(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 658 963**

(51) Int. CI.:

G06K 9/62 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **11.04.2013 PCT/EP2013/057540**

(87) Fecha y número de publicación internacional: **24.10.2013 WO13156374**

(96) Fecha de presentación y número de la solicitud europea: **11.04.2013 E 13714965 (4)**

(97) Fecha y número de publicación de la concesión europea: **13.12.2017 EP 2839410**

(54) Título: **Procedimiento de reconocimiento de un contexto visual de una imagen y dispositivo correspondiente**

(30) Prioridad:

16.04.2012 FR 1253495

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
13.03.2018

(73) Titular/es:

COMMISSARIAT À L'ÉNERGIE ATOMIQUE ET AUX ÉNERGIES ALTERNATIVES (100.0%)
Bâtiment le Ponant D, 25 rue Leblanc
75015 Paris, FR

(72) Inventor/es:

SHABOU, AYMEN y
LE BORGNE, HERVÉ

(74) Agente/Representante:

CARPINTERO LÓPEZ, Mario

DESCRIPCIÓN

Procedimiento de reconocimiento de un contexto visual de una imagen y dispositivo correspondiente

La presente invención se refiere a un procedimiento de reconocimiento del contexto visual de una imagen y un dispositivo que le corresponde. La presente invención se inscribe en el dominio de la detección y el reconocimiento automático de objetos en unas imágenes o unos flujos de vídeo, denominado dominio de la clasificación semántica. Más precisamente, la presente invención se inscribe en el dominio denominado de la clasificación supervisada. Se aplica principalmente a la videovigilancia, a la visión artificial por ejemplo en un robot o en vehículos, así como a la búsqueda de imágenes.

Diversas aplicaciones requieren que se puedan identificar unos objetos en unas imágenes o en unos flujos de vídeo. El reconocimiento de objetos en imágenes constituye una problemática fundamental en el dominio de la visión artificial.

La clasificación supervisada es una técnica particularmente apropiada para el reconocimiento de objetos en imágenes. La clasificación supervisada implica dos fases principales: una primera fase se realiza fuera de línea, y constituye una fase de aprendizaje que permite establecer un modelo; una segunda fase, denominada fase de ensayo, se realiza en línea y permite determinar una previsión de una etiqueta o “label” de una imagen de ensayo. Estas fases se explican más en detalle a continuación con referencia a la figura 1.

En las aplicaciones a las que se dirige la presente invención, se busca reconocer una categoría de un concepto visual, más que una instancia particular. Por ejemplo, si el concepto visual es el concepto “vehículo”, se busca reconocer cualquier vehículo, incluso si el vehículo presente en la imagen ensayada no está presente en la base de aprendizaje.

Cada una de las dos fases antes citadas comprende principalmente una etapa denominada de extracción de características, comúnmente designada según la terminología inglesa “feature extraction”, que se dirige a describir una imagen, o una parte de una imagen, mediante un juego de características que forman un vector de dimensión determinada. La calidad de la extracción de las características se basa en la pertinencia y la robustez de las informaciones extraídas de las imágenes, en el sentido de los conceptos visuales que se busca reconocer. La robustez se considera principalmente con relación a unas variaciones de las imágenes en términos de punto de vista, de luminosidad, de rotación, traslación y de zoom.

Unas técnicas de extracción de características conocidas implican una etapa de extracción de descriptores locales de la imagen, para reconstruir una firma final, mediante un enfoque denominado de “saco de palabras visuales”, comúnmente designada por la sigla BOV correspondiente a la terminología inglesa “Bag Of Visual terms”, o “Bag Of Visterms”. La figura 2 descrita en detalle a continuación ilustra el principio de funcionamiento de una etapa de extracción de características. Típicamente, se extraen de la imagen considerada uno o una pluralidad de descriptores locales, a partir de píxeles o de parches densos en la imagen, o más generalmente de sitios en la imagen. En otros términos, se asocian unos descriptores locales a otros tantos parches, que pueden definirse principalmente por su localización o situación, por ejemplo por unas coordenadas (x, y) en una referencia cartesiana en la que se define igualmente el dominio de la imagen considerada, pudiendo limitarse un parche a un píxel, o consistir en un bloque de una pluralidad de píxeles. En lo que sigue se hará referencia a la localización de descriptores locales para designar la localización de los sitios a los que están asociados, y de una manera similar se podrá cualificar unos descriptores locales de vecinos espacialmente en una imagen, cuando los sitios a los que están asociados son espacialmente vecinos en la imagen. Se registran entonces los descriptores locales durante una etapa de codificación o “coding” en un espacio de las características o “feature space” según la terminología inglesa, en función de un diccionario de referencia, comúnmente designado por el término inglés “codebook”. Se agregan entonces los vectores decodificados, durante una etapa de agregación o de “pooling” en un único vector que forma la firma. Estas etapas pueden repetirse para varias partes de la imagen considerada, y posteriormente concatenarse las firmas, por ejemplo de acuerdo con un esquema de pirámide espacial, designada por el acrónimo SPM que designa la terminología inglesa “Spatial Pyramid Matching”, que consiste en recortar la imagen considerada en sub-bloques, por ejemplo unos cuadrados de 2x2 o 4x4 bloques, o unos rectángulos de 1x3 bloques, etc., en determinar la firma para cada sub-bloque y posteriormente concatenar todas las firmas determinadas ponderándolas por un factor dependiente de la escala de los recortes en sub-bloques. Se describe una técnica de tipo SPM por ejemplo en la publicación de S. Lazebnik, C. Schmid y J. Ponce “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories” en CVPR, 2006.

Diferentes técnicas conocidas forman la base de las etapas de agregación y de codificación antes citadas. La etapa de codificación puede basarse principalmente en una técnica denominada de “codificación dura”, comúnmente designada según la terminología inglesa “Hard Coding” o según el acrónimo correspondiente HC. Se describen unas técnicas de codificación dura por ejemplo en la publicación de S. Lazebnik, C. Schmid y J. Ponce “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories” anteriormente citada, o también en la publicación de J. Sivic y A. Zisserman “Video google: a text retrieval approach to object matching in videos” en ICCV, 2003. Según una técnica de codificación dura, se recodifica un descriptor local en un vector que incluye un único “1” en la dimensión correspondiente al índice de su vecino más próximo en el diccionario de referencia, y una pluralidad

de “0” en otros lugares. Asociado con una etapa de agregación basada en la determinación de un promedio, una etapa de codificación mediante codificación dura conduce de ese modo a la realización de un histograma de aparición de palabras visuales del diccionario de referencia más presentes, siendo considerada una palabra visual del diccionario de referencia como presente cuando es la más próxima a un descriptor local de la imagen considerada.

La etapa de codificación puede basarse igualmente en una técnica denominada de “codificación suave”, comúnmente designada según la terminología inglesa “Soft Coding” o según el acrónimo correspondiente SC. Se describe una técnica de codificación suave principalmente en la publicación de J. Van Gemert, C. Veenman, A. Smeulders y J. Geusebroek “Visual word ambiguity” - PAMI, 2009. Según la técnica de codificación suave, se recodifica un descriptor local según su similitud con cada una de las palabras visuales del diccionario de referencia. La similitud se calcula por ejemplo como una función decreciente de la distancia, típicamente una exponencial de la inversa de la distancia.

La etapa de codificación puede basarse igualmente en una técnica denominada de “codificación lineal localmente restringida”, comúnmente designada según la terminología inglesa “Locally constrained Linear Coding” o según el acrónimo correspondiente LLC. Se describen principalmente unas técnicas de tipo LLC en la publicación de S. Gao, I. Tsang, L. Chia y P. Zhao, “Local features are not lonely - Laplacian sparse coding for image classification” en CVPR, 2011, en la publicación de L.Liu, L. Wang y X. Liu “In defense of soft-assignment coding” en CVPR, 2011, o también en la publicación de J. Yang, K. Yu, Y. Gong y T. Huang “Linear spatial pyramid matching using sparse coding for image classification” in CVPR, 2009. El principio de esta técnica consiste en restringir la codificación de tipo suave a los vecinos más próximos de los descriptores de las características, por ejemplo de 5 a 20 vecinos más próximos del diccionario de referencia. De esa forma, puede reducirse de manera significativa el ruido de codificación.

La etapa de codificación puede basarse igualmente en una técnica denominada de “codificación saliente localmente restringida”, comúnmente designada según la terminología inglesa “Locally constrained Salient Coding” en la que cada descriptor no está codificado más que por su vecino más próximo asociándole una respuesta, denominada de pertinencia “saliency”, que depende de las distancias relativas de los vecinos más próximos al descriptor. En otros términos, cuanto más reducida es la distancia del vecino más próximo al descriptor con relación a la distancia de los otros vecinos más próximos a este mismo descriptor, más grande es la pertinencia. Una técnica de tipo “saliency coding” se describe principalmente la publicación de Y. Huang, K. Huang, Y. Yu, y T. Tan “Salient coding for image classification” en CVPR, 2011.

Un inconveniente de los procedimientos conocidos de reconocimiento del contexto visual, que implementan unas etapas de extracción de características conocidas, está vinculado al hecho de que aunque los descriptores locales estén próximos en una imagen, y tengan en consecuencia probablemente unas características semejantes, su codificación en el espacio de las características por la etapa de codificación puede conducir a una gran variabilidad. Resulta de ahí que grandes zonas relativamente homogéneas de la imagen pueden codificarse de modo muy diferente.

Un objeto de la presente invención es paliar al menos los inconvenientes antes citados, proponiendo un procedimiento de reconocimiento de un contexto visual de una imagen basado en la técnica de los sacos de palabras, que asegura un alisado de los códigos de los descriptores locales en función de su proximidad espacial mutua y de su similitud.

Se propone principalmente con este fin por la presente invención incluir una nueva etapa de codificación en un procedimiento de reconocimiento, o bien unos medios apropiados para implementar en un dispositivo de reconocimiento del contexto visual, que implemente por ejemplo una codificación que tenga en cuenta el contexto espacial de los descriptores locales, permitiendo elegir las dimensiones no nulas de los vectores recodificados y precediendo a una técnica de codificación en sí misma conocida.

Considerando los descriptores locales como distribuidos según una estructura predeterminada, por ejemplo una rejilla regular, la presente invención propone que la toma en consideración del contexto espacial de los descriptores locales se realice por medio de una función objetiva que tenga en cuenta a la vez la restricción de la situación en el dominio espacial, es decir que dos descriptores locales próximos en la rejilla son candidatos para ser codificados idénticamente según la similitud entre ellos, o incluso según la similitud de las palabras visuales del diccionario de referencia a las que se asocian unos de descriptores locales vecinos, y una codificación local en el espacio de características.

Otra ventaja de la invención es que puede basarse indiferentemente en una de las técnicas de codificación en sí mismas conocidas, de las que se han presentado anteriormente unos ejemplos.

Con este fin, la invención tiene por objeto un procedimiento de reconocimiento de un contexto visual según la reivindicación 1. Según un modo de realización de la invención, dicho segundo término puede ponderarse mediante un parámetro de regularización global.

Según un modo de realización de la invención, dicha función objetiva puede definirse por la relación (1) presentada a

continuación en el presente documento.

Según un modo de realización de la invención, la minimización de la función objetiva puede operarse mediante un algoritmo de optimización.

Según un modo de realización de la invención, dicho algoritmo de optimización puede basarse en un enfoque de tipo por corte de grafos.

Un procedimiento de reconocimiento según uno de los modos de realización de la invención puede aplicarse al reconocimiento de un contexto visual en un flujo de vídeo formado por una pluralidad de imágenes, formando al menos una imagen entre la pluralidad de imágenes el objeto de un procedimiento de reconocimiento de un contexto visual en una imagen según uno de los modos de realización de la invención.

La presente invención tiene igualmente por objeto un sistema de clasificación supervisado que comprende al menos una fase de aprendizaje realizada fuera de línea, y una fase de ensayo realizada en línea, comprendiendo la fase de aprendizaje y la fase de ensayo cada una al menos una etapa de extracción de características tal como se define en un procedimiento de reconocimiento según uno cualquiera de los modos de realización de la invención.

La presente invención tiene igualmente por objeto un dispositivo de reconocimiento de un contexto visual en una imagen que comprende unos medios adaptados para la implementación de un procedimiento de reconocimiento según uno cualquiera de los modos de realización de la invención.

La presente invención tiene igualmente por objeto un programa informático que incluye unas instrucciones para implementar un procedimiento según uno de los modos de realización de la invención.

Surgirán otras características y ventajas de la invención con la lectura de la descripción, dada a título de ejemplo, realizada con relación a unos dibujos adjuntos que representan:

- la figura 1, un diagrama que ilustra la técnica de clasificación supervisada;
- la figura 2, un diagrama que ilustra el principio de la técnica de extracción de características de la imagen según el enfoque BOV;
- las figuras 3a y 3b, un esquema que ilustra de manera sinóptica el principio de asociación de los descriptores locales a unas palabras visuales de un diccionario de referencia, según un ejemplo de realización de la presente invención;
- la figura 4, un esquema que ilustra de manera sinóptica un dispositivo de reconocimiento del contexto visual según un ejemplo de realización de la presente invención.

La figura 1 presenta un diagrama que ilustra la técnica de clasificación supervisada, presentada anteriormente.

Un sistema de clasificación supervisada comprende principalmente una fase 11 de aprendizaje realizada fuera de línea y una fase 13 de ensayo realizada en línea.

Cada una de la fase 11 de aprendizaje y la fase 13 de ensayo comprende una etapa 111, 131 de extracción de características que permite describir una imagen mediante un vector de dimensión determinada. La etapa 11 de aprendizaje consiste en extraer las características sobre un gran número de imágenes 113 de aprendizaje; una serie de firmas y las etiquetas 112 correspondientes alimentan un módulo 115 de aprendizaje, que produce entonces un modelo 135.

La etapa 13 de ensayo consiste en describir, por medio de la etapa 131 de extracción de características, una imagen denominada imagen 133 de ensayo mediante un vector de la misma naturaleza que durante la fase 11 de aprendizaje. Este vector se aplica a la entrada del modelo 135 antes citado. El modelo 135 produce en su salida una predicción 137 de la etiqueta de la imagen 133 de ensayo. La predicción asocia la etiqueta (o las etiquetas) a la (o a las) más pertinente(s) a la imagen de ensayo entre el conjunto de las etiquetas posibles.

Esta pertenencia se calcula por medio de una función de decisión asociada al modelo de aprendizaje aprendido sobre la base del aprendizaje dependiente del algoritmo de aprendizaje utilizado.

La etiqueta de una imagen indica su grado de pertenencia a cada uno de los conceptos visuales considerados. Por ejemplo, si se consideran tres clases, por ejemplo las frases “playa”, “ciudad” y “montaña”, la etiqueta es un vector de tres dimensiones en el que cada componente es un número real. Por ejemplo, cada componente puede ser un número real comprendido entre 0 si la imagen no contiene el concepto, y 1 si la imagen contiene el concepto de navegación.

La técnica de aprendizaje puede basarse en una técnica en sí misma conocida, tal como la técnica de los separadores de amplio margen, comúnmente designada por la sigla SVM correspondiente a la terminología inglesa “Support Vector Machine”, en una técnica denominada de “boosting”, o también en una técnica del tipo designado por las siglas MKL correspondientes a la terminología inglesa “Multiple Kernel Learning”.

La presente invención se refiere más particularmente a la etapa 111, 131 de extracción de características, y se

integra por tanto en un procedimiento de reconocimiento del contexto visual que comprende además principalmente una etapa de adquisición de una imagen o de un flujo de vídeo, una etapa eventual de recorte de la imagen en sub-imágenes, la predicción de la etiqueta para la imagen de ensayo o una sub-imagen.

La figura 2 presenta un diagrama que ilustra el principio de la técnica de extracción de características de una imagen.

Una etapa 211 de extracción de características de una imagen I comprende una etapa de extracción de una pluralidad n de descriptores 2111 locales para una pluralidad de parches de la imagen I. Los descriptores 2111 locales son unos vectores de un tamaño d. La extracción de los descriptores locales puede realizarse según diferentes técnicas en sí mismas conocidas. Por ejemplo, según la transformada de características visuales invariantes con la escala, comúnmente designada por la terminología inglesa “Scale-Invariant Feature Transform” o por las siglas “SIFT” correspondientes, o bien según la técnica de los histogramas de gradientes, comúnmente designada por las siglas HOG que corresponde a la terminología inglesa “Histograms Of Gradients”, o también según la técnica de las características robustas aceleradas, comúnmente designada según las siglas SURF procedentes de la terminología inglesa “Speeded Up Robust Features”.

Un descriptor local implica una transformación matemática o estadística a partir de al menos un píxel de una imagen. Se encontrarán ejemplos, bien conocidos por el experto en la materia, de descriptores locales en los artículos Mikolajczyk, K.; Schmid, C., “A performance evaluation of local descriptors”, Pattern Analysis and Machine Intelligence, IEEE Transactions en, vol. 27, n.º 10, págs.1615, 1630, octubre de 2005. doi: 10.1109/TPAMI.2005.188 o en los párrafos “localisation des caractéristiques” y “caractéristiques locales” del artículo http://fr.wikipedia.org/wiki/Extraction de caract%C3%A9ristique en vision par ordinateur.

En particular, un descriptor local no es una zona simple de la imagen obtenida por segmentación.

Los n de descriptores 2111 locales se registran en el espacio de las características durante una etapa 2112 de codificación para formar una matriz 2113 de codificación, en función de un diccionario 2101 de referencia y de al menos un criterio de similitud, siendo construido el diccionario 2101 de referencia durante una fase 210 de elaboración del diccionario de referencia descrita en detalle en el presente documento a continuación. El diccionario 2101 de referencia comprende una pluralidad K de palabras visuales de tamaño d, y su tamaño es por tanto igual a dxK (d líneas y K columnas). Los vectores registrados que forman la matriz 2113 de codificación se agregan entonces durante una etapa 2114 de agregación o de “pooling” para formar un vector único o firma 2115 de longitud K. El proceso definido por las etapas 2112 de codificación y 2114 de agregación se reitera eventualmente para varias partes de la imagen I o sub-imágenes, y posteriormente las firmas concatenadas durante una etapa 2116 de concatenación para formar una firma 2117 concatenada según un esquema de tipo pirámide espacial.

El diccionario 2101 de referencia se establece durante la fase 210 de elaboración a partir de una gran colección de descriptores locales procedentes de una pluralidad de imágenes 1 a N. Para un número n_i de descriptores locales por imagen 1 a N, se construye una matriz 2100 de Nxn_i descriptores locales de tamaño d. El diccionario 2101 de referencia puede construirse por ejemplo a partir de la matriz 2100 empleando una clasificación supervisada, designada por el término inglés “clustering”, por ejemplo según el algoritmo de los K-promedios o “K-means” según la terminología inglesa, permitiendo particionar los Nxni descriptores locales en una pluralidad k de conjuntos con el fin de minimizar el error de reconstrucción de los descriptores por el centroide en el interior de cada partición. Es posible igualmente recurrir a otros procedimientos de aprendizaje del diccionario de referencia, tales como por ejemplo la extracción aleatoria de descriptores locales o la codificación parsimoniosa.

La etapa 2114 de agregación puede consistir por ejemplo en realizar un promedio, o bien la agregación puede efectuarse según el máximo por columna de la matriz 2113 de codificación de dimensión nxK.

La etapa 2112 de codificación puede basarse en parte en una técnica de codificación tal como la técnica de codificación dura, o bien de codificación suave, o bien de codificación lineal localmente restringida o LLC, o también de codificación saliente LSC, técnicas estas conocidas en sí mismas que han sido introducidas anteriormente.

A partir de una hipótesis Markoviana para una imagen I dada, es posible afirmar que los parches vecinos en una zona unida o lisa de la imagen deben codificarse en unas bases similares a partir del diccionario 2101 de referencia. Por el contrario, para unos parches vecinos en una zona de la imagen I que presente discontinuidades, las bases correspondientes en el diccionario 2101 de referencia pueden variar, en función de las informaciones de los descriptores locales correspondientes.

Sea $P = \{1; 2; \ldots; N_I\}$ el conjunto de los índices de los píxeles o de los parches densos, o más generalmente de los sitios, en la imagen I. Se extrae un conjunto de descriptores locales $X = \{x_p; x_p \in \Re^d; p \in P\}$ de la imagen I para todos los sitios considerados.

El diccionario 2101 de referencia puede denotarse por $B = \{b_i; b_i \in \Re^K; i \in \aleph\}$. Se considera que cada descriptor local se atribuye a un subconjunto de vectores locales de referencia o “palabras visuales” que pertenecen al

diccionario 2101 de referencia.

Por razones de simplicidad de la notación, es posible indicar como $Y = \{y_p ; y_p \in \aleph^m ; p \in \mathrm{P}\}$ al conjunto de los índices de los vectores locales de referencia a los que se asocian unos descriptores x_p locales.

En el caso ejemplificado en el que la etapa de codificación se basa en parte en una codificación lineal localmente restringida o LLC, cada vector y_p representa los índices de las m palabras visuales más próximas a x_p.

El conjunto de los vectores locales de referencia relativos a los índices en el vector y_p puede indicarse por $\hat{B}_p = \{\hat{b}_{p,i} ; i \in \{1; \ldots; m\}\}$ y es posible definir el conjunto $\hat{B}$ de los conjuntos $\hat{B}_p$, es decir: $\hat{B} = \{ \hat{B}_p; p \in P\}$.

De ese modo, según una hipótesis denominada de situación, cada descriptor local puede asociarse a unas palabras visuales comprendidas entre el conjunto de las k palabras visuales más próximas del diccionario 2101 de referencia, siendo k superior a m. El número k restringe la búsqueda de los m vectores óptimos del diccionario de referencia a la vecindad local en el espacio de los descriptores locales con el fin de tener en cuenta la hipótesis de la situación en el espacio de los descriptores locales y acelerar el proceso de búsqueda de la base óptima para un descriptor local dado.

En otros términos, k designa el número máximo de los vectores de la base cuyos m vectores se considerarán como los óptimos para codificar un descriptor local dado.

El valor de k puede elegirse suficientemente grande de manera que considere una vecindad relativamente grande en el espacio de las características.

El conjunto de los índices de las k palabras visuales más próximas al descriptor local x_p puede indicarse por: $L_p = \{l_p^1 ; l_p^2 ; \ldots ; l_p^k\}$.

Este conjunto puede designarse conjunto de las etiquetas que un sitio p puede llevar.

El problema que la presente invención propone resolver puede asimilarse a un problema de etiquetado que consiste en determinar la referencia o “asociación” objetiva para cada descriptor local, entre las k palabras visuales más próximas, según la hipótesis espacial contextual.

Es posible con este fin introducir una función objetiva, o “función de energía” E(Y) que permita formalizar este problema, presentándose como una suma de un primer término que expresa la asociación de un descriptor local dado a unas palabras visuales próximas del diccionario de referencia, y de un segundo término que expresa la asociación de palabras visuales del diccionario de referencia similares para unos descriptores locales próximos espacialmente en la imagen I, y que presentan entre ellos una similitud suficiente.

Ventajosamente un parámetro global de regularización permite ponderar el segundo término con relación al primer término, de manera que permite un ajuste del alisado permitido por la presente invención.

Las figuras 3a y 3b ilustran de manera sinóptica el principio de asociación de descriptores locales a unas palabras visuales del diccionario de referencia, según un ejemplo de realización de la presente invención.

Con referencia a la figura 3a, pueden extraerse tres descriptores locales x_p, x_q y x_r de tres sitios de una imagen I. La imagen I puede subdividirse en una rejilla regular, en el ejemplo ilustrado por la figura. Los descriptores locales x_p, x_q y x_r son vecinos en el sentido de la rejilla regular. El primer descriptor local x_p presenta una primera tasa de similitud, igual a 0,9 en el ejemplo ilustrado por la figura, con el segundo descriptor local x_q, y una segunda tasa de similitud, igual a 0,7 en el ejemplo ilustrado por la figura, con el tercer descriptor local x_r.

Se ha de observar que el uso de una rejilla regular no constituye más que un ejemplo no limitativo de la presente invención. Pueden utilizarse otras técnicas de segmentación de la imagen. Por ejemplo los tratamientos previos pueden permitir elegir unos sitios dispuestos en unos puntos particulares de la imagen en función de diferentes criterios tales como unos gradientes de contraste, etc.

Ahora con referencia a la figura 3b, los descriptores locales x_p, x_q y x_r se asocian, durante la etapa de codificación, a unas palabras visuales del diccionario 2101 de referencia. El mismo diccionario de referencia se ilustra en la figura por dos referencias en A y B: la primera referencia A que ilustra un principio de asociación según una técnica de codificación conocida en la técnica, por ejemplo una de las técnicas de codificación conocidas introducidas anteriormente; la segunda referencia B ilustra un principio de asociación según un ejemplo de realización de la presente invención.

Según una de las técnicas de codificación conocidas en el estado de la técnica, los tres descriptores locales x_p, x_q y x_r se asocian cada uno a una pluralidad de vecinos más próximos entre las palabras visuales del diccionario 2101 de referencia, en número de tres en el ejemplo ilustrado por la figura. La asociación de un descriptor local a unas

palabras visuales del diccionario 2101 de referencia se realiza en función de criterios de similitud, es decir de distancia. Los tres descriptores locales x_p, x_q y x_r se asocian así a unas palabras visuales del diccionario de referencia sin consideración a la proximidad espacial de la imagen I de los sitios de los que proceden. De ese modo, unos descriptores locales vecinos espacialmente en la imagen y similares, tales como el primer y segundo descriptores locales x_p, x_q pueden verse asociados a unas palabras visuales diferentes del diccionario 2101 de referencia. En el ejemplo ilustrado por la figura, el primer descriptor local x_p y el segundo descriptor local x_q no comparten más que una única palabra visual del diccionario 2101 de referencia, a pesar de su vecindad espacial y su similitud.

Considerando ahora la segunda referencia B en la figura 3b, el principio de asociación según la presente invención asegura que unos descriptores locales procedentes de sitios espacialmente vecinos de la imagen I, y que presentan una tasa de similitud suficiente, se asocien a unas palabras visuales idénticas del diccionario 2101 de referencia. En el ejemplo ilustrado en la figura 3b, el primer descriptor local x_p y el segundo descriptor local x_q comparten así tres palabras visuales comunes del diccionario 2101 de referencia.

De ese modo, según la presente invención, la etapa 2112 de codificación puede elaborar la matriz 2113 de codificación mediante una asociación de cada descriptor local con una o una pluralidad de palabras visuales del diccionario 2101 de referencia en función de al menos un criterio de similitud, realizándose además la asociación en función de la similitud entre cada descriptor 2111 local y las palabras visuales del diccionario 2101 de referencia asociadas a al menos un descriptor 2111 local espacialmente próximo en el dominio de la imagen I de cada descriptor 2111 local considerado.

Es posible por ejemplo introducir una función de energía que sintetiza el principio de asociación presentado anteriormente, y que puede formularse por la relación siguiente:

$$E(Y) = \underbrace{\sum_{p \in P} f_{datos}\left(x_p, \hat{B}_p\right)}_{E_p(y_p)} + \beta \underbrace{\sum_{p \sim q} w_{p,q} f_{previo}\left(\hat{B}_p, \hat{B}_q\right)}_{E_{p,q}(y_p, y_q)} \qquad (1),$$

en la que:

- el término $f_{datos}(x_p, \hat{B}_p)$ representa la distancia total entre un descriptor local x_p y una pluralidad m elegida de palabras visuales de referencia, es decir $f_{datos}\left(x_p, \hat{B}_p\right) = \sum_{i=1}^{m} \left\| x_p - \hat{b}_{p,i} \right\|_2^2$;
- p~q representan los índices de dos descriptores locales espacialmente vecinos según un esquema de vecindad determinado, pudiendo definirse el sistema de vecindad, por ejemplo, mediante una rejilla de los 4 vecinos más próximos;
- el término $f_{previo}(\hat{B}_p, \hat{B}_q)$ representa una suma de las distancias entre las palabras visuales asociadas a los descriptores locales vecinos x_p y x_q, es decir $f_{previo}\left(\hat{B}_p, \hat{B}_q\right) = \sum_{i=1}^{m} \left\| \hat{b}_{p,i} - \hat{b}_{q,i} \right\|$;
- el término $w_{p,q}$ representa un parámetro local denominado de regularización que corresponde al nivel de similitud entre los parches locales x_p y x_q. De ese modo, cuanto más elevado sea el nivel de similitud que tienen los parches locales, más regularizada es la operación de selección de la referencia. Pueden emplearse diferentes procedimientos de medición de la similitud entre unas características locales. Es posible por ejemplo recurrir a una técnica de núcleo de intersección de histogramas, conocida según la denominación inglesa “histogram intersection kernel”, particularmente eficaz cuando las características locales se basan en unos histogramas. Esta técnica consiste en medir la similitud entre dos vectores de histogramas sumando los valores mínimos entre los dos vectores en cada dimensión. Un núcleo de intersección de histogramas puede indicarse por K(., .). Los parámetros de regularización local pueden establecerse según la relación siguiente:

$$w_{p,q} = \begin{cases} K\left(x_p, x_q\right) & \text{si} \quad K\left(x_p, x_q\right) \geq T, \\ 0 & \text{si no} \end{cases} \qquad (2).$$

Esta forma binaria de los parámetros de regularización permite una regularización solamente sobre unos parches vecinos similares, por debajo de un valor de umbral de similitud indicado por T. Además, permite reducir la sensibilidad del modelo al parámetro de regularización global β. El parámetro de regularización global permite ajustar la influencia del alisado.

- La relación (1) anterior comprende dos términos principales indicados por E_{datos} y E_{previo}. El primer término E_{datos} es un término denominado de probabilidad, que penaliza la asociación de las palabras visuales alejadas de los descriptores locales a estos últimos, mientras que el segundo término E_{previo} es un término denominado “a priori” que penaliza la asociación de diferentes palabras visuales a los parches vecinos similares.

Minimizando la función objetiva definida por la relación (1) anterior, es posible entonces obtener una configuración óptima de asociación de descriptores locales a unas palabras visuales del diccionario de referencia. Esta minimización se expresa por la relación siguiente:

$$\tilde{Y} = \arg\min E\left(Y \middle| X, \hat{B}, W\right) \qquad (3).$$

Las asociaciones retenidas constituyen la configuración óptima que asocia los vectores de referencia óptimos a los descriptores locales de la imagen.

Se ha de observar que unos casos particulares de la función objetiva así definida corresponden a unos casos en sí mismos conocidos de técnicas de codificación, principalmente:

- en el caso particular en el que el parámetro de regularización global β es nulo y en el que m es igual a 1, la técnica de asociación utilizada es entonces la técnica de codificación dura o de codificación saliente;
- en el caso particular en el que el parámetro de regularización global β es nulo y en el que m es superior a 1, la técnica de asociación es entonces una técnica de tipo LLC.

La función objetivo así definida es no convexa para las formas generales de las funciones de distancia f_{previo} y f_{datos} introducidas anteriormente. Su minimización puede realizarse de manera apropiada, por ejemplo a partir de técnicas derivadas de técnicas en sí mismas conocidas de optimización rápida habitualmente dedicadas a unas funciones de energía de tipo campos de Markov aleatorios multi-etiquetas de orden 1, es decir que hacen intervenir al máximo unos pares de interacciones entre sitios vecinos, comúnmente designados “pairwise multi-label MRF energies”. Más particularmente, unos algoritmos basados en un enfoque por cortes de grafos, más comúnmente designados “graph-cuts” según la terminología inglesa, que permite la resolución de numerosos problemas de etiquetado en el dominio de la visión por ordenador.

Por ejemplo, el algoritmo designado bajo la denominación “alfa-expansión” es un algoritmo basado en un enfoque por cortes de grafos particularmente apropiado para la presente aplicación. Este algoritmo es un algoritmo iterativo basado en las transiciones binarias de la configuración deseada. En una iteración dada i, cada sitio puede conservar su etiqueta actual o bien cambiar de etiqueta, cambiando a una nueva etiqueta $\alpha^{(i)} \in L$, siendo L un conjunto discreto de etiquetas. Este “movimiento binario” se realiza de manera óptima estableciendo un grafo apropiado en el que se calculan una sección mínima/ondulación máxima. Las particiones binarias sufren varios movimientos —o cambios— de particiones hasta que se llega a una convergencia en un óptimo local de energía. De tal manera, es posible obtener unos óptimos locales de funciones de energía no convexas, incluso un óptimo global de una función de energía no convexa, con una buena precisión. Además, un algoritmo de ese tipo presenta no solamente la ventaja de ser eficaz, sino igualmente la de poder aplicarse a diversos problemas de etiquetado, incluso cuando las etiquetas no están ordenadas.

Ventajosamente, se propone por la presente invención resolver el problema de optimización planteado por medio de un novedoso algoritmo de optimización por aproximación que extiende el principio de la alfa-expansión, en dos direcciones: por una parte de manera que se aplique a unas etiquetas vectoriales, es decir que se asocia a cada sitio un conjunto finito de etiquetas; por otra parte para obligar a cada sitio a endosar una etiqueta entre un subconjunto apropiado de etiquetas.

Este nuevo algoritmo puede designarse α_{knn}-expansión. El principio sobre el que se basa este algoritmo de optimización consiste en realizar, en cada iteración, un movimiento de expansión binario hacia $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_m\} \in \aleph^m$, solamente para un subconjunto de sitios $S_\alpha \subset P$, estando α comprendido entre el conjunto de los índices de las k palabras visuales más próximas, es decir: $S_\alpha = \{p \in P$, tal que $\alpha \subset L_p\}$. Estos sitios se designan en el presente documento a continuación sitios activos.

Con cada iteración del algoritmo de optimización, se realiza un movimiento binario global de todos los sitios activos. Los movimientos binarios se integran para un número de etiquetas vectoriales $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ en el seno de un ciclo. Si la función de energía continúa decreciendo, se inicia un nuevo ciclo, hasta que el algoritmo converge hacia un óptimo.

Se definen a continuación las etapas del algoritmo:

- Los datos de entrada son: $Y^{(0)}$, la configuración inicial, por ejemplo correspondiente a una asociación sobre la base de los m vecinos más próximos tal como se utiliza en un enfoque basado en una técnica de codificación de tipo LLC; $C_{máx}$, un valor correspondiente a un número máximo de ciclos; W la matriz de los parámetros de regularización $w_{p,q}$ introducida anteriormente; $\hat{B}$, conjunto introducido anteriormente; X, igualmente introducido anteriormente.
- Para cada ciclo $c \leq C_{máx}$, proceder a las etapas siguientes:
    - Seleccionar n vectores de etiquetas $\{\alpha_i\}_{i=1}^{n}$ entre el conjunto de los índices de los k vecinos más próximos

a los descriptores locales;

○ Para cada iteración i ≤ n, realizar las sub-etapas siguientes:

▪ Realizar un movimiento de expansión binaria óptimo hacia α_i: $Y^{(i)} = \arg\min_{Y} E\left(Y \middle| X, \hat{B}, W\right)$, tal que:

$$y_p^{(i)} \in \left\{y_p^{(i-1)}, \alpha_i\right\}, \forall p \in P \ ;$$

▪ Atribuir a $\tilde{Y}$ el valor Y(i);

○ Si $E(\tilde{Y}) < E(Y^{(c-1)})$, entonces atribuir a Y(c) el valor $\tilde{Y}$, si no devolver el valor $\tilde{Y}$, dado de salida del algoritmo.

Con el fin de realizar un movimiento de expansión binaria óptimo en la sub-etapa presentada anteriormente, puede elaborarse un grafo dirigido $G_\alpha = (A_\alpha, \varepsilon_\alpha)$, designando A_α el conjunto de los nodos relativos a los sitios activos, designando ε_α el conjunto de los bordes orientados que conectan unos nodos vecinos.

Se añaden dos nodos auxiliares s y t para el cálculo de la ondulación máxima.

Cuando se ha elaborado de ese modo el grafo dirigido, puede calcularse la ondulación máxima en tiempos polinomiales gracias a la propiedad de la sub-modularidad de la función objetiva propuesta, que es una propiedad matemática de la función a minimizar necesaria para su optimización en tiempos polinomiales. En efecto, para los movimientos de expansión binaria, la restricción de sub-modularidad en la función de energía se verifica para todos los términos de datos y los términos métricos a priori. Este es el caso para la función objetiva según la presente invención formulada por la relación (1) anterior, en la que se utiliza una métrica como a priori para el cálculo de la distancia entre las referencias asociadas a los parches locales vecinos.

La minimización de esta función objetiva a partir del algoritmo propuesto es rápida, puesto que los movimientos de expansión binaria están restringidos a los sitios activos en cada iteración; además la utilización de un algoritmo de ondulación máxima en tiempos polinomiales tal como se ha descrito anteriormente presenta la ventaja de estar bien adaptada a unos grafos de estructura de rejilla.

Una vez se ha elegido la base óptima en términos de situación en el espacio de las características y en el dominio espacial de la imagen, es decir una vez que se han elegido los vectores de base de cada descriptor de manera óptima, según el modelo establecido anteriormente en lo que se refiere a las etapas de codificación y de agregación antes citadas, puede realizarse la asociación de una codificación o “respuesta” a cada uno de los vectores de base por medio de diversas técnicas.

Por ejemplo, unas técnicas de codificación dura proporcionan unas respuestas denominadas duras, o bien unas técnicas de codificación saliente proporcionan unas respuestas salientes, o bien unas técnicas de codificación suave proporcionan unas respuestas blandas que pueden utilizarse, o bien resolviendo el sistema lineal o bien calculando las probabilidades a posteriori de una característica local que pertenece al conjunto de las palabras visuales elegidas.

Para cada imagen, se agregan los códigos finales para producir un vector único que forma una firma. Por ejemplo, puede utilizarse una técnica de agregación por máximo, comúnmente designada por la terminología inglesa “max-pooling”, presentando esta técnica la ventaja de producir mejores resultados que una técnica de agregación por promedio, mientras es más rápida desde un punto de vista del tiempo de cálculo.

La figura 4 presenta un esquema que ilustra de manera sinóptica un dispositivo de reconocimiento del contexto visual según un ejemplo de realización de la presente invención.

Un dispositivo de reconocimiento del contexto visual de una imagen puede implementarse mediante unos medios dedicados de cálculo, o bien a través de unas instrucciones de software ejecutadas por un microprocesador unido a una memoria de datos. Por razones de claridad de la exposición, el ejemplo ilustrado por la figura 4 describe de manera no limitativa el dispositivo de reconocimiento en términos de módulos de software, considerándose que ciertos módulos descritos pueden subdividirse en varios módulos, o bien reagruparse.

El dispositivo 40 de reconocimiento recibe en la entrada una imagen I digitalizada, por ejemplo introducida por unos medios de introducción de datos dispuestos anteriormente, no representados en la figura. Un microprocesador 400 unido a una memoria 402 de datos permite la implementación de módulos de software cuyas instrucciones de programa se almacenan en la memoria 402 de datos o en una memoria dedicada. Las imágenes, los descriptores pueden almacenarse en una memoria 404 que forma una base de datos.

El dispositivo de reconocimiento del contexto visual puede configurarse para implementar un procedimiento de reconocimiento según uno de los modos de realización descritos

La implementación de un procedimiento de reconocimiento puede realizarse por medio de un programa informático que incluye unas instrucciones previstas con este fin.

Se ha de observar que la presente invención puede aplicarse igualmente al reconocimiento del contexto visual en un flujo de vídeo, pudiendo aplicarse un procedimiento según uno de los modos de realización descritos a las imágenes extraídas del flujo de vídeo.

Se ha de observar igualmente que la presente invención puede afectar a todas las aplicaciones que incluyen el reconocimiento del contexto visual. En particular, un procedimiento o un dispositivo según la presente invención puede permitir:

- facilitar la instalación de una red de cámaras de videovigilancia a gran escala;
- reconocer automáticamente el contexto de un vehículo con el fin de extraer las consecuencias para un procesamiento posterior, por ejemplo una adaptación de la velocidad cuando el vehículo entra en una zona urbana;
- un robot compañero, o un robot de asistencia a personas a domicilio, de reconocimiento de la estancia en la que se encuentra;
- reconocer un objeto fabricado, un logotipo, un lugar o un ambiente dado en un vídeo o un flujo televisual.

REIVINDICACIONES

1. Procedimiento de reconocimiento del contexto visual en una imagen (I) que se inscribe en un dominio, comprendiendo al menos una etapa (211) de extracción de características de la imagen (I) y que comprende al menos:

- una etapa de extracción de una pluralidad de descriptores (2111) locales para una pluralidad de sitios de la imagen (I),
- una etapa (2112) de codificación de la pluralidad de descriptores (2111) locales elaborando una matriz (2113) de codificación mediante una asociación de cada descriptor local con una o una pluralidad de palabras visuales de un diccionario (2101) de referencia en función de al menos un criterio de similitud,
- una etapa (2114) de agregación que forma una firma única a partir de la matriz (2113) de codificación,

estando el procedimiento **caracterizado porque** la asociación realizada durante la etapa (2112) de codificación se realiza además por medio de la minimización de la función objetivo definida por una suma de un primer término que expresa la suma para cada descriptor (x_p) local dado de la distancia total de dicho descriptor local a unas palabras visuales del diccionario de referencia asociadas a dicho descriptor local y un segundo término que expresa para unos descriptores locales vecinos en la imagen y que presentan un nivel de similitud suficiente, la distancia entre palabras visuales del diccionario de referencia asociadas a estos descriptores locales.

2. Procedimiento de reconocimiento según la reivindicación 1, **caracterizado porque** dicho segundo término está ponderado por un parámetro (β) de regularización global.

3. Procedimiento de reconocimiento según la reivindicación 2, **caracterizado porque** dicha función objetivo se define por la relación siguiente: $E(Y) = \underbrace{\sum_{p \in P} f_{datos}\left(x_p, \hat{B}_p\right)}_{E_p(y_p)} + \beta \underbrace{\sum_{p \sim q} w_{p,q} f_{previo}\left(\hat{B}_p, \hat{B}_q\right)}_{E_{p,q}(y_p, y_q)}$, en la que el término $f_{datos}(x_p, \hat{B}_p)$ representa la distancia total entre un descriptor local (x_p) y una pluralidad m de palabras visuales del diccionario (2101) de referencia, el término $f_{previo}(\hat{B}_p, \hat{B}_q)$ representa una suma de las distancias entre las palabras visuales asociadas a los descriptores locales vecinos xp y xq, el término p~q representa los índices de dos parches espacialmente vecinos según un sistema de vecindad determinado, $Y = \left\{ y_p ; y_p \in \aleph^m ; p \in P \right\}$ representa el conjunto de los índices de las palabras visuales de referencia a las que se asocian unos descriptores x_p locales, designando el término $\hat{B}_p = \{\hat{b}_{p,i} ; i \in \{1; \ldots; m\}\}$ el conjunto de los vectores locales de referencia relativos a los índices en el vector y_p.

4. Procedimiento de reconocimiento según la reivindicación 3, **caracterizado porque** la minimización de la función objetivo se opera mediante un algoritmo de optimización.

5. Procedimiento de reconocimiento según la reivindicación 4, **caracterizado porque** dicho algoritmo de optimización se basa en un enfoque del tipo por corte de grafos.

6. Procedimiento de reconocimiento de un contexto visual en un flujo de vídeo formado por una pluralidad de imágenes, **caracterizado porque** al menos una imagen entre la pluralidad de imágenes forma el objeto de un procedimiento de reconocimiento según una cualquiera de las reivindicaciones anteriores.

7. Sistema de clasificación supervisada que comprende al menos una fase (11) de aprendizaje realizada fuera de línea, y una fase (13) de ensayo realizada en línea, comprendiendo cada una de la fase (11) de aprendizaje y la fase (13) de ensayo al menos una etapa de extracción de características tal como se define en un procedimiento de reconocimiento según una cualquiera de las reivindicaciones anteriores.

8. Dispositivo de reconocimiento de un contexto visual en una imagen (I) que comprende unos medios adaptados para la implementación de un procedimiento de reconocimiento según una cualquiera de las reivindicaciones 1 a 6.

9. Programa informático que incluye unas instrucciones para la ejecución del procedimiento de reconocimiento según una cualquiera de las reivindicaciones 1 a 6, cuando se ejecuta el programa por un procesador.

10. Soporte de registro legible por un procesador en el que se registra un programa que incluye unas instrucciones para la ejecución del procedimiento de reconocimiento según una cualquiera de las reivindicaciones 1 a 6, cuando el programa se ejecuta por un procesador.

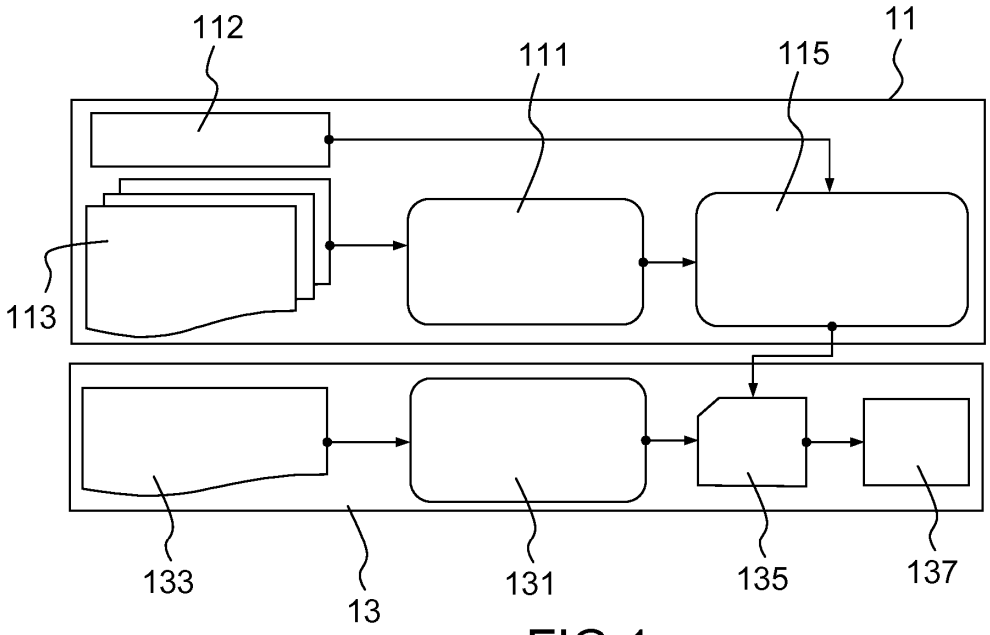
112
111
115
11
113
133
13
131
135
137

FIG.1

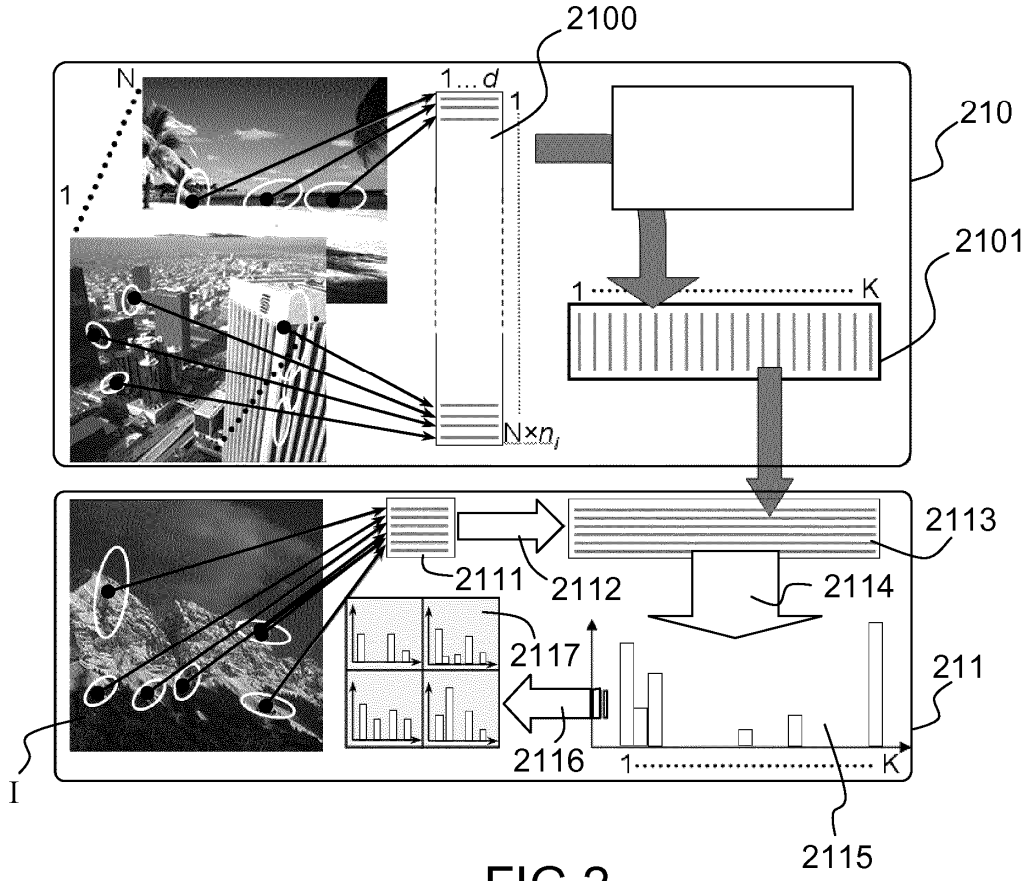
2100
N
1
1...d
1
210
2101
1
K
N×n_i
2113
2111
2112
2114
2117
211
2116
1
K
I
2115

FIG.2

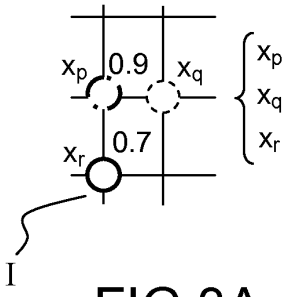

FIG.3A

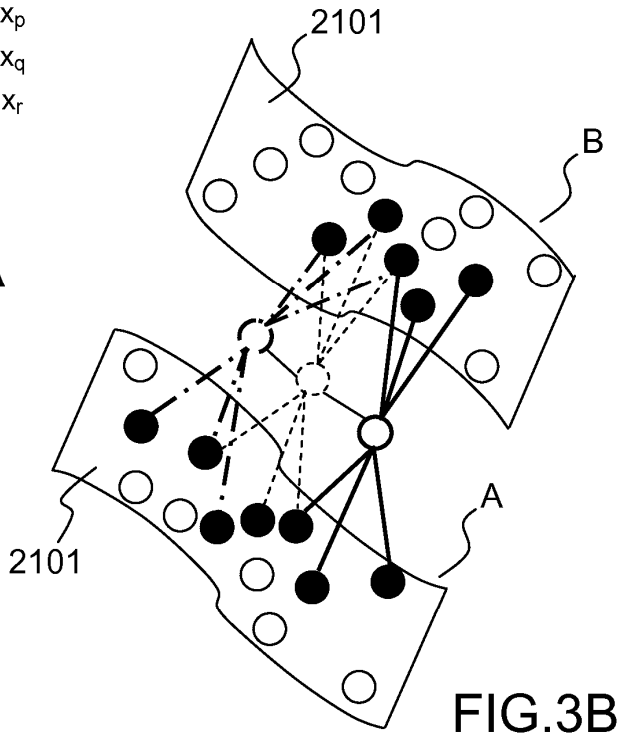

FIG.3B

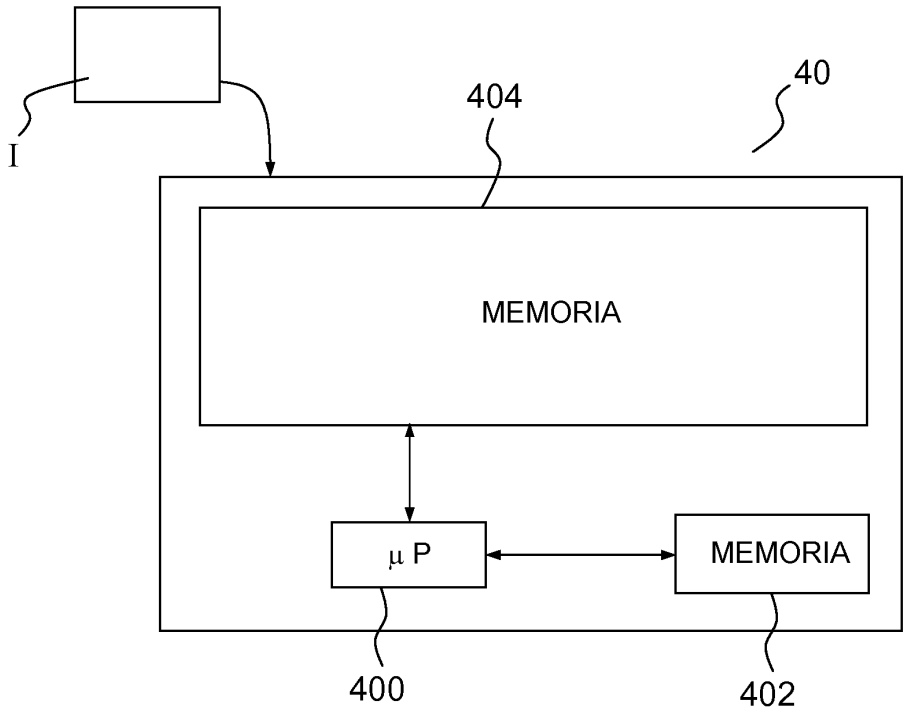

FIG.4