sistema da presente invenção, que compreende quatro módulos para criar o vídeo final: detecção de objeto, ajuste de resolução espacial, composição de quadros, e codificação de vídeo.

[0019] Os objetivos da presente invenção são alcançados através de um método para composição e compressão de vídeo com base em contexto a partir de objetos com resolução espacial normalizada que compreende as etapas de:

receber como dados de entrada um conjunto de quadros de vídeo digital ou imagem, com a maior resolução possível, e os parâmetros, que informam as categorias de objetos alvo e uma resolução espacial para cada categoria;

detectar e extrair os objetos desejados, em cada quadro do vídeo de entrada, considerando as categorias informadas como parâmetro;

ajustar a resolução espacial dos objetos extraídos de acordo com os parâmetros;

compor quadros finais, cada um correspondente a um quadro do vídeo de entrada, com os objetos extraídos e ajustados espacialmente em uma grade;

gerar um vídeo final através do processamento de todos os quadros finais com um algoritmo de codificação que utiliza as similaridades visuais e correlações locais nos quadros (tanto espacialmente em cada quadro quanto temporalmente através dos quadros);

transmitir os vídeos finais e os dados de coordenadas correspondentes para um sistema de análise baseado em visão, onde são armazenados e analisados.

[0020] Adicionalmente, a concretização preferida da invenção descreve um sistema para composição e compressão de vídeo com base em contexto a partir de objetos com resolução espacial normalizada que compreende as etapas de:

um módulo de detecção de objeto que detecta uma categoria de objetos alvo e extrai seus dados de coordenada;

um módulo de ajuste de resolução espacial que ajusta a amostragem do objeto detectado para coincidir com a resolução informada como parâmetro;

um módulo de composição de quadro que organiza os objetos detectados de cada quadro de entrada em uma grade para criar um quadro final; e

um módulo de codificação de vídeo que codifica o vídeo final utilizando correlações espaciais e temporais de objetos similares em posições similares nos quadros finais subsequentes.

[0021] O sistema recebe como entrada um vídeo ou uma imagem digital com a maior resolução possível, e os parâmetros que informam as categorias de objetos alvo e uma resolução espacial para cada categoria. Com base nestes dados de entrada, o sistema executa para cada categoria informada as etapas de: (i) detectar e extrair, para cada quadro de entrada, os objetos desejados relacionados com a categoria considerada; (ii) ajustar a resolução espacial dos objetos extraídos de acordo com os parâmetros; (iii) para cada quadro de entrada, compor um quadro final correspondente com os objetos extraídos e ajustados espacialmente dispostos em uma grade; (iv) gerar um vídeo final pelo processamento dos quadros finais com um algoritmo de codificação que poderia se beneficiar das semelhanças visuais e correlações locais nos quadros (tanto espacialmente em cada quadro quanto temporalmente entre os quadros). As similaridades visuais melhoram

consideravelmente a eficácia do algoritmo de codificação, com consequente aumento da capacidade de compressão.

Breve Descrição das Figuras

[0022] Os objetivos e vantagens da presente invenção se tornarão mais claros através da seguinte descrição detalhada de uma concretização exemplar e não limitativa em conjunção com as figuras em anexo, nas quais:

[0023] A Figura 1 descreve um cenário no qual uma concretização do método da presente invenção é aplicada.

[0024] A Figura 2 mostra as entradas para um sistema que implementa uma concretização do método da presente invenção.

[0025] A Figura 3 representa uma perspectiva geral do sistema implementando uma concretização do método da presente invenção.

[0026] A Figura 4 representa o fluxograma de uma concretização do método da presente invenção (funcionamento da invenção), implementado pelo sistema.

Descrição Detalhada da Invenção

Cenário e aplicação da presente invenção

[0027] A Figura 1 descreve um cenário no qual uma concretização do método da presente invenção é aplicada. O cenário é composto por pelo menos uma câmera 100 que pode tirar fotos/vídeo da cena desejada completa 102, retratando

os objetos necessários. O método da presente invenção pode ser executado na câmera 100 ou em qualquer dispositivo externo 101 com capacidade de processamento, fixado na câmera 100. A cena 102 pode ser uma sala de aula com estudantes, um estacionamento com carros, um lugar público (por exemplo, aeroporto, estádio) ou qualquer cena relacionada onde é necessário analisar uma ou mais categorias de objetos, tais como faces, mãos, placas de carro, carros, etc.

As entradas para o sistema proposto

[0028] A Figura 2 ilustra a entrada de dados 200 necessária para a concretização do método da presente invenção, que é composta de quadros de vídeo 201 e parâmetros 202. Os quadros de vídeo 201 são aqueles capturados pela câmera 100 com a mais alta resolução disponível, preferencialmente em formato RAW, em que os dados provenientes do sensor da câmera são minimamente processados. Os parâmetros 202 são especificados pelo usuário do sistema e representam as exigências da tarefa de visão computacional final, que compreendem: (i) uma ou mais categorias de objetos alvo a serem detectados nos quadros de entrada, fornecendo nomes predefinidos, como “face” e “mão”; ou fornecendo uma imagem modelo dos objetos alvo; ou pelo fornecimento de coordenadas específicas de objetos

alvo fixos; (ii) uma resolução espacial, em pixels, para cada categoria. Por exemplo, considerando quadrados múltiplos de 16 pixels, parâmetros “face/5; mão/3” significa que “faces” serão detectadas e representadas com 80x80 pixels (5x16 = 80), e “mãos” serão detectadas e representadas com 48x48 pixels (3x16 = 48).

Visão geral do sistema

[0029] A Figura 3 ilustra a visão geral do sistema 300 exemplar de acordo com uma concretização da presente invenção. A finalidade do sistema 300 é compor e codificar um vídeo para cada categoria de objetos alvo informada como parâmetro 202 de entrada de dados 200, a fim de ser transmitido para qualquer sistema de análise baseado em visão computacional 350. Para cada quadro de vídeo de entrada 201, o sistema 300 cria um quadro final 331 com uma grade de objetos na resolução desejada, de acordo com os parâmetros 202 informados pelo usuário, e todos esses quadros finais 331 são utilizados para gerar uma sequência de vídeo codificado final 341.

[0030] Assim, o sistema 300 exemplar compreende quatro módulos: detecção de objetos 310, ajuste de resolução espacial 320, composição de quadros 330 e codificação de vídeo 340.

[0031] Para cada quadro de vídeo de entrada 201, o módulo de detecção de objetos 310 detecta a primeira categoria de objetos alvo 311 e extrai os seus dados de coordenadas 312. O módulo de ajuste de resolução espacial 320 realiza o aumento ou redução de amostragem de cada objeto detectado 311 para coincidir com a resolução desejada informada como parâmetro 202. O módulo de composição de quadro 330 organiza os objetos detectados 311 de cada quadro de entrada 201 em uma grade para criar um quadro final 331. O módulo final, ou seja, de codificação de vídeo 340, codifica o vídeo final 341 através da aplicação de um codec que aproveita as correlações espaciais e temporais de objetos similares em posições similares em quadros finais subsequentes 331. Todo o processo é repetido para criar um vídeo final 341 – composto por uma pluralidade de quadros finais 331 – para cada categoria de objetos alvo. Os vídeos finais 341 e os dados de coordenadas correspondentes 312 são transmitidos (por exemplo, através da Internet) de forma eficiente para um sistema de análise baseado em visão computacional 350, onde são armazenados e analisados. O sistema baseado em visão computacional 350 pode ser relacionado a uma variedade de cenários: análise da expressão facial em estudantes durante uma classe; procura de carros roubados

em um estacionamento (ou ruas) por suas placas; análise visual de pragas nas plantações de frutas/vegetais; análise visual do desempenho dos atletas em campo; análise visual de áreas perigosas em câmeras de vigilância, reconhecimento facial em locais públicos, etc.

[0032] Ainda com referência à Figura 3, cada módulo do sistema 300 será descrito em mais detalhes abaixo.

Módulo de detecção de objetos

[0033] O módulo de detecção de objetos 310 recebe como entrada 200 os quadros de vídeo 201 e parâmetros 202 especificando as categorias de objetos alvo e uma resolução alvo para cada categoria. Cada categoria pode ser informada de três maneiras distintas: (a) o sistema é previamente treinado para detectar algumas categorias de objetos e, neste caso, apenas o nome da categoria precisa ser informado; (b) o usuário pode fornecer uma imagem modelo do objeto a ser detectado; (c) o usuário pode fornecer as coordenadas de imagem de objetos fixos.

[0034] No primeiro caso (a), o sistema precisa ser treinado para detectar algumas categorias de objetos. Uma possível solução é usar *OverFeat*, um reconhecedor de objeto baseado em rede convolucional (Sermanet et al. "*OverFeat: Integrated Recognition, Localization and Detection Using Convolutional Networks*" em International Conference on

Learning Representations, ICLR 2014, disponível também online em arXiv preprint 1312.6229v4).

[0035] No caso em que o usuário fornece imagens modelo (b), há vários descritores de imagem que podem ser utilizados para descrever, detectar e casar as características locais que representam o objeto alvo. O descritor mais popular é o SIFT divulgado por David G. Lowe, intitulado: "*Object Recognition from Local Scale-Invariant Features*", em *IEEE International Conference on Computer Vision*, 1999. No entanto, outros descritores podem ser considerados, tais como SURF, divulgado por Herbert Bay et al., intitulado: "*SURF: Speed Up Robust Features*", em *European Conference on Computer Vision*, 2006. Caso o usuário forneça as coordenadas de imagem dos objetos alvo (c), a detecção de objetos é direta e imediata, uma vez que o módulo 310 vai delimitar os objetos de acordo com as coordenadas específicas informadas.

[0036] Este módulo 310 gera dois tipos diferentes de informação: os objetos alvo detectados 311 como diferentes blocos de imagens e suas coordenadas de imagem 312. Para cada quadro do vídeo de entrada 201, este módulo 310 detecta e extrai os objetos alvo 311, como descrito acima. As coordenadas 312 podem ser os pixels do canto superior esquerdo e do canto inferior direito do retângulo

(“caixa delimitadora”) que envolve o objeto detectado no quadro do vídeo de entrada original 201. Além da “caixa delimitadora”, a detecção de objetos pode ser conseguida por diferentes modos, tais como o conjunto de coordenadas de todos os pixels que delimitam o objeto. O módulo de detecção de objetos também pode tirar proveito de um processo de rastreamento, que pode rastrear objetos detectados nos quadros. Se o mesmo objeto é frequentemente posicionado no mesmo local da grade final (após o módulo de composição de quadro), há uma possibilidade de obtenção de taxas de compressão ainda mais elevadas no vídeo final.

Módulo de ajuste de resolução espacial

[0037] O módulo de ajuste de resolução espacial 320 processa os objetos 311 detectados no módulo 310 anterior, de modo que eles sejam representados na resolução espacial alvo informada pelos parâmetros do usuário 202. Se a resolução atual do objeto é menor do que a resolução desejada, um processo de aumento de amostragem é realizado. Caso contrário, um processo de redução de amostragem é realizado.

[0038] A operação de redução de amostragem (*down-sampling*) é mais simples do que a de aumento de amostragem (*up-sampling*) e pode ser feita de várias maneiras. Considerando cada objeto detectado como um bloco de imagem

diferente, pode-se simplesmente retirar cada outra coluna e/ou linha da imagem. Se várias colunas e/ou linhas tiverem que ser desconsideradas a fim de conseguir a resolução espacial final desejada, processos de suavização podem ser aplicados para reduzir as descontinuidades que podem ser geradas em um bloco de imagem com muitos detalhes, por exemplo. Uma possibilidade, neste caso, seria calcular a média das colunas consecutivas que devem ser eliminadas e então substituir as colunas remanescentes pela coluna média.

[0039] Por outro lado, a operação de aumento de amostragem (*up-sampling*) é mais complexa porque deve preservar a nitidez e a estrutura do bloco de imagem ao criar novas colunas/linhas, a fim de aumentar a resolução espacial. Este processo é frequentemente referido como ampliação (*up-sampling*). As abordagens de ampliação tradicionais são baseadas em interpolação bilinear, bicúbica ou de *spline*. Estes métodos são rápidos, mas normalmente geram imagens borradas. Para evitar esse problema, métodos de interpolação que tentam preservar características de imagem, como a direcionalidade, podem ser aplicados (Vladan Velisavljecic et al., intitulado: "*Image Interpolation with Directionlets*", *IEEE International Conference on Acustics, Speech and Signal*

Processing, 2008; Hao Jiang et al., intitulado "*A New Direction Adaptive Scheme for Image Interpolation*", *IEEE International Conference on Acustics, Speech and Signal Processing*, 2002; Sadik Bayrakeri et al., intitulado: "*A New Method for Directional Image Interpolation*", *IEEE International Conference on Acustics, Speech and Signal Processing*, 1995). Outra possibilidade de aumento de amostragem é usar métodos baseados em super resolução (Freeman et al., intitulado: "Example-based Super-resolution", *IEEE Computer Graphics and Applications*, 2002).

Módulo de composição do quadro

[0040] O módulo de composição de quadro 330 compõe os quadros finais 331. Para cada quadro do vídeo de entrada 201, os blocos de imagem com os objetos detectados (já ajustados espacialmente pelo módulo anterior) são organizados em uma grade. Uma possibilidade para determinar a configuração da grade (largura e altura) é considerar a informação sobre o número máximo de objetos que podem ser detectados no vídeo. Por exemplo, em uma sala de aula, pode-se saber de antemão o número máximo de alunos. Para melhor compressão, a grade deve ser tão quadrada quanto possível, por exemplo, com largura e altura correspondente à raiz quadrada do número de objetos detectados no quadro.

No entanto, dependendo do algoritmo de codificação de vídeo a ser utilizado, a grade pode ter diferentes formas, como uma única linha ou única coluna, por exemplo.

Módulo de codificação de vídeo

[0041] O módulo de codificação de vídeo 340 junta, inicialmente, todos os quadros gerados anteriormente numa sequência de vídeo bruto e, em seguida, aplica-se um codec de vídeo padrão - tal como H.264/AVC ou HEVC - a fim de gerar uma sequência de vídeo codificado final (341 - pluralidade de quadros finais 331) pronta para ser armazenada e/ou transmitida e/ou analisada por sistemas baseados em visão computacional 350. Existem várias vantagens na aplicação de tais codecs de vídeo para as sequências de vídeo bruto. Todos os quadros gerados anteriormente têm objetos da mesma categoria e, devido às técnicas de previsão espacial incluídas nos codecs de vídeo mencionados, toda essa correlação espacial é reduzida. Além disso, uma vez que os quadros diferentes têm objetos similares em posições similares, a correlação temporal inerente também é reduzida devido às técnicas de estimativa de movimento e compensação de movimento que fazem parte dos codecs de vídeo mencionados. Finalmente, cada bloco de imagem correspondente a cada objeto 311 dentro de cada quadro 331 pode ser codificado com uma resolução de

qualidade diferente. Por exemplo, um bloco com uma alta resolução espacial inicial antes de passar pelo módulo de ajuste de resolução espacial 320 tem muitas informações e pode ser mais comprimido que outro bloco com uma resolução espacial inicial mais baixa que não podem ter perdas de ainda mais informação durante o processo de codificação. O processo de aplicação de um alto nível de compressão para um bloco de imagem significa codificar este bloco com um alto parâmetro de quantização (QP), enquanto que a aplicação de um baixo nível de compressão significa comprimir este bloco com um baixo QP. Ambos os codecs de vídeo mencionados – H.264/AVC e HEVC – permitem a codificação de cada bloco de imagem com um QP diferente, o que significa que o quadro comprimido final 331 é composto por blocos codificados com diferentes resoluções de qualidade e que o processo de compressão total de quadro é otimizado.

[0042] O vídeo final 341, juntamente com os dados de coordenadas (312) correspondentes gerados pelo módulo 310, é eficientemente transmitido para um sistema de análise baseado em visão computacional 350, onde são armazenados e analisados.

Concretização do método da presente invenção

[0043] Como descrito acima, o principal objetivo do sistema 300 é implementar o método 400, o que corresponde à operação genérica da invenção. De acordo com a figura 4, o método 400 compreende as etapas de:

- receber 405 como dados de entrada 200 quadros 201 de um vídeo digital ou imagem, com a maior resolução possível, e os parâmetros 202, que informam as categorias de objetos alvo e uma resolução espacial para cada categoria;
- para cada categoria de objeto informado como parâmetros 202 e para cada quadro 201 do vídeo de entrada:
    - detectar e extrair 410 os objetos desejados 311, considerando as categorias informadas [Esta etapa 410 é implementada pelo módulo 310 do sistema 300];
    - ajustar 420 a resolução espacial dos objetos extraídos 311 de acordo com os parâmetros 202 [Esta etapa 420 é implementada pelo módulo 320 do sistema 300];
    - compor 430 um quadro final correspondente 331 com os objetos 311 extraídos e ajustados espacialmente agrupados numa grade [Esta

etapa 430 é implementada pelo módulo 330 do sistema 300];

- gerar 440 um vídeo final 341 através do processamento de todos os quadros finais 331 com um algoritmo de codificação que poderia se beneficiar das semelhanças visuais e correlações locais nos quadros (tanto espacialmente em cada quadro quanto temporalmente entre vários quadros). As semelhanças visuais melhoram consideravelmente a eficácia do algoritmo de codificação, aumentando consequentemente a capacidade de compressão [Esta etapa 440 é implementada pelo módulo 340 do sistema 300];
- transmitir 450 eficientemente os vídeos finais 341 e os dados de coordenadas 312 correspondentes para um sistema de análise baseado em visão computacional 350, onde são armazenados e analisados [Esta etapa 450 corresponde à interface entre o sistema 300 e o sistema baseado em visão computacional externo 350].

[0044] Embora a presente invenção tenha sido descrita em conexão com certa concretização preferencial, deve ser entendido que não se pretende limitar a invenção àquela concretização particular. Ao contrário, pretende-se

cobrir todas as alternativas, modificações e equivalentes possíveis dentro do espírito e do escopo da invenção, conforme definido pelas reivindicações em anexo.

REIVINDICAÇÕES

1. Sistema para composição e compressão de vídeo com base em contexto a partir de objetos com resolução espacial normalizada que compreende:

pelo menos um processador que compreende:

um módulo de entrada de dados para receber uma categoria de um objeto e uma entrada de parâmetro de resolução espacial por um usuário de sistema;

um módulo de detecção de objeto que detecta objetos da categoria de entrada pelo usuário de sistema e extrai dados de coordenada dos objetos detectados;

um módulo de ajuste de resolução espacial que ajusta o objeto detectado para coincidir com a entrada de parâmetro de resolução espacial pelo usuário de sistema;

um módulo de composição de quadro que organiza os objetos detectados de cada quadro de entrada em uma grade para criar um quadro final; e

um módulo de codificação de vídeo que codifica o conjunto de quadros de saída em um vídeo final utilizando correlações espaciais e temporais de objetos similares em uma posição similar em quadros de saída subsequentes;

em que o vídeo final, e os dados de coordenadas, são transmitidos para um sistema de análise baseado em visão para serem armazenados e analisados;

caracterizado pelo fato de que:

o módulo de detecção de objeto recebe como entrada os quadros de vídeo e parâmetros especificando as categorias dos

objetos alvo e uma resolução espacial alvo para cada categoria, e para cada quadro de vídeo, esse módulo detecta e extrai os objetos alvo e suas coordenadas de imagem correspondentes; e

o módulo de ajuste de resolução espacial processa os objetos detectados, de modo que os objetos são representados em uma resolução espacial alvo especificada pelos parâmetros, e

se uma resolução atual do objeto for menor do que a resolução especificada pelo parâmetro, um processo de aumento de amostragem é realizado, caso contrário, um processo de redução de amostragem é realizado.

2. Sistema, de acordo com a reivindicação 1, **caracterizado pelo** fato de que a entrada inclui um vídeo digital ou um conjunto de quadro de imagem.

3. Sistema, de acordo com a reivindicação 2, **caracterizado pelo** fato de que o conjunto de quadros de vídeo digital é obtido por uma câmera; e

o parâmetro representa um requisito do sistema de análise com base em visão que compreende:

um ou mais tipos de objetos alvo a serem detectados nos quadros de entrada, fornecendo nomes predefinidos; ou fornecendo uma imagem modelo do objeto alvo; ou fornecendo coordenadas fixas específicas de objetos alvo; e

uma resolução espacial em pixels para a categoria.

4. Sistema, de acordo com a reivindicação 1, **caracterizado pelo** fato de que a detecção e extração dos objetos alvo é implementada por um ou mais dentre um

reconhecedor de objeto baseado em rede convolucional, vários descritores de imagem, e delimitação do objeto alvo de acordo com coordenadas específicas.

5. Sistema, de acordo com a reivindicação 1, **caracterizado pelo** fato de que o módulo de composição de quadro organiza blocos de imagem com os objetos detectados já ajustados espacialmente em uma grade que corresponde ao quadro de saída, considerando o número máximo de objetos que podem ser detectados no vídeo.

6. Sistema, de acordo com a reivindicação 1, **caracterizado pelo** fato de que o módulo de codificação de vídeo junta todos os quadros gerados anteriormente numa sequência de vídeo bruto e aplica um codec de vídeo padrão, a fim de gerar uma sequência de vídeo codificado final pronta para ser armazenada e/ou transmitida e/ou analisada por sistemas de análise baseados em visão computacional.

7. Sistema para composição e compressão de vídeo com base em contexto a partir de objetos com resolução espacial normalizada que compreende:

pelo menos um processador que compreende:

um módulo de detecção de objeto que detecta objetos da categoria especificada como entrada e extrai dados de coordenada dos objetos alvo detectados;

um módulo de ajuste de resolução espacial que ajusta o objeto detectado para coincidir com o parâmetro de resolução especificado por um usuário de sistema;

um módulo de composição de quadro que organiza os objetos

detectados de cada quadro de entrada em uma grade para criar um quadro de saída; e

um módulo de codificação de vídeo que codifica o conjunto de quadros de saída em um vídeo final utilizando correlações espaciais e temporais de objetos similares em uma posição similar em quadros de saída subsequentes;

em que o vídeo final, e os dados de coordenadas, são transmitidos para um sistema de análise baseado em visão para serem armazenados e analisados,

em que o módulo de codificação de vídeo junta todos os quadros gerados anteriormente em uma sequência de vídeo bruta e aplica um codec de vídeo padrão, a fim de gerar uma sequência de vídeo codificada final pronta para ser armazenada e/ou transmitida e/ou analisada pela análise baseada em visão sistema, e

em que cada bloco de imagem correspondente a cada objeto dentro de cada quadro pode ser codificado com uma qualidade diferente, aplicando diferentes parâmetros de quantização, resultando em um quadro compactado final compreendendo blocos com qualidades diferentes, otimizando o procedimento de compressão de vídeo;

caracterizado pelo fato de que:

o módulo de detecção de objeto recebe como entrada os quadros de vídeo e parâmetros especificando as categorias dos objetos alvo e uma resolução espacial alvo para cada categoria, e para cada quadro de vídeo, esse módulo detecta e extrai os objetos alvo e suas coordenadas de imagem correspondentes; e

o módulo de ajuste de resolução espacial processa os objetos detectados, de modo que os objetos são representados em uma resolução espacial alvo especificada pelos parâmetros, e

se uma resolução atual do objeto for menor do que a resolução especificada pelo parâmetro, um processo de aumento de amostragem é realizado, caso contrário, um processo de redução de amostragem é realizado.

8. Sistema, de acordo com a reivindicação 7, **caracterizado pelo** fato de que o codec de é H.264/AVC ou HEVC.

9. Método para composição e compressão de vídeo com base em contexto a partir de objetos com resolução espacial normalizada que compreende as etapas de:

receber, usando um módulo de entrada de dados, como dados de entrada, um vídeo digital ou um conjunto de quadros de imagem e parâmetros, especificados por um usuário de sistema, para uma categoria de um objeto alvo e uma resolução espacial para a categoria;

detectar e extrair os objetos alvo a partir do vídeo de entrada, com base na categoria e o parâmetro;

ajustar a resolução espacial dos objetos extraídos de acordo com o parâmetro;

compor um quadro final com os objetos extraídos e ajustados espacialmente agrupados numa grade;

gerar um vídeo final através do processamento de todos os quadros finais com um algoritmo de codificação que utiliza as similaridades visuais e correlações locais em um quadro; e

preparar o vídeo final e os dados de coordenadas para transmitir para um sistema de análise baseado em visão computacional para armazenamento e análise;

caracterizado pelo fato de que:

o módulo de detecção de objeto recebe como entrada os quadros de vídeo e parâmetros especificando as categorias dos objetos alvo e uma resolução espacial alvo para cada categoria, e para cada quadro de vídeo, esse módulo detecta e extrai os objetos alvo e suas coordenadas de imagem correspondentes; e

o módulo de ajuste de resolução espacial processa os objetos detectados, de modo que os objetos são representados em uma resolução espacial alvo especificada pelos parâmetros, e

se uma resolução atual do objeto for menor do que a resolução especificada pelo parâmetro, um processo de aumento de amostragem é realizado, caso contrário, um processo de redução de amostragem é realizado.

10. Método, de acordo com a reivindicação 9, **caracterizado pelo** fato de que o parâmetro inclui pelo menos um de um nome de categoria predefinido, uma imagem de modelo e coordenadas fixas específicas de objetos alvo.

11. Método, de acordo com a reivindicação 9, **caracterizado pelo** fato de que:

a etapa detectar e extrair os objetos alvo é realizada por um módulo de detecção de objeto;

a etapa de ajustar a resolução espacial dos objetos extraídos é realizada por um módulo de ajuste de resolução

espacial;

a etapa de compor o quadro final é realizada por um módulo de composição de quadro; e

a etapa de gerar um vídeo final é realizada por um módulo de codificação de vídeo.

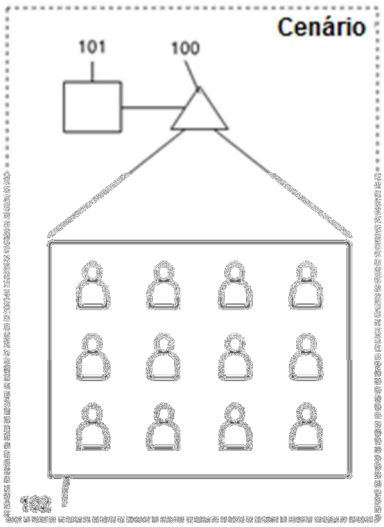

Figura 1

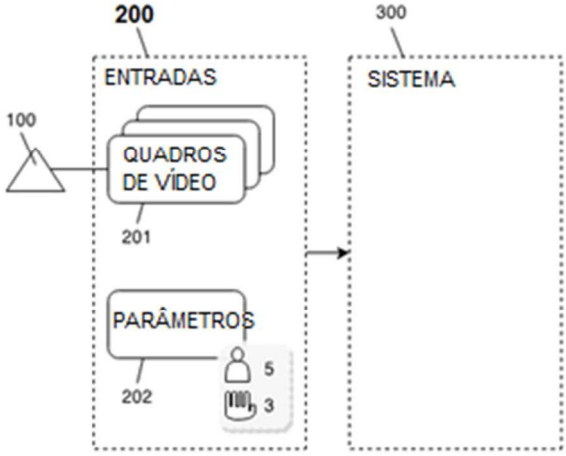

Figura 2

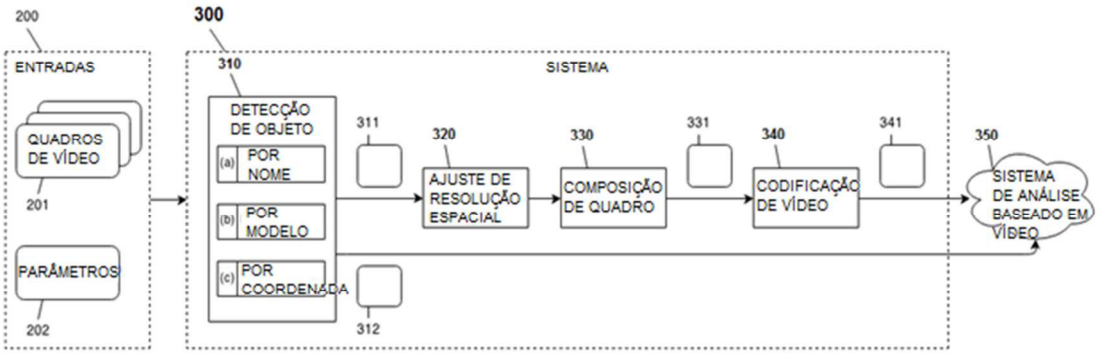

Figura 3

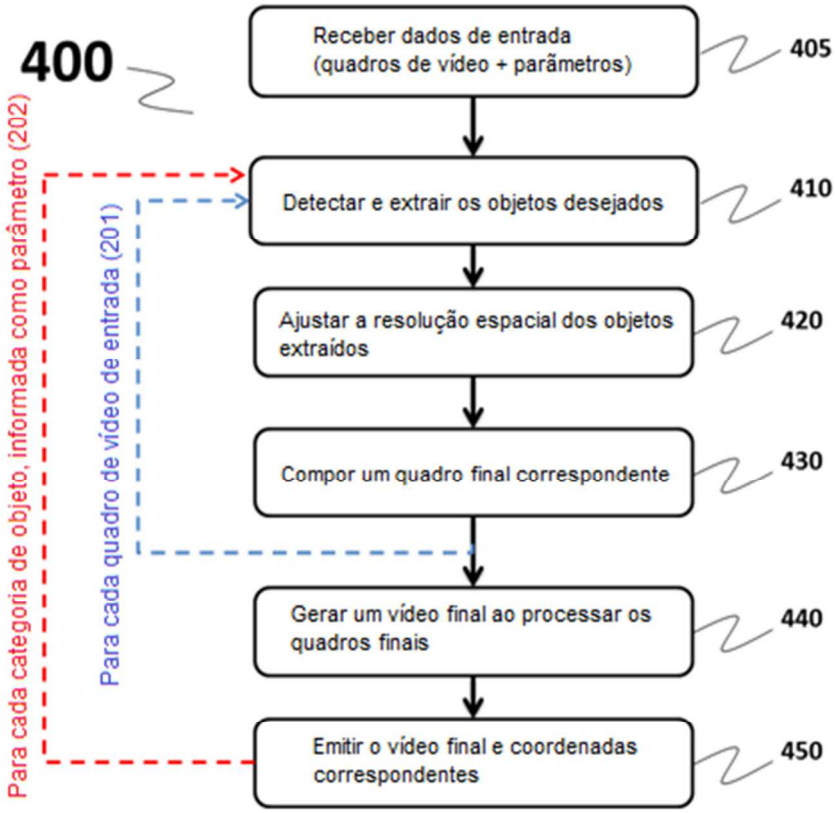

Figura 4