(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

COURBEVOIE

(11) **N° de publication :** **3 073 311**

(à n'utiliser que pour les commandes de reproduction)

(21) **N° d'enregistrement national :** **17 60541**

(51) Int Cl[8] : ***G 06 T 7/80*** *(2018.01), G 06 K 9/46, G 06 N 3/08*

(12) DEMANDE DE BREVET D'INVENTION A1

(22) **Date de dépôt :** 09.11.17.

(30) **Priorité :**

(43) **Date de mise à la disposition du public de la demande :** 10.05.19 Bulletin 19/19.

(56) **Liste des documents cités dans le rapport de recherche préliminaire :** *Se reporter à la fin du présent fascicule*

(60) **Références à d'autres documents nationaux apparentés :**

( ) **Demande(s) d'extension :**

(71) **Demandeur(s) :** *B<>COM* — FR.

(72) **Inventeur(s) :** DUONG NAM-DUONG et KACETE AMINE.

(73) **Titulaire(s) :** B<>COM.

(74) **Mandataire(s) :** AVOXA.

(54) PROCEDE D'ESTIMATION DE POSE D'UNE CAMERA DANS LE REFERENTIEL D'UNE SCENE TRIDIMENSIONNELLE, DISPOSITIF, SYSTEME DE REALITE AUGMENTEE ET PROGRAMME D'ORDINATEUR ASSOCIE.

(57) L'invention concerne un procédé d'estimation de pose d'une caméra dans un référentiel d'une scène tridimensionnelle, comprenant les étapes suivantes:

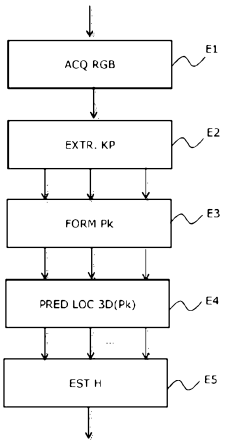
Obtention (E1) d'une image d'intensités de couleur de la scène capturée par la caméra, dite image courante;

Extraction (E2) d'une pluralité de points d'intérêt de l'image courante, un dit point étant invariant par transformation géométrique de l'image;

Formation (E3) d'une pluralité d'imagettes dans l'image d'intensités de couleur, une imagette comprenant un point d'intérêt de la pluralité extraite;

Prédiction (E4) des localisations 3D des points d'intérêt de la pluralité d'imagettes dans un référentiel de la scène, par application d'un système de prédiction automatique, ledit système ayant été entraîné à l'aide d'un ensemble d'apprentissage comprenant des imagettes issues d'une pluralité d'images de la scène acquises par la caméra depuis une pluralité de points de vue, une imagette étant associée à une position 2D de son point d'intérêt dans un référentiel de l'image et à une position 3D de son point d'intérêt dans le référentiel de la scène;

Estimation (E5) d'une pose de la caméra pour l'image courante, par mise en correspondance des positions 2D de la pluralité de points d'intérêt et de reprojections dans le référentiel de l'image courante des localisations 3D prédites.



FR 3 073 311 - A1

Procédé d'estimation de pose d'une caméra dans le référentiel d'une scène tridimensionnelle, dispositif, système de réalité augmentée et programme d'ordinateur associé

1. Domaine de l'invention

Le domaine de l'invention est celui de l'estimation de la pose d'une caméra dans le référentiel d'une scène tridimensionnelle (3D).

L'invention peut notamment, mais non exclusivement, trouver une application dans le domaine de la réalité augmentée, pour l'insertion d'un ou plusieurs objets virtuels ou réels dans l'image de la scène réelle vue par la caméra.

2. Présentation de l'art antérieur

On connait du document de Shotton *et al.*, intitulé « Scene Coordinate Regression Forests for Camera Relocalisation in RGB-D images », publié par la Conférence IEEE Conference on Computer Vision and Pattern Recognition, en 2013, une solution permettant de calculer la pose d'une caméra RGB-D (pour « Red Green Blue – Depth », en anglais) à l'aide d'un système de prédiction automatique, pour (« machine learning », en anglais), qui, après une phase d'apprentissage, est capable de prédire, à partir d'une image d'intensités de couleurs et une image de profondeur acquises par la caméra, un nuage de points correspondants dans un référentiel de la scène 3D. La pose de la caméra est ensuite estimée sur la base du nuage de points prédit.

Un avantage de cette solution est qu'elle prédit la pose de la caméra de façon complètement automatique sans aucune hypothèse géométrique.

Un premier inconvénient de cette solution est qu'elle impose de manipuler des nuages de points, ce qui la rend complexe à mettre en œuvre, notamment parce qu'elle nécessite des ressources importantes de calcul et de stockage.

Un deuxième inconvénient de cette méthode est qu'elle comprend, aussi bien dans la phase d'apprentissage que dans la phase de test, une transformation préalable de l'image destinée à prendre en compte des paramètres intrinsèques de la caméra, tels que des focales ou des centres de projection. Cette étape nécessite une calibration préalable de la caméra.

Un troisième inconvénient de cette solution est qu'elle nécessite une caméra de profondeur dans la phase de tests, ce qui exclut son utilisation pour une application de réalité virtuelle embarquée dans

des équipements terminaux de type mobile, comme par exemple un téléphone intelligent (pour « smartphone », en anglais) ou une tablette.

On connait aussi du document de Kendall et al., intitulé « PoseNet : A convolutional Network for Rel-Time 6-DOF Camera Relocalization », publié dans les Proceedings de la conférence IEEE International Conference on Computer Vision, pages 2938-2946, en 2015, une méthode d'estimation directe de la pose de la caméra à partir d'images d'une caméra RGB 2D et d'un réseau de neurones convolutif. Le réseau de neurones utilisé est adapté d'un réseau de neurones connu et entraîné pour classifier des images, en modifiant certaines couches. On appellera cette méthode « PoseNet-1 ». L'apprentissage du réseau est réalisé à partir d'une base d'images RGB entières et leurs poses correspondantes. Il s'appuie sur la minimisation d'une fonction de perte qui prend en compte la somme pondérée d'une erreur de translation et d'une erreur de rotation de la pose. La pondération comprend un facteur d'échelle qui dépend de la scène.

Un inconvénient majeur de la solution PoseNet-1 est que la valeur du facteur d'échelle est difficile à déterminer. Son évaluation se fait par voie empirique ce qui nécessite plusieurs apprentissages à partir desquels une valeur optimale est déterminée. La configuration du réseau est donc complexe.

Dans une deuxième publication plus récente, intitulée « Geometric loss functions for camera pose regression with deep learning » et publiée sur ArXiv.2017, les mêmes auteurs ont proposé une nouvelle version appelée « PoseNet-2 » de leur méthode qui utilise une fonction de perte purement géométrique qui n'utilise pas de facteur d'échelle.

Un avantage de ces deux solutions est qu'elles réalisent une estimation de la pose de la caméra en temps réel.

Un inconvénient de ces méthodes PoseNet-2 est qu'elles n'estiment pas toujours la pose de la caméra avec une précision suffisante pour des applications de réalité augmentée et qu'elles ne fournissent aucune indication sur la précision ou la confiance associée à l'estimation de pose.

3. Objectifs de l'invention

L'invention vient améliorer la situation.

L'invention a notamment pour objectif de pallier ces inconvénients de l'art antérieur.

Plus précisément, un objectif de l'invention est de proposer une solution d'estimation de pose plus précise tout en gardant une complexité adaptée à des contraintes temps réel et une configuration simple.

Un autre objectif de l'invention est de fournir une mesure de confiance associé à la pose de la caméra estimée.

Encore un autre objectif de l'invention est de proposer une solution robuste aux occultations de la scène.

4. Exposé de l'invention

Ces objectifs, ainsi que d'autres qui apparaîtront par la suite, sont atteints à l'aide d'un procédé d'estimation de pose d'une caméra dans un référentiel d'une scène tridimensionnelle, ledit procédé comprenant les étapes suivantes :

- Obtention d'une image d'intensités de couleur de la scène capturée par la caméra, dite image courante ;
- Extraction d'une pluralité de points d'intérêt de l'image courante, un dit point étant invariant par transformation géométrique de l'image ;
- Formation d'une pluralité d'imagettes dans l'image d'intensités de couleur, une imagette comprenant un point d'intérêt de la pluralité extraite;
- Prédiction des localisations 3D des points d'intérêt de la pluralité d'imagettes dans un référentiel de la scène, par application d'un système de prédiction automatique, ledit système ayant été entraîné à l'aide d'un ensemble d'apprentissage comprenant des imagettes issues d'une pluralité d'images de la scène acquises par la caméra depuis une pluralité de points de vue, une imagette étant associée à une position 2D de son point d'intérêt dans un référentiel de l'image et à une localisation 3D de son point d'intérêt dans le référentiel de la scène; et
- Estimation d'une pose de la caméra pour l'image courante, par mise en correspondance des positions 2D de la pluralité de points d'intérêt et de reprojections dans le référentiel de l'image courante des localisations 3D prédites.

L'invention repose sur une approche tout-à-fait nouvelle et inventive de l'estimation de pose d'une caméra, qui s'appuie sur une prédiction directe de la localisation 3D de petites zones de l'image ou imagettes centrées sur des points d'intérêts de l'image à partir de leur position 2D dans l'image d'entrée, sans nécessité de recourir à une carte de profondeur, et sur une mise en correspondance 3D/2D des localisations des points d'intérêt de l'image, indépendante des images précédemment acquises.

Plutôt que d'utiliser l'ensemble des points de l'image, l'invention restreint la prédiction à des petites zones de l'image, dites imagettes, qui concentrent l'information pertinente, ce qui permet de réduire la complexité du système sans compromis sur son efficacité.

Le fait de prédire la localisation 3D d'imagettes (3 composantes) plutôt que directement la pose de la caméra (6 composantes) contribue en outre à simplifier la structure du système de prédiction automatique.

Selon un aspect de l'invention, l'estimation de pose comprend la mise en œuvre d'au moins une itération des sous-étapes suivantes :

- Détermination d'un sous-ensemble de la pluralité de points d'intérêt ;
- Calcul d'au moins une hypothèse de pose à partir des localisations 3D prédites pour le sous-ensemble et des positions 2D correspondantes; et
- Evaluation d'une erreur de reprojection des positions 2D de la pluralité de points d'intérêt dans le référentiel de la scène à l'aide de l'hypothèse de pose calculée par rapport aux localisations 3D prédites ;

et une sélection de l'hypothèse de pose qui minimise l'erreur de reprojection.

Un avantage de ce mode de réalisation est de ne prendre en compte que les meilleures prédictions de localisation 3D dans l'estimation de pose.

Selon un autre aspect de l'invention, une mesure de confiance de la pose estimée est évaluée au moins en fonction d'un nombre de points d'intérêt, pour lesquels l'erreur de reprojection est inférieure à un seuil prédéterminé.

Un avantage de cette mesure de confiance est qu'elle apporte une dimension probabiliste à la pose estimée.

Selon encore un autre aspect de l'invention, le procédé comprend une phase préalable d'apprentissage comprenant les étapes suivantes :

- Obtention d'un ensemble d'apprentissage comprenant une pluralité de d'images d'intensité de couleur de la scène acquises par la caméra, depuis une pluralité de points de vue, un point de vue étant associé à une pose connue de la caméra;
- Extraction d'une pluralité de points d'intérêt de ladite image d'intensités de couleurs, un dit point étant associé à une position 2D dans l'image d'intensités;
- Obtention de localisations 3D de la pluralité de points d'intérêt dans le référentiel de la scène, dites de vérité terrain;
- Entraînement du système automatique de prédiction de pose à partir des couples d'imagettes, un couple d'imagettes étant associé à la position 2D de son point d'intérêt dans un référentiel de l'image et à la localisation 3D de son point d'intérêt dans le référentiel de la scène.

Selon un premier mode de réalisation, dans la phase d'apprentissage, la localisation 3D d'un point d'intérêt de l'image d'intensités associée à une pose connue est obtenue par triangulation géométrique des positions 2D du point dans l'image et dans une image précédente.

Un avantage de cette méthode est qu'elle ne nécessite pas de caméra de profondeur pour labelliser les données d'entrées.

Selon un deuxième mode de réalisation, dans la phase d'apprentissage, la caméra est configurée pour acquérir simultanément une image de profondeur associée à une image d'intensités et :

- l'ensemble d'apprentissage obtenu comprend une pluralité de couples d'images d'intensité de couleur et de profondeur de la scène, depuis une pluralité de points de vue, un point de vue étant associé à une pose connue de la caméra;
- les localisations 3D des points d'intérêt sont obtenues par projection perspective de leurs position 2D et profondeur à l'aide d'un modèle prédéterminé de la caméra et de la pose connue, dans le référentiel de la scène.

Un avantage de ce mode de réalisation est qu'il permet de labelliser de façon simple et efficace les imagettes grâce à l'utilisation d'une caméra de profondeur limitée à la phase d'apprentissage.

Selon un autre aspect de l'invention, la prédiction est réalisée par un réseau de neurones comprenant 5 étages de convolution.

Un avantage de cette architecture particulière de réseau de neurones est qu'elle est de faible profondeur, tout en permettant une prédiction efficace des localisations 3D des points d'intérêts de l'image dans le référentiel de la scène directement à partir des imagettes d'intensité.

L'invention concerne également un dispositif adapté pour mettre en œuvre le procédé d'estimation de pose selon l'un quelconque des modes particuliers de réalisation définis ci-dessus. Ce dispositif pourra bien sûr comporter les différentes caractéristiques relatives au procédé selon l'invention. Ainsi, les caractéristiques et avantages de ce dispositif sont les mêmes que ceux du procédé d'estimation de pose, et ne sont pas détaillés plus amplement.

Selon un mode particulier de réalisation de l'invention, un tel dispositif est compris dans un équipement terminal, par exemple de type tablette ou téléphone intelligent (pour « smartphone », en anglais).

L'invention concerne aussi un système de réalité augmentée comprenant une caméra apte à acquérir une image d'intensités de couleurs d'une scène tridimensionnelle réelle, un module) de composition d'images apte à composer une image de sortie à partir d'une image d'entrée acquise de la scène par la caméra et au moins un objet réel ou virtuel, à l'aide d'une localisation 3D initiale dudit au moins un objet dans la scène et d'une pose estimée de la caméra, un module d'affichage apte à restituer l'image de sortie et un dispositif d'estimation de pose de la caméra qui vient d'être décrit.

L'invention concerne aussi un programme d'ordinateur comportant des instructions pour la mise en œuvre des étapes d'un procédé d'estimation de pose d'une caméra tel que décrit précédemment, lorsque ce programme est exécuté par un processeur.

Ce programme peut utiliser n'importe quel langage de programmation. Il peut être téléchargé depuis un réseau de communication et/ou enregistrés sur un support lisible par ordinateur.

L'invention se rapporte enfin à un support d'enregistrement, lisible par un processeur, intégrés ou non au dispositif d'estimation de pose d'une caméra selon l'invention, éventuellement amovible, mémorisant un programme d'ordinateur mettant en œuvre un procédé d'estimation de pose, tel que décrit précédemment.

5. Liste des figures

D'autres avantages et caractéristiques de l'invention apparaîtront plus clairement à la lecture de la description suivante d'un mode de réalisation particulier de l'invention, donné à titre de simple exemple illustratif et non limitatif, et des dessins annexés, parmi lesquels :

- la figure **1** décrit de façon schématique les étapes d'un procédé d'estimation de pose selon l'invention, dans une phase d'apprentissage ;
- la figure **2** illustre une image RGB d'entrée à partir de laquelle on a extrait des points clés et formé des imagettes autour de ces points clés ;
- la figure **3** illustre un exemple d'image RGB d'entrée dont on a labellisé les points clés à partir d'une image de profondeur acquise par une caméra RGB-D et les caractéristiques de la caméra ;
- la figure **4** illustre de façon schématique un modèle de projection de sténopé d'un point de l'image d'entrée dans le référentiel de la caméra à l'aide de la carte de profondeur puis dans le référentiel monde à l'aide de la pose $H = [R|t]$;
- la figure **5** compare la complexité du réseau de neurones convolutif mis en œuvre dans un mode de réalisation de l'invention en termes de nombres de paramètres à celles des réseaux de l'art antérieur ;
- la figure **6** présente de façon schématique la structure en couche du réseau de neurones convolutif selon un mode de réalisation de l'invention ;
- la figure **7** illustre le principe de filtrage par une couche de convolution d'un réseau de neurones convolutif ;
- la figure **8** illustre de façon schématique le principe d'une couche de pooling d'un réseau de neurones convolutif ;
- la figure **9** décrit de façon schématique les étapes d'un procédé d'estimation de pose selon un mode de réalisation de l'invention, dans une phase de test ;
- la figure **10** décrit de façon plus détaillée l'étape d'estimation de pose de l'image d'entrée par mise en correspondance des localisations 2D et 3D de ses imagettes selon un mode de réalisation de l'invention et élimination des prédictions incorrectes (« outliers ») ;
- la figure **11** présente une reconstruction visuelle de la trajectoire de la caméra à partir de sa pose estimée et la compare à celle de la vérité terrain ;
- la figure **12** illustre le rapport entre le nombre d'inliers retenu par le procédé selon l'invention et l'erreur d'estimation de pose de la caméra sur plusieurs scènes ;
- la figure **13** illustre la robustesse de la méthode d'estimation de pose selon l'invention face à une situation d'occultation partielle ;

- la figure **14** illustre le lien entre temps de calcul de la prédiction, précision de l'estimation de pose et nombre d'imagettes extraites de l'image d'entrée ; et
- la figure **15** décrit de façon schématique la structure matérielle d'un dispositif d'estimation de pose selon un mode de réalisation de l'invention.

6. Description d'un mode de réalisation particulier de l'invention

Le principe général de l'invention repose sur la prédiction directe des localisations 3D dans le référentiel de la scène d'une pluralité d'imagettes extraites d'une image 2D d'intensités de couleur de la scène et sur l'estimation de la pose de la caméra par mise en correspondance des localisations 3D prédites avec les positions 2D des imagettes. Cette prédiction est réalisée à l'aide d'un système de prédiction automatique qui nécessite d'être entraîné à partir de données labellisées avec le résultat de la prédiction, au cours d'une phase préalable, dite d'apprentissage. Une fois entraîné, il peut être utilisé sur des données d'entrée non labellisées, dans un mode de fonctionnement normal, dit phase de test.

En relation avec la Figure **1**, on décrit les étapes d'un procédé d'estimation de pose d'une caméra dans une phase d'apprentissage selon un premier mode de réalisation de l'invention.

Au cours d'une étape A1, des données d'apprentissage sont collectées. Ces données sont constituées d'une collection de N images (pour « frames », en anglais), avec N entier non nul, qui sont soit acquises directement par une caméra RGB-D, apte à fournir une image d'intensité de couleur I^C et son image de profondeur I^D associée, soit obtenues d'une base de données publiques. Ces données d'apprentissage comprennent aussi une pose de la caméra dans le référentiel monde, associée à chacune des images de la collection. N est par exemple compris entre 100 et 2000.

Par exemple, les données d'apprentissage sont acquises par un module d'acquisition MACQ constitué par exemple d'un système « Kinect.v2 ® », marque déposée, apte à acquérir simultanément les images d'intensité de couleur et de profondeur et la pose de caméra associée. Ce système comprend un capteur RGB apte à générer une image de résolution w=1920xh=1080 pixels à 30 Hz et un capteur D de profondeur apte à capturer une image de résolution 512x424 avec la même fréquence. L'image d'intensité de couleur et l'image de profondeur sont ensuite alignées de telle sorte qu'elles présentent les mêmes dimensions w, h et correspondent à un unique point de vue de la caméra RGB-D.

Le système Kinect.v2 comprend en outre un émetteur laser et une caméra infrarouge. Le laser génère une onde modulée qui est capturée par la caméra infrarouge. Un temps de trajet de l'onde entre l'émetteur et un objet de la scène est calculé puis exploité pour en déduire une distance entre l'émetteur et l'objet.

Dans la suite, on désigne par (I_i^C, I_i^D) avec i entier compris entre 1 et N, un couple d'images acquis par la caméra RGB-D. Un exemple de couple d'images (I_i^C, I_i^D) est illustré par la Figure **2**.

Selon l'exemple précédent, la pose de la caméra est obtenue en A2 à l'aide d'un module d'annotation apte à calculer une pose associée à chaque paire d'images de la caméra. Il s'agit par exemple d'un module « KinectFusion ®», marque déposée, inclus dans le système « Kinect.v2 ® », dont le principe de fonctionnement est par exemple décrit dans le document de Newcombe *et al.* intitulé « KinectFusion : Real-time Dense Surface Mapping and Tracking », par la conférence IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2011. Ce système est basé sur une méthode de reconstruction dense 3D et temps-réel qui s'appuie sur une technique de Localisation et Mise en correspondance simultanées ou SLAM (pour « Simultaneous Localization and Mapping », en anglais). Une telle technique fait des hypothèses géométriques et temporelles sur la scène. Elle considère une paire d'images en entrée, en extrait des points d'intérêts et les met en correspondance, ce qui lui permet, en résolvant un système linéaire d'équations, de déterminer précisément la pose de la caméra et de reconstruire un modèle 3D de la scène $\mathcal{L}_M$.

Ce module fournit donc les valeurs de pose H_i de la caméra associées à chaque instant d'acquisition d'un couple d'images RGB-D I_i^C, I_i^D. Ces informations constituent une « vérité terrain » nécessaire à l'apprentissage du système de prédiction de localisation 3D des imagettes qui va être décrit ci-après.

On notera qu'il existe d'autres systèmes d'annotation de poses, qui utilisent des marqueurs positionnés sur la caméra RGB-D. Par exemple, ils sont composés d'un matériau qui offre une réponse maximale à un module de segmentation comprenant un laser, apte à les localiser. Par exemple, la base de données CORBS décrite dans le document de Wasenmüller *et al.*, intitulé « Corbs : Comprehensive RGB-D Benchmark for SLAM using Kinect v2 », publié par la conférence Applications of Computer Vision, en 2016, pages 1-7, a été annotée de cette manière.

Selon une alternative, on peut aussi obtenir directement d'une base d'images le couple d'images I_i^C, I_i^D et sa pose Hi associée.

De façon connue en soi, la pose d'un couple d'images s'exprime par exemple sous la forme $H_i = (Q_i, T_i)$, avec Q_i un quaternion unitaire comprenant 4 composantes de rotation q_w, q_x, q_y, q_z et T_i un vecteur comprenant 3 composantes de translation t_x, t_y, t_z de la caméra dans le référentiel monde (O, x, y, z).

Au cours d'une étape A3, on extrait K points d'intérêt KP de l'image d'intensité de couleur I_i^c, avec K entier non nul, inférieur à au nombre w.h de pixels contenus dans l'image I_i^c. On désigne par points d'intérêt, ou points clés, des points invariants aux rotations/translations/ changements d'échelle. Ce module détecte des points isolés (pour « sparse », en anglais) par exemple à l'aide d'une méthode dite SURF et décrite dans le document de Bay *et al.*, intitulé « Speeded-up Robust Features (SURF) », publié dans la revue Computer Vision and Image Understanding , numéro 110, pages 346-359, en 2008.

La méthode SURF exploite une matrice Hessienne $\mathcal{H}(x,\sigma)$ définie comme suit :

$$\mathcal{H}(x,\sigma) = \begin{pmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{yx}(x,\sigma) & L_{yy}(x,\sigma) \end{pmatrix}$$

avec

$$L_{xx}(x,\sigma) = I(x) \times \frac{\partial^2}{\partial^2 x^2} g(\sigma),\ L_{xy}(x,\sigma) = I(x) \times \frac{\partial^2}{\partial xy} g(\sigma)$$

où I(x) représente l'image dans laquelle on cherche à extraire les points clés. $g(\sigma)$ définit une gaussienne avec un noyau σ. La convolution de l'image avec la gaussienne a pour but de représenter l'image à plusieurs résolutions, sous la forme d'une pyramide d'échelles. Une dérivation seconde est appliquée à aux images de la pyramide $\left(\frac{\partial^2}{\partial^2 x^2}, \frac{\partial^2}{\partial^2 y^2}\right)$ ce qui correspond à une intensité de variation de contraste. Concrètement, pour calculer les dérivées secondes sur l'image on utilise des noyaux discrets convolutifs. En relation avec la Figure **3**, on présence des exemples de filtres permettant de calculer L_{xx} et L_{xy} respectivement.

On calcule ensuite le déterminant de $\mathcal{H}$ qui estdéfini comme suit :

$$Det(\mathcal{H}) = D_{xx}D_{yy} - \left(0.9D_{xy}\right)^2$$

où D_{xx} est l'approximation de L_{xx} par une convolution avec un noyau discret. Une réponse maximale correspond à un point clé $KP(x,y,s)$ dont la position dans l'image vaut x, y et s correspond à l'échelle à laquelle il a été extrait. Une orientation de ce point clé est calculée à partir d'une ondelette de Haar dans les directions x, y sur un voisinage prédéterminé. Une orientation principale est calculée comme étant la somme de toute les réponses des ondelettes sur un secteur de $\pi/3$.

En relation avec la Figure **3**, on présente un exemple de points d'intérêt extraits d'une image d'intensités de couleur I_i^C. On note qu'il s'agit par exemple de points correspondant à des angles et des changements de contraste importants.

On extrait au maximum 500 points d'intérêt par image. On associe à chaque point sa valeur d'échelle et son orientation. La valeur d'échelle indique le niveau de détails et d'importance du point clé extrait. L'orientation indique la nature du changement de contraste.

Au cours d'une étape A4, on obtient une localisation 3D des points d'intérêt $P_{k,i}$ chaque couple d'imagettes (I_i^C, I_i^D) dans le référentiel de la scène. Il s'agit d'une labellisation des échantillons de l'ensemble d'apprentissage. Un exemple est illustré par la Figure **2**.

Cette étape, illustrée par la Figure **4**, comprend d'abord une première projection des images (I_i^C, I_i^D) dans un référentiel de la caméra, à l'aide de paramètres intrinsèques Q et t de cette caméra, comprenant un centre de projection (c_x,c_y), une focale horizontale f_x et une focale verticale f_y. On notera que dans le cas de pixels carrés, on a fx= fy.

A partir des paramètres intrinsèques du capteur de profondeur de la caméra RGB-D, chaque valeur de profondeur d (représentée par deux coordonnées de pixel u, v) est projetée dans un référentiel 3D de la caméra, selon un modèle de projection dit de sténopé (pour « pinhole », en anglais) connu en soi, en 3 coordonnées (x, y, z) selon les formules suivantes :

$$\begin{cases} x = \dfrac{d(u - c_x)}{f_x} \\ y = \dfrac{d(v - c_y)}{f_y} \\ z = d \end{cases}$$

Il s'agit d'une modélisation simple et linéaire du processus de formation des images au sein d'une caméra. Ce modèle suppose que le système optique de la caméra, c'est-à-dire sa lentille respecte les conditions de Gauss.

On obtient un triplet $Loc3D_i^{Cam}\left(x_i^{Cam}, y_i^{Cam}, z_i^{Cam}\right)$ qui correspond à la localisation 3D du point d'intérêt dans le référentiel de la caméra.

A l'aide de la pose H_i de la caméra, comprenant les paramètres extrinsèques de la caméra, les paramètres intrinsèques Q et t déjà cités, correspondant à la vérité terrain, ce nuage de points est ensuite projeté dans le référentiel monde de la scène (O,x,y,z), selon une deuxième projection basée sur une transformation rigide. Un triplet $Loc3D_i^{World}\left(x_i^{World}, y_i^{World}, z_i^{World}\right)$ est obtenu.

Les données d'entrée du système automatique d'apprentissage prennent alors la forme suivante : { I_i^C, I_i^D, $Loc3D_i^{World}$} avec i allant de 1 à N, N étant le nombre d'images de la collection de données d'apprentissage.

Au cours d'une étape A5, illustrée par la Figure **3**, on forme ensuite des imagettes (pour « patches », en anglais) centrées sur les points d'intérêt KP extraits, dans chacune des images du couple (I_i^C, I_i^D). Dans cet exemple de réalisation, elles sont toutes choisies avec les mêmes dimensions, égales à 49x49. Plus généralement, on choisit avantageusement le nombre de points d'intérêts et la dimension des imagettes de telle sorte qu'au maximum deux imagettes se recouvrent de 70 %. Par exemple, on considère jusqu'à 500 imagettes dans une image d'entrée de dimensions par exemple égales à 640x380.

A l'issue de cette étape, on dispose d'un ensemble de K couples d'imagettes $(P_{i,k}^C, P_{i,k}^D)$ avec k entier compris entre 1 et K, annotés par leur localisation 3D$Loc3D_i^{World}(x_i^{World}, y_i^{World}, z_i^{World})$

associée à leur couple d'images d'origine (I_i^C, I_i^D). Dans la suite, on désigne par échantillon $E_{i,k}$ un couple d'imagettes et sa localisation 3D associée : $E_{i,k} = \{(P_{i,k}^C, P_{i,k}^D), Loc3D_i^{World}\}$.

Les étapes A2 à A5 sont répétées pour les N couples d'images d'entrée.

Selon une variante des étapes précédentes, on collecte les données d'apprentissage à partir d'un ensemble d'images RGB-D acquises à partir d'une caméra calibrée. Ceci permet d'obtenir directement pour chaque image un nuage de points 3D dans un référentiel de la caméra. On obtient la pose de la caméra à l'aide d'une méthode géométrique de pose, de type « Structure for Motion », qui réalise une triangulation entre deux images 2D successives. Les points clés sont extraits de façon similaire à celle décrite précédemment. On détermine ensuite la localisation 3D des points clés dans le référentiel de la scène à l'aide de la pose obtenue.

A l'issue de l'étape A5, on dispose donc d'une collection de N.K échantillons d'apprentissage $\{E_{i,k}\}$.

Au cours d'une étape A6, on présente cet ensemble de N.K échantillons $\{E_{i,k}\}$ en entrée d'un système de prédiction automatique. Dans cet exemple de réalisation de l'invention, il s'agit d'un réseau de neurones convolutif dont un exemple est illustré par la figure **6**.

Ce réseau de neurones convolutif a été spécialement conçu pour prédire directement la localisation 3D des éléments de la pluralité d'imagettes issue d'une image RGB dans le référentiel de la scène, à partir de leur position 2D dans l'imagette. Il utilise uniquement des imagettes d'image RGB de dimensions fixes, par exemple égales à 49×49 pixels.

Les inventeurs ont constaté que les réseaux de neurones convolutifs connus, comme AlexNet, décrit par Krizhevsky et al. dans le document intitulé «Imagenet classification with deep convolutional neural networks,» publié dans la revue « *Advances in neural information processing systems* », en 2012, VGG-Net décrit par Simonyan et al., dans le document intitulé : «Very Deep Convolutional Networks for Large-Scale Image Recognition» publié dans *ICLR*, en 2014, GoogleNet décrit par Szegedy et al., dans un article intitulé «Going deeper with convolutions» publié dans les *Proceedings of the IEEE conference on computer vision and pattern recognition*, en 2015 ou ResNet décrit par He et al., dans un article intitulé «Deep residual learning for image recognition», publié dans les *Proceedings of the IEEE conference on computer vision and pattern recognition*, en 2016, sont conçus pour classifier des objets et ne conviennent pas au traitement de plusieurs imagettes, en en termes de temps de calcul, de l'ordre de 10 ms à quelques secondes par image). Ils sont généralement très profonds, c'est-à-dire qu'ils comprennent un nombre élevé de couches et de ce fait mettent en œuvre un nombre élevé de paramètres, comme illustré par le graphique de la Figure **5**. Une conséquence de cette complexité est un temps d'apprentissage élevé, de l'ordre de quelques semaines pour un million d'images.

Selon un mode de réalisation, l'invention propose un réseau neuronal convolutif léger, que l'on désignera par « xyzNet », pour localiser les K points d'intérêt d'une image d'entrée dans le système de coordonnées du monde. Ce réseau de neurones convolutif se compose, comme illustré par la Figure **6**, de cinq couches de convolution CONV qui constituent 5 étages successifs de la structure du réseau de neurones. Les deux premières couches de convolution sont suivies d'une couche dite « RELU » réalisant une opération non linéaire et d'une couche de sous-échantillonnage dite « POOL ». Seul le premier étage comprend, suite à la couche de sous-échantillonnage POOL une couche de normalisation de réponse dite « LRN ». Les 5 étages ainsi formés sont suivis de deux étages comprenant chacun une couche de connexion neuronale complète dite « FC » et d'une couche RELU. Pour superviser le réseau, on utilise une fonction de perte (pour « LOSS », en anglais**)** définie par une différence de distance euclidienne entre la la localisation 3D réelle d'un point d'intérêt (ou vérité) et sa localisation 3D prédite**.**

On décrit maintenant les différents types de couches de façon plus détaillée :

- CONV pour « Convolution » en anglais) : Chacune des cinq convolutions effectue une opération de filtrage spatiale avec ensemble de filtres de taille 3×3 (conv1 : 32 filtres ; conv2 : 64 filtres ; conv3, conv4, conv5 : 128 filtres). On décrit plus en détails une couche de convolution, en relation avec la Figure **7**. Il s'agit d'un bloc de base d'un réseau CNN, qui

est utilisé pour extraire des caractéristiques des données d'entrée. Une couche comprend généralement plusieurs filtres. Un filtre est entraîné au cours de la phase d'apprentissage pour détecter une même caractéristique à différentes positions dans l'image d'entrée et produire une carte. Le produit de convolution de l'image $\boldsymbol{x}$ et le filtre $\boldsymbol{h}$ est définie par l'expression suivante.

$$y[m,\ n] = \ x[m,n] * h[m,n] = \sum_{j:-\infty}^{+\infty} \sum_{i:-\infty}^{+\infty} x[i,j].h[m-i,n-j]$$

- POOL *(pour « Pooling », en anglais) :* les deux couches de pooling permettent de réaliser un sous-échantillonnage avec une métrique basée sur la valeur maximum dans un voisinage de taille 3x3. Cette opération permet, d'un côté d'ajouter une forme d'invariance aux translations et aux rotations. D'un autre côté, elle réduit la quantité de paramètres permettant de définir les poids du réseau en question. La Figure **8** illustre cette opération.

- LRN (*pour « Local Response Normalization », en anglais)*, est une opération permettant de normaliser la sortie d'une couche donnée (en imposant la moyenne à zéro et la déviation standard à 1) sur l'ensemble d'imagettes envoyées en parallèle pendant une itération (pour « batch », en anglais). L'importance de cette opération réside dans qualité de convergence de l'apprentissage du réseau. En effet, en normalisant, on réduit le taux de variabilité des données entre elles permettant un comportement plus restreint de l'optimisation de la fonction (par descente de gradient par exemple). Cette opération peut être formalisée de la manière suivante :

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B)^2$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta$$

Où μ_B représente la moyenne de toutes les imagettes du batch B, et σ_B désigne la déviation standard relative au même batch. $\hat{x}_i$ représente le signal centré normalisé et y_i définit le signal de sortie final shifté avec deux paramètres γ et β respectivement.

- FC (pour « *Fully connected », en anglais)*, est une opération permettant de regrouper toutes les réponses neuronales obtenues à partir d'une donnée d'entrée initiale pour constituer une représentation encodée de cette donnée d'entrée. Notre réseau comporte deux couches FC (FC6, FC7). La première permet de constituer une représentation de dimension (1x1024), et la deuxième reconstitue une représentation finale de même dimension que la localisation 3D qu'on veut régresser ou prédire, à savoir 1x3. Cette opération est, analytiquement, définie comme suit :

$$y_i = W.x_i + b^T$$

Où x et y représentent respectivement les représentations de la donnée à l'entrée et à la sortie de la couche FC respectivement, W définit les poids à apprendre au niveau de cette couche, et b^T est le biais.

- RELU (pour « *Rectified Linear Unit », en anglais)*, est un opérateur non linéaire qui calcule les activations des neurones par une excitation donnée. Dans notre réseau, chaque convolution est suivie d'une couche relu pour introduire une non-linéarité importante dans le comportement du réseau et obtenir une meilleure discrimination. Cette rectification s'applique sur la fonction d'activation standard utilisée précédemment qui est par exemple la fonction sigmoïde représentée comme suit :

$$f(x) = \frac{1}{1 + e^{-x}}$$

Cette fonction représente une limitation importante pour la convergence de l'apprentissage. On peut la rectifier de la manière suivante :

$$f(x) = \max(0, x)$$

ce qui permet un résultat plus consistant et rapide en terme de convergence.

- Fonction de perte ou LOSS **:** Dans la phase d'apprentissage, les poids du réseau xyzNet sont appris en minimisant une fonction de perte d'objectif euclidienne avec un algorithme d'optimisation descendant gradient stochastique. La fonction de perte est alors définie comme suit:

$$loss = \sum_{k \in B} norm(loc3D_k^{world} - \widehat{loc3D_k^{world}})$$

Cette fonction de perte évalue la qualité de la localisation 3D prédite. Cette localisation se réduit à 3 coordonnées (x,y,z), ce qui la rend plus facile à optimiser que celle associée à un système de prédiction automatique de la pose de la caméra, du type « posenet », dont la donnée de sortie comprend 6 paramètres (3 translations et 3 rotations).

Dans ce mode de réalisation de l'invention, 500 imagettes sont extraites de chaque image avec la contrainte que deux imagettes ne se chevauchent pas plus de 70%. Les imagettes d'entraînement sont normalisées par soustraction de la valeur d'intensité moyenne de tous les imagettes.

En relation avec la Figure **9**, on décrit maintenant le procédé d'estimation de pose dans sa phase de test, selon un mode de réalisation de l'invention.

- Au cours d'une étape E1, on obtient une image d'entrée I acquise par une caméra RGB 2D, par exemple de dimensions VGA 640x480. Chaque élément d'image comprend 3 intensités R, G et B.
- Au cours d'une étape E2, on extrait de l'image I un nombre K prédéterminé, par exemple égal à 500, de points clés KP, par exemple selon la méthode SURF précédemment décrite.
- Au cours d'une étape E3, on forme autant d'imagettes ou patches P_k que de points clés extraits, une imagette étant centrée sur un point clé KP_k et de dimensions prédéterminées, par exemple égales à 49x49. A l'issue de cette étape, on dispose donc pour l'image I de K imagettes P_k avec k entier compris entre 1 et K, à présenter en entrée du système de prédiction automatique, qui est dans cet exemple de réalisation, le réseau de neurones convolutif xyznet qui a subi la phase d'apprentissage précédemment décrite.
- Au cours d'une étape E4, les K imagettes $\{P_k\}$ sont traitées par le réseau xyznet, qui produit une prédiction de localisation 3D par imagette P_k : $\widehat{Loc3D_k^{World}}(x_k^{World}, y_k^{World}, z_k^{World})$

- En E5, on estime la pose $H = [R|T]$ (R : rotation,T : translation) de l'image I à partir des K prédictions de localisation 3D des K imagettes extraites de l'image I.

Dans ce mode de réalisation, illustré par la Figure **10**, l'estimation de pose E5 est réalisée en générant en $E5_1$ plusieurs hypothèses de poses à partir de sous-ensembles choisis aléatoirement parmi les K localisations prédites par le système de prédiction automatique et en sélectionnant en $E5_2$ l'hypothèse de pose qui minimise une erreur de reprojection.

Plus précisément, la sous-étape $E5_1$ comprend :

- Une détermination aléatoire $E5_{11}$ d'un sous-ensemble SE de M avec 2<M<K points d'intérêt parmi les K points d'intérêt KP_k de l'image courante;
- Un calcul $E5_{12}$ d'une hypothèse de pose Hi par mise en correspondance 2D/3D des positions 2D Loc2Dm des M points d'intérêt avec leurs localisations 3D Loc3Dm prédites. Seules les prédictions exactes de trois pixels sont théoriquement nécessaires pour déduire la pose de la caméra. On utilise avantageusement une méthode de vision par ordinateur bien connue et basée sur ce principe, à savoir l'algorithme Perspective-n-Points (PnP) ;
- Une évaluation d'une erreur de reprojection des localisations 3D des K points d'intérêt KP dans le référentiel de la scène à l'aide de l'hypothèse de poseH. Cette erreur est calculée comme suit :

$$Error(H) = \sum norm(Loc2D_k - reproj(\widehat{Loc3D_k^{world}}))$$

Ou $reproj()$ est une fonction de reprojection 3D->2D définie comme suit :

$$reproj(X) = K.H^{-1}.X$$

Avec

$$H_n : \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix}$$

Où R et t définissent la rotation et la translation de la caméra avec ses paramètres intrinsèques K.

On définit les points d'intérêts correctement prédits (pour « inliers », en anglais) en optimisant l'expression suivante :

$$\max_{H_i} \sum_{j \in B} \rho(\alpha_{ij})$$

Où

$$\alpha_{ij} = norm(Loc2D - K.H_i^{-1}\widehat{Loc3D_j^{world}}) - \tau$$

On définit notre fonction ρ comme suit :

$$\rho(\alpha) = \begin{cases} 1 \; if \; \alpha < 0 \\ 0 \; otherwise \end{cases}$$

τ est le seuil maximal d'erreur de reprojection qui définit les inliers.

Ces étapes sont itérées N fois, avec N entier non nul, selon un algorithme dit de Ransac (pour « Random Sample Access », en anglais).

En $E5_2$, on sélectionne parmi les hypothèses de pose H_n la meilleure hypothèse de pose H parmi N obtenues, comme celle qui obtient l'erreur de reprojection Er la plus faible sur ses J points d'intérêts définis comme inliers.
La prise en compte des inliers seulement permet d'éliminer les prédictions bruyantes ou erronées fournies par le réseau de neurones convolutif.

Avantageusement, le nombre d'inliers associés à une hypothèse est utilisé pour évaluer une mesure de confiance associée à la pose de caméra estimée.

Par exemple, on la définit en considérant qu'un nombre d'inliers égal à 80 représente approximativement une erreur de translation de 5 cm et une erreur de rotation de 2 degrés en moyenne. Ces valeurs représentent un seuil de précision important en réalité augmentée. On associe de ce fait une mesure de confiance maximale de 100% à un nombre supérieur ou égal à 80 inliers.

Une telle mesure de confiance empirique permet d'injecter un aspect probabiliste à la méthode d'estimation de pose selon l'invention.

Un avantage de l'invention est de combiner l'apprentissage en profondeur de multiples imagettes avec une méthode de filtrage basée sur les algorithmes PnP et Ransac pour éliminer les outliers, par exemple localisés sur les objets en mouvement dans la scène ou encore les objets partiellement occultés.

On présente maintenant des mesures de performances du procédé d'estimation de pose selon le mode de réalisation de l'invention qui vient d'être décrit.

On considère 7 scènes distinctes à l'échelle d'une pièce. Ces scènes comportent des trajectoires géométriquement complexes comportant des rotations pures, des changements brusques de direction et des mouvements rapides de la caméra.

Chaque scène comprend des séquences qui sont capturées autour d'une seule pièce et annotées en utilisant Kinect Fusion, comme précédemment décrit.

On considère en outre la base de données CORBS décrite dans le document de Wasenmüller *et al.*, intitulé « Corbs : Comprehensive RGB-D Benchmark for SLAM using Kinect v2 », publié par la conférence Applications of Computer Vision, en 2016, pages 1-7. Elle comprend un ensemble de données annotées de façon plus précise grâce à un système de capteurs multiples. Les données visuelles sont archivées à l'aide d'un Kinect v2. La vérité terrain pour la trajectoire est obtenue par un système de capture de mouvement externe. Chaque scène contient un modèle de scène 3D dense obtenu via un scanner 3D externe.

Scene	Chess	Fire	Heads	Office	Pumpkin	RedKitchen	Stairs
Err_P	0.25m	0.19m	0.14m	0.65m	0.27m	0.44m	0.34m
Err_I	0.13m	0.11m	0.06m	0.26m	0.11m	0.14m	0.13m

Table 1

La table 1 évalue les performances de prédiction du réseau de neurones xyznet sur les 7 scènes. Nous calculons la moyenne de l'erreur de distance entre les prédictions obtenues et les vérités terrains sur tous les prédictions (Err_p) et sur les inliers Err_i seulement.

Scène	**PoseNet-1**	**PoseNet-2**	**xyzNet**
Chess	0,32 m – 8,12 °	**0,13 m – 4,48 °**	0,18 m – 4,80 °
Fire	0,47 m – 14,4 °	0,27 m – 11,3 °	**0,21 m – 6,72 °**
Heads	0,29 m – 12, 0 °	0,17 m – 13,0 °	**0,15 m – 8,08 °**
Office	0,48 m – 7,68 °	**0,19 m – 5,55 °**	0,47 m – 8,08 °

Pumpkin	0,47 m – 8,42 °	0,26 m – 4,47 °	**0,17 m – 4,18 °**
Red Kitchen	0,59 m – 8,84 °	0,23 m – 5,35 °	**0,20 m – 4,65 °**
Stairs	0, 47 m – 13, 8 °	0,35 m – 12, 4 °	**0,19 m – 4,63 °**
Moyenne	0,44 m – 10, 4 °	0,23 m – 8,13 °	**0,22 m – 6,09 °**

Table 2

En relation avec la Table **2**, on compare les performances du procédé selon le premier et le deuxième mode de réalisation de l'invention avec les méthodes posenet de l'état de l'art pour 7 scènes.

On constate que le procédé selon le premier mode de réalisation de l'invention donne en moyenne des résultats plus précis que cet art antérieur.

En relation avec la Figure **11**, on présente une reconstruction de la trajectoire TE de la caméra à partir des poses estimées selon l'invention et on la compare visuellement à celle TVT de la vérité terrain. L'écart observé entre les deux trajectoires correspond à l'erreur d'estimation de l'ordre de quelques centimètres.

En relation avec la Figure **12**, on illustre l'impact du nombre d'inliers sur la précision de la pose estimée. Elle montre la relation proportionnelle entre le nombre d'inliers et la précision sur l'estimation de la pose de la pose. Pour une scène donnée, par ex « Chess » l'erreur pour la translation est de 0.17 cm pour 10 inliers contre 0.07cm pour 80 inliers. Ce résultat est directement à la minimisation de l'erreur par PnP qui estime une meilleure hypothèse de pose de la caméra avec un plus grand nombre d'inliers.

Ceci confirme le fait que le nombre d'inliers est une information pertinente pour définir une mesure de confiance de l'estimation de pose de caméra réalisée.

En relation avec la figure **13**, on illustre la capacité du procédé selon l'invention à traiter une occultation partielle Pour ce faire, on présente des images d'entrée dégradées comprenant une zone rectangulaire noire ZRN de taille 200x200. La zone change de position d'une image à l'autre. On constate que quelle que soit la position de la zone de masquage, l'estimation de la pose de la caméra reste correcte. En effet, on constate que l'objet synthétique OS en forme de parallélépipède rectangle reste inséré correctement dans la scène quelle que soit la position de la zone de masquage ZRN.

Ces bons résultats s'expliquent par le fait que le procédé selon l'invention s'appuie sur des points inliers correctement prédits et issus d'imagettes localisées en dehors de la zone dégradée. Cette approche non dense (pour « sparse », en anglais) apporte une robustesse de la solution à une occlusion partielle de l'image et permet de continuer à fournir une estimation robuste et précise de la pose de la caméra alors que l'image d'entrée est visuellement fortement dégradée.

On notera que le résultat obtenu sera probablement associé à une mesure de confiance plus faible du fait d'un nombre plus réduit d'inliers.

Le problème d'occultation partielle étant très fréquent dans les scènes réelles, la robustesse de l'approche selon l'invention est un atout pour le d'un environnement dynamique.

En relation avec la figure **14**, on présente le temps de calculs de prédiction en fonction du nombre d'imagettes extraites. Elle illustre également l'impact du nombre de patches utilisé sur la qualité de la prédiction. Environ 130 imagettes produisent une erreur d'environ 0.05m pour un temps d'éxécution de 80ms correspondant à environ 13 fps. Ce temps d'exécution est court ce qui démontre la faible complexité de l'approche selon l'invention de notre approche due directement à l'architecture du réseau de neurones convolutif permettant un temps d'inférence rapide et une discrimination et généralisation suffisante pour une bonne localisation 3D.

L'invention propose ainsi une approche basée sur un réseau de neurones convolutif qui permet d'estimer de façon régressive la localisation 3D de points d'une image 2D acquise par une caméra RGB dans un environnement non contraint. Contrairement à l'art antérieur, les échantillons d'apprentissage sont collectés de façon non dense sous la forme d'une pluralité d'imagettes centrées sur des points clés d'une image d'intensités acquise par la caméra. Les résultats obtenus, notamment en termes d'erreurs en translation et rotation réalisées sur des bases de données publiques valident cette approche et montrent notamment qu'elle est plus précise que les méthodes de l'art antérieur, tout en restant temps réel et simple à configurer.

On notera que l'invention qui vient d'être décrite, peut être mise en œuvre au moyen de composants logiciels et/ou matériels. Dans cette optique, les termes « module » et « entité », utilisés dans ce document, peuvent correspondre soit à un composant logiciel, soit à un composant matériel, soit encore à un ensemble de composants matériels et/ou logiciels, aptes à mettre en œuvre la ou les fonctions décrites pour le module ou l'entité concerné(e).

En relation avec la Figure **15,** on présente maintenant un exemple de structure simplifiée d'un dispositif 100 d'estimation de pose d'une caméra selon l'invention. Le dispositif 100 met en œuvre le procédé d'estimation de pose selon l'invention qui vient d'être décrit.

Cette figure **15** illustre seulement une manière particulière, parmi plusieurs possibles, de réaliser l'algorithme détaillé ci-dessus. En effet, la technique de l'invention se réalise indifféremment sur une machine de calcul reprogrammable (un ordinateur PC, un processeur DSP ou un microcontrôleur) configurée pour exécuter un programme comprenant une séquence d'instructions, ou sur une machine de calcul dédiée (par exemple un ensemble de portes logiques comme un FPGA ou un ASIC, ou tout autre module matériel).

Dans le cas où l'invention est implantée sur une machine de calcul reprogrammable, le programme correspondant (c'est-à-dire la séquence d'instructions) pourra être stocké dans un médium de stockage amovible (tel que par exemple une disquette, un CD-ROM ou un DVD-ROM) ou non, ce médium de stockage étant lisible partiellement ou totalement par un ordinateur ou un processeur.

Par exemple, le dispositif 100 comprend une unité de traitement 110, équipée d'un processeur µ1, et pilotée par un programme d'ordinateur Pg1 120, stocké dans une mémoire 130 et mettant en œuvre le procédé de selon l'invention.

A l'initialisation, les instructions de code du programme d'ordinateur Pg_1 120 sont par exemple chargées dans une mémoire RAM avant d'être exécutées par le processeur de l'unité de traitement 110. Le processeur de l'unité de traitement 110 met en œuvre les étapes du procédé décrit précédemment, selon les instructions du programme d'ordinateur 120.
Dans cet exemple de réalisation de l'invention, le dispositif 100 comprend une machine de calcul reprogrammable ou une machine de calcul dédiée, apte à et configurée pour :

- Obtenir ACQ une image d'intensités de couleur de la scène capturée par la caméra, dite image courante ;
- Extraire EXTR une pluralité de points d'intérêt de l'image courante, un dit point étant invariant par transformation géométrique de l'image ;
- Former FORM PT une pluralité d'imagettes dans l'image d'intensités de couleur, une imagette comprenant un point d'intérêt de la pluralité extraite;

- Prédire PRED des localisations 3D des points d'intérêt de la pluralité d'imagettes dans un référentiel de la scène, par application d'un système de prédiction automatique, ledit système ayant été entraîné à l'aide d'un ensemble d'apprentissage comprenant des imagettes issues d'une pluralité d'images de la scène acquises par la caméra depuis une pluralité de points de vue, une imagette étant associée à une position 2D de son point d'intérêt dans un référentiel de l'image et à une position 3D de son point d'intérêt dans le référentiel de la scène;
- Estimer EST H une pose de la caméra pour l'image courante, par mise en correspondance des positions 2D de la pluralité de points d'intérêt et de reprojections dans le référentiel de l'image courante des localisations 3D prédites.

Avantageusement, la machine de calcul est configurée pour mettre en œuvre les modes de réalisation de l'invention qui viennent d'être décrits en relation avec les Figures **1** et **9**.

En particulier, elle est en outre apte à mettre en œuvre la phase d'apprentissage et la phase de test du système de prédiction automatique selon l'invention telles que précédemment décrites. Elle est alors configurée pour :

- Obtenir un ensemble d'apprentissage comprenant une pluralité de d'images d'intensité de couleur de la scène acquises par la caméra, depuis une pluralité de points de vue, un point de vue étant associé à une pose connue de la caméra;
- Extraire une pluralité de points d'intérêt de ladite image d'intensités de couleurs, un dit point étant associé à une position 2D dans l'image d'intensités et à une profondeur dans l'image de profondeur ;
- Obtenir des localisations 3D de la pluralité de points d'intérêt dans le référentiel de la scène;
- Entraîner le système automatique de prédiction de pose à partir des couples d'imagettes, un couple d'imagettes étant associé à la position 2D de son point d'intérêt dans un référentiel de l'image et à la localisation 3D de son point d'intérêt dans le référentiel de la scène.

Le dispositif 100 comprend en outre une unité M_1 140 de stockage, telle qu'une mémoire, par exemple de type mémoire tampon (pour « buffer », en anglais), apte à stocker par exemple les localisations 3D prédites de la pluralité de points clés et/ou les hypothèses de pose Hn estimées et les point inliers obtenus selon la technique PnP – Ransac décrite en relation avec la Figure **10**.

Ces unités sont pilotées par le processeur μ1 de l'unité de traitement 110.

De façon avantageuse, un tel dispositif 100 d'estimation de pose peut être intégré à un système 10 de réalité augmentée.

Un tel système 10 comprend, en plus du dispositif 100, au moins un module d'acquisition MACQ d'images d'entrée, comprenant par exemple une caméra RGB apte à capturer une image d'une scène réelle, un module d'annotation ANNOT apte à produire des localisations 3D de points d'intérêt des images, un module de composition COMP apte à composer une image de sortie, dite « augmentée » à partir d'une image d'entrée de la scène acquise par la caméra et au moins un objet réel ou virtuel, comme illustré par la Figure 14, à l'aide d'une position initiale dudit au moins un objet dans la scène et d'une pose estimée de la caméra et un module DISP d'affichage apte à restituer l'image de sortie.

Selon une variante, le dispositif 100, une fois entraîné, peut être intégré à un équipement terminal ET, par exemple un ordinateur personnel, qui peut être mobile, comme une tablette ou un téléphone intelligent (pour « smartphone », en anglais), est lui-même compris dans le système 10.

Le dispositif 100 est alors agencé pour coopérer au moins avec les modules suivants du système 10 ou de l'équipement terminal ET:

- un module E/R d'émission/réception de données, par l'intermédiaire duquel une image RGB est obtenue, par exemple en provenance d'une base de données distante; et/ou
- le module d'acquisition MACQ de la séquence d'images d'entrée, tel que par exemple une caméra vidéo RGB, par exemple via un câble HDMI ;
- le module d'annotation ANNOT apte à produire les localisations 3D de la pluralité de points d'intérêt extraits d'une image d'entrée, par exemple de type kinectfusion® ;
- le dispositif d'affichage DISP, configuré pour restituer une composition d'une image RGB 2D avec la scène 3D virtuelle ou réelle à l'aide de la pose estimée de la caméra.

Grâce à ses bonnes performances et à sa simplicité de mise en œuvre, l'invention qui vient d'être décrite permet plusieurs usages. Une première application est d'augmenter la réalité d'une scène filmée par la caméra RGB-D, en y injectant des objets supplémentaires, virtuels ou réels. On connait par exemple une application de décoration intérieure, qui permet à un client de tester virtuellement l'agencement d'un mobilier dans une pièce de son appartement, avant se décider à l'achat. Cette application nécessite une estimation de la pose de la caméra dans un référentiel de la pièce, de façon à localiser l'image qu'elle acquiert dans la scène et à y insérer, lors de leur restitution sur un dispositif d'affichage, le mobilier virtuel avec les bonnes dimensions et la bonne perspective. Une position spatiale du mobilier virtuel est initialisée dans la scène. Elle nécessite une connaissance

a priori d'une structure 3D de la pièce. Ensuite, un suivi de la trajectoire de la caméra est réalisé en estimant sa pose dans un référentiel de la scène selon l'invention, ce qui permet, pour chaque nouvelle image acquise, de projeter le mobilier virtuel dans la scène, à la bonne position et avec la bonne perspective. Avec l'invention, du fait que le traitement est moins complexe et ne nécessite pas dans sa phase de test de caméra de profondeur, il devient envisageable de mettre en œuvre cette application sur un équipement terminal mobile, de type tablette ou téléphone intelligent (pour « smartphone », en anglais).

Une deuxième application envisagée est l'assistance d'un opérateur de maintenance, par exemple d'avions. On suppose qu'il acquiert une image de pièces du moteur à partir d'un équipement terminal mobile, de type tablette. Le système selon l'invention lui permet, dans sa phase de test, d'estimer la pose de la caméra dans la scène constituée par le moteur de l'avion à partir de l'image courante. La connaissance au préalable de la structure 3D du moteur permet d'initialiser un rendu d'informations supplémentaires relatives à une de ses pièces. Par exemple, on affiche une référence du modèle, des informations relatives à sa qualité, une date d'installation etc. Avec l'invention, il est possible de suivre la trajectoire de la caméra et d'estimer sa pose à chaque nouvelle image acquise. De cette manière, les informations supplémentaires sont projetées dans chaque nouvelle image acquise par la caméra avec la bonne perspective, ce qui garantit de maintenir au cours du temps un réalisme du rendu de la scène vue par la caméra.

Il va de soi que les modes de réalisation qui ont été décrits ci-dessus ont été donnés à titre purement indicatif et nullement limitatif, et que de nombreuses modifications peuvent être facilement apportées par l'homme de l'art sans pour autant sortir du cadre de l'invention.

REVENDICATIONS

1. Procédé d'estimation de pose d'une caméra dans un référentiel d'une scène tridimensionnelle, comprenant les étapes suivantes :

- Obtention (E1) d'une image d'intensités de couleur de la scène capturée par la caméra, dite image courante ;
- Extraction (E2) d'une pluralité de points d'intérêt de l'image courante, un dit point étant invariant par transformation géométrique de l'image ;
- Formation (E3) d'une pluralité d'imagettes dans l'image d'intensités de couleur, une imagette comprenant un point d'intérêt de la pluralité extraite;
- Prédiction (E4) des localisations 3D des points d'intérêt de la pluralité d'imagettes dans un référentiel de la scène, par application d'un système de prédiction automatique, ledit système ayant été entraîné à l'aide d'un ensemble d'apprentissage comprenant des imagettes issues d'une pluralité d'images de la scène acquises par la caméra depuis une pluralité de points de vue, une imagette étant associée à une position 2D de son point d'intérêt dans un référentiel de l'image et à une position 3D de son point d'intérêt dans le référentiel de la scène;
- Estimation (E5) d'une pose de la caméra pour l'image courante, par mise en correspondance des positions 2D de la pluralité de points d'intérêt et de reprojections dans le référentiel de l'image courante des localisations 3D prédites.

2. Procédé d'estimation de pose d'une caméra selon la revendication **1**, caractérisé en ce que l'estimation de pose comprend la mise en œuvre d'au moins une itération des sous-étapes suivantes :

- Détermination ($E5_1$) d'un sous-ensemble de la pluralité de points d'intérêt ;
- Calcul ($E5_2$) d'au moins une hypothèse de pose à partir des localisations 3D prédites pour le sous-ensemble et des positions 2D correspondantes; et
- Evaluation ($E5_3$) d'une erreur de reprojection des positions 2D de la pluralité de points d'intérêt dans le référentiel de la scène à l'aide de l'hypothèse de pose calculée par rapport aux localisations 3D prédites ;

et en ce qu'elle comprend une sélection (E6) de l'hypothèse de pose qui minimise l'erreur de reprojection.

3. Procédé selon la revendication **2**, caractérisé en ce qu'une mesure de confiance de la pose estimée est évaluée au moins en fonction d'un nombre de points d'intérêt, pour lesquels l'erreur de reprojection est inférieure à un seuil prédéterminé.

4. Procédé selon l'une des revendications précédentes, caractérisé en ce qu'il comprend une phase préalable d'apprentissage comprenant les étapes suivantes :

- Obtention (A1) d'un ensemble d'apprentissage comprenant une pluralité de d'images d'intensité de couleur de la scène acquises par la caméra, depuis une pluralité de points de vue, un point de vue étant associé à une pose connue de la caméra;
- Extraction (A2) d'une pluralité de points d'intérêt de ladite image d'intensités de couleurs, un dit point étant associé à une position 2D dans l'image d'intensités ;
- Obtention (A3) de localisations 3D de la pluralité de points d'intérêt dans le référentiel de la scène, dites de vérité terrain;
- Entraînement (A4) du système automatique de prédiction de pose à partir des couples d'imagettes, un couple d'imagettes étant associé à la position 2D de son point d'intérêt dans un référentiel de l'image et à la localisation 3D de son point d'intérêt dans le référentiel de la scène.

5. Procédé selon la revendication **4**, caractérisé en ce que, dans la phase d'apprentissage, la localisation 3D de vérité terrain d'un point d'intérêt de l'image d'intensités associée à une pose connue est obtenue par triangulation géométrique des positions 2D du point dans l'image et dans une image précédente.

6. Procédé selon la revendication **4**, caractérisé en ce que, dans la phase d'apprentissage, la caméra étant configurée pour acquérir simultanément une image de profondeur associée à une image d'intensités :

- l'ensemble d'apprentissage obtenu comprend une pluralité de couples d'images d'intensité de couleur et de profondeur de la scène, depuis une pluralité de points de vue, un point de vue étant associé à une pose connue de la caméra;
- les localisations 3D des points d'intérêt sont obtenues par projection perspective de leurs position 2D et profondeur à l'aide d'un modèle prédéterminé de la caméra et de la pose connue, dans le référentiel de la scène.

7. Procédé selon l'une des revendications **4** à **6**, caractérisé en ce que caractérisé en ce que, la prédiction étant réalisée par un réseau de neurones, ledit réseau comprenant 5 étages de convolution.

8. Dispositif (100) d'estimation de pose d'une caméra dans un référentiel d'une scène tridimensionnelle, ledit dispositif comprenant une machine de calcul dédiée à ou configurée pour :

- Obtenir une image d'intensités de couleur de la scène capturée par la caméra, dite image courante ;
- Extraire une pluralité de points d'intérêt de l'image courante, un dit point étant invariant par transformation géométrique de l'image ;
- Former une pluralité d'imagettes dans l'image d'intensités de couleur, une imagette comprenant un point d'intérêt de la pluralité extraite;
- Prédire des localisations 3D des points d'intérêt de la pluralité d'imagettes dans un référentiel de la scène, par application d'un système de prédiction automatique, ledit système ayant été entraîné à l'aide d'un ensemble d'apprentissage comprenant des imagettes issues d'une pluralité d'images de la scène acquises par la caméra depuis une pluralité de points de vue, une imagette étant associée à une position 2D de son point d'intérêt dans un référentiel de l'image et à une position 3D de son point d'intérêt dans le référentiel de la scène;
- Estimer une pose de la caméra pour l'image courante, par mise en correspondance des positions 2D de la pluralité de points d'intérêt et de reprojections dans le référentiel de l'image courante des localisations 3D prédites.

9. Système (10) de réalité augmentée comprenant :

- une caméra (MACQ) apte à acquérir une image d'intensités de couleurs d'une scène tridimensionnelle réelle,
- un module (COMP) de composition d'images apte à composer une image de sortie à partir d'une image d'entrée acquise de la scène par la caméra et au moins un objet réel ou virtuel, à l'aide d'une localisation 3D initiale dudit au moins un objet dans la scène et d'une pose estimée de la caméra,
- un module (DISP) d'affichage apte à restituer l'image de sortie,

caractérisé en ce qu'il comprend un dispositif d'estimation de pose de la caméra selon la revendication **8**.

10. Programme d'ordinateur (Pg1) comportant des instructions pour la mise en œuvre du procédé de suivi de cible selon l'une quelconque des revendications **1** à **7**, lorsque ledit programme est exécuté par un processeur.

11. Support d'enregistrement lisible par un ordinateur, sur lequel est enregistré un programme d'ordinateur comprenant des instructions de code de programme pour l'exécution des étapes du procédé selon l'une des revendications **1** à **7**.

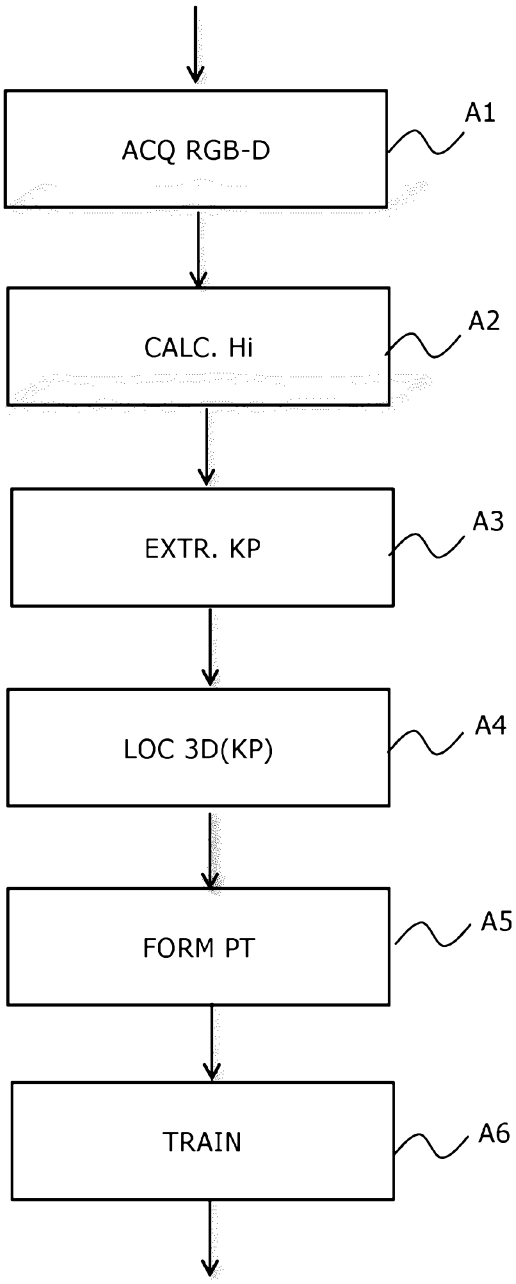


FIG. 1

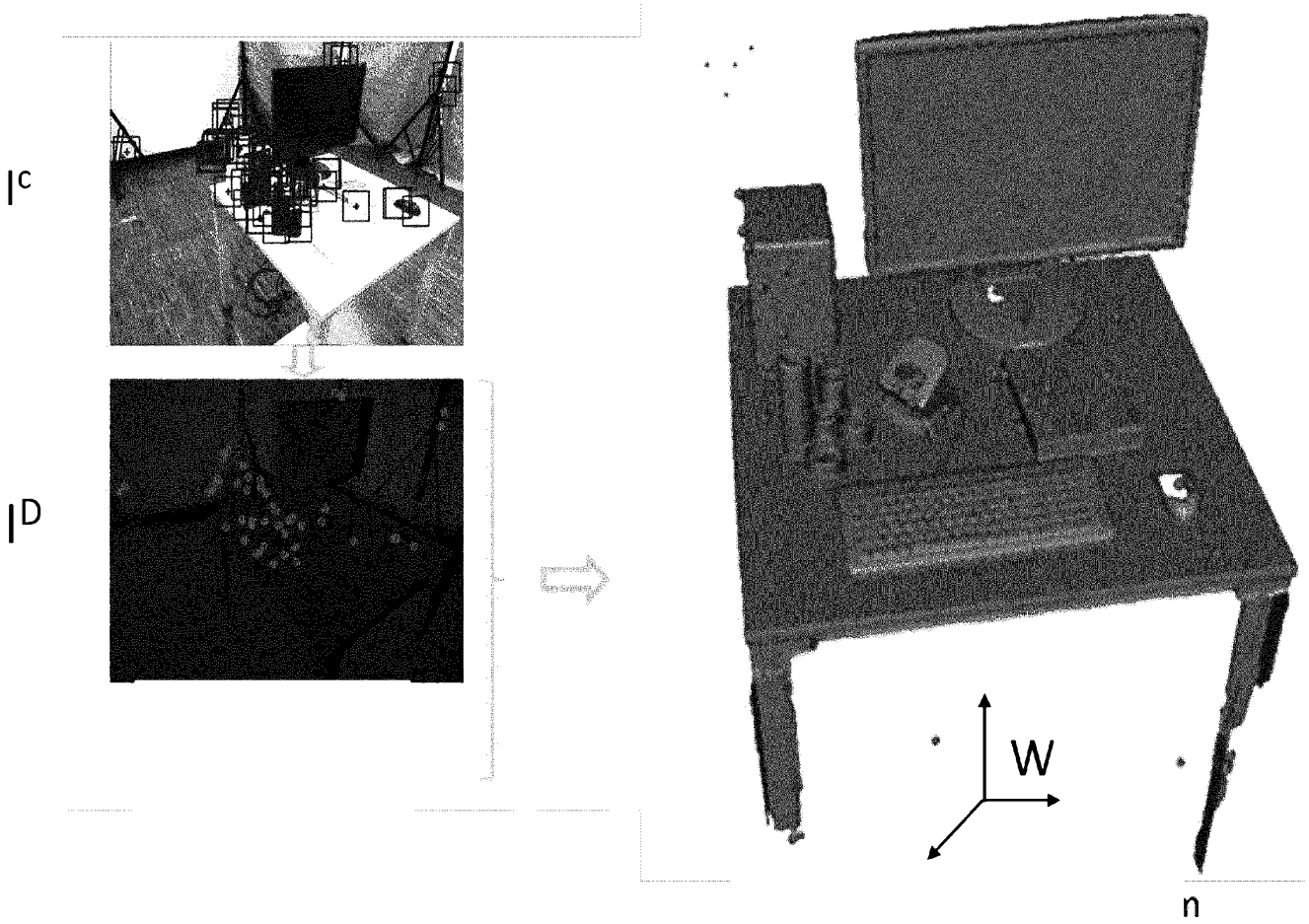


FIG. 2

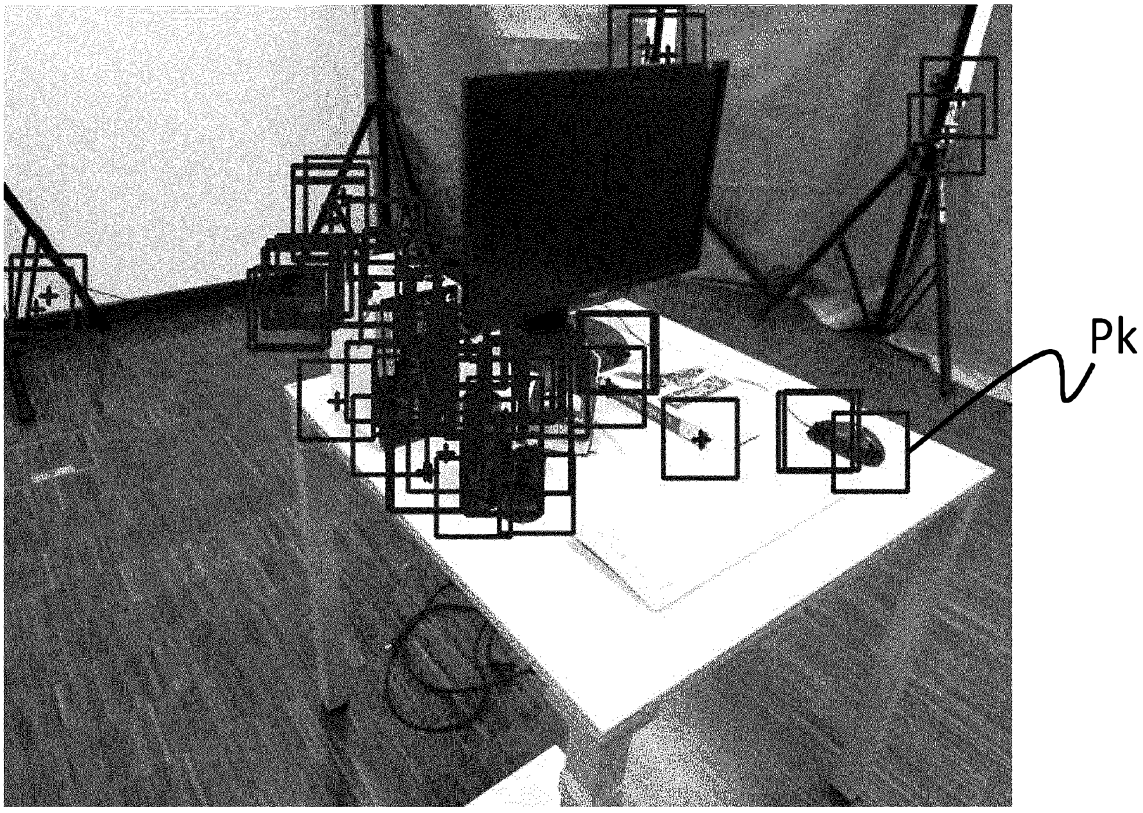


FIG. 3

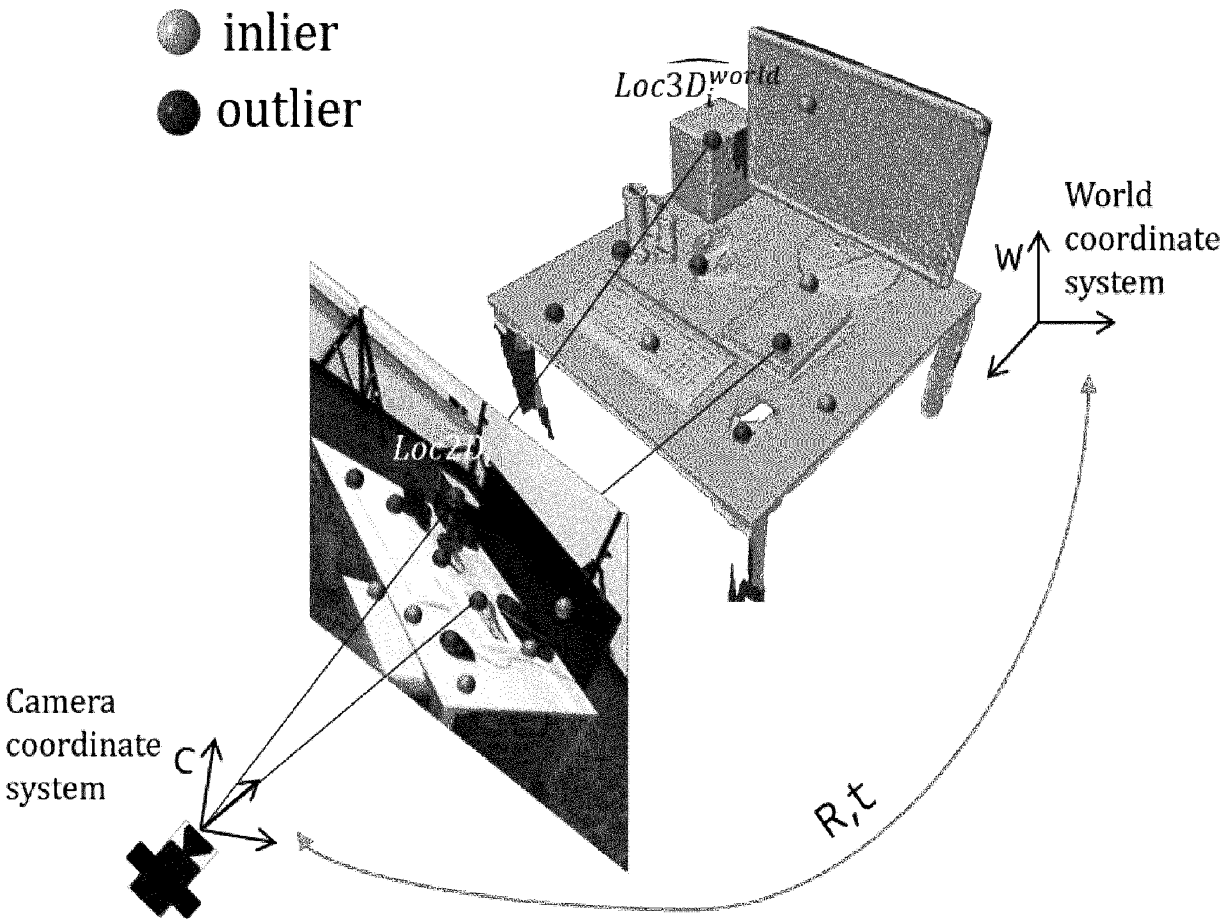

inlier
outlier
$\widehat{Loc3D}_i^{world}$
World
coordinate
system
W
$Loc2D_i$
Camera
coordinate
system
C
R,t

FIG. 4

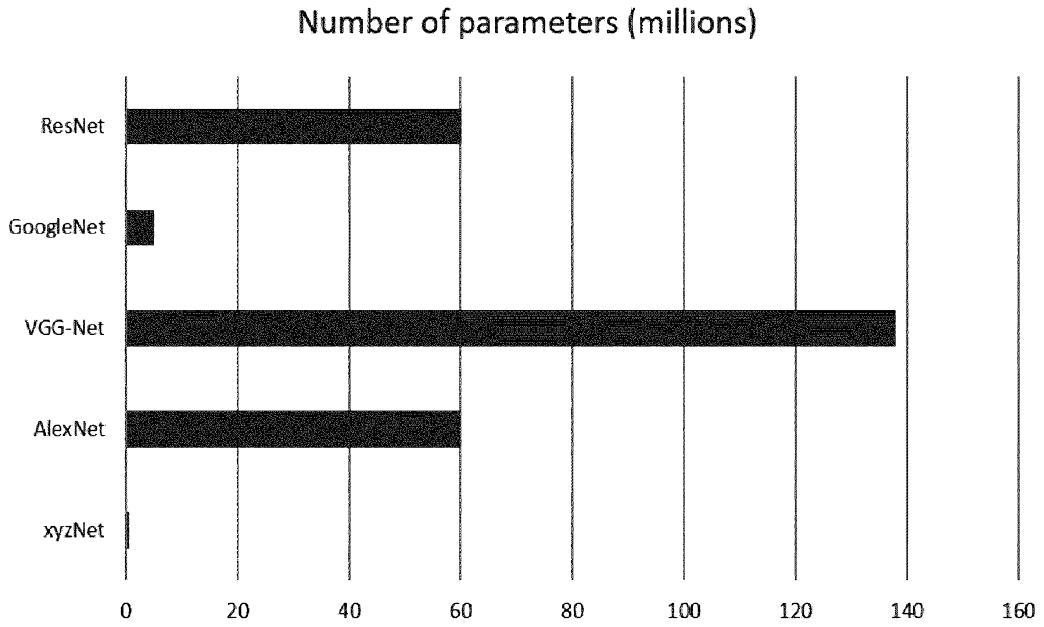

Number of parameters (millions)
ResNet
GoogleNet
VGG-Net
AlexNet
xyzNet
0
20
40
60
80
100
120
140
160

FIG. 5

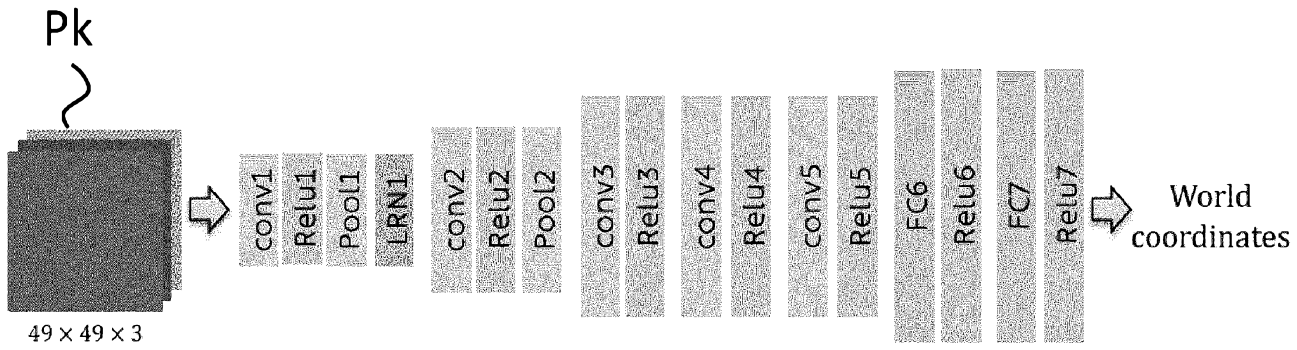


FIG. 6

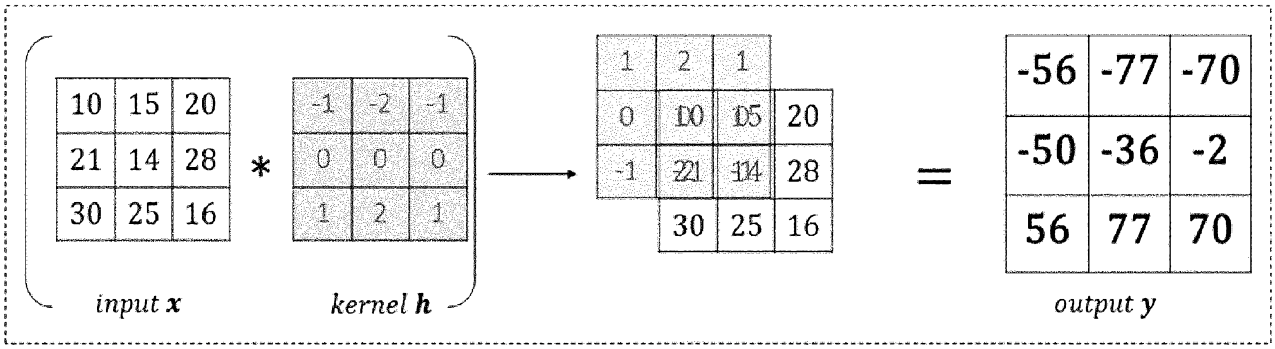


FIG. 7

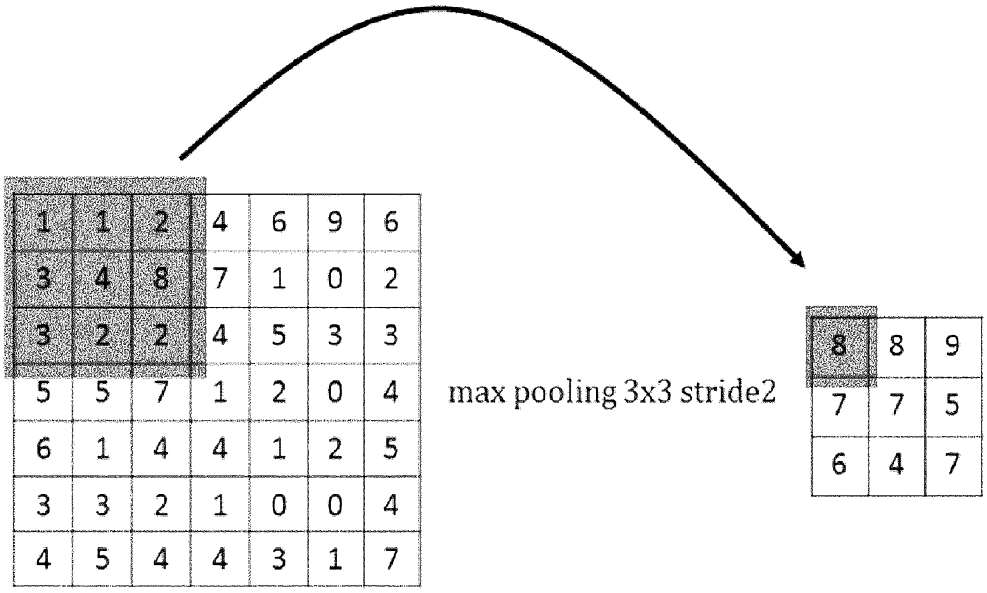


FIG. 8

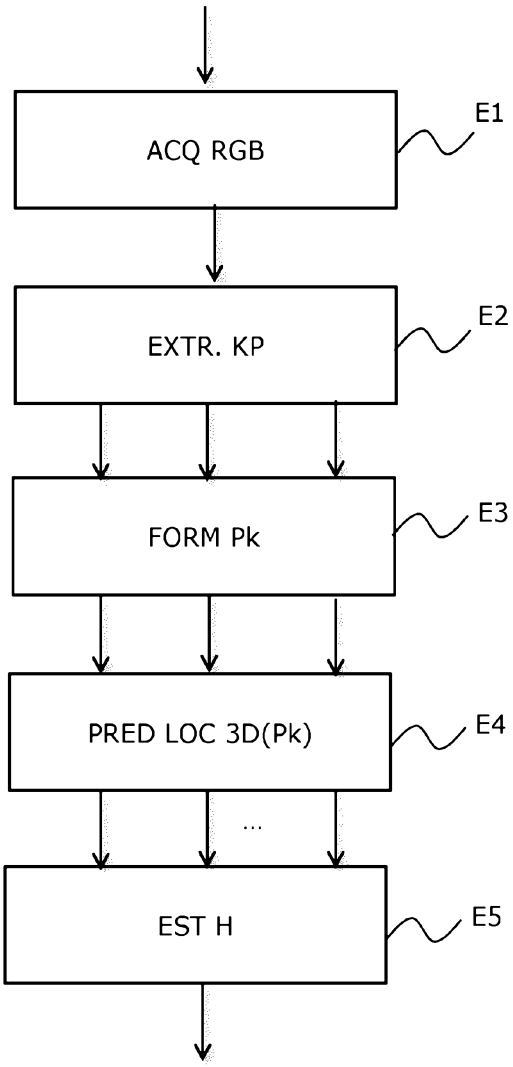

ACQ RGB
E1
EXTR. KP
E2
FORM Pk
E3
PRED LOC 3D(Pk)
E4
...
EST H
E5

FIG. 9

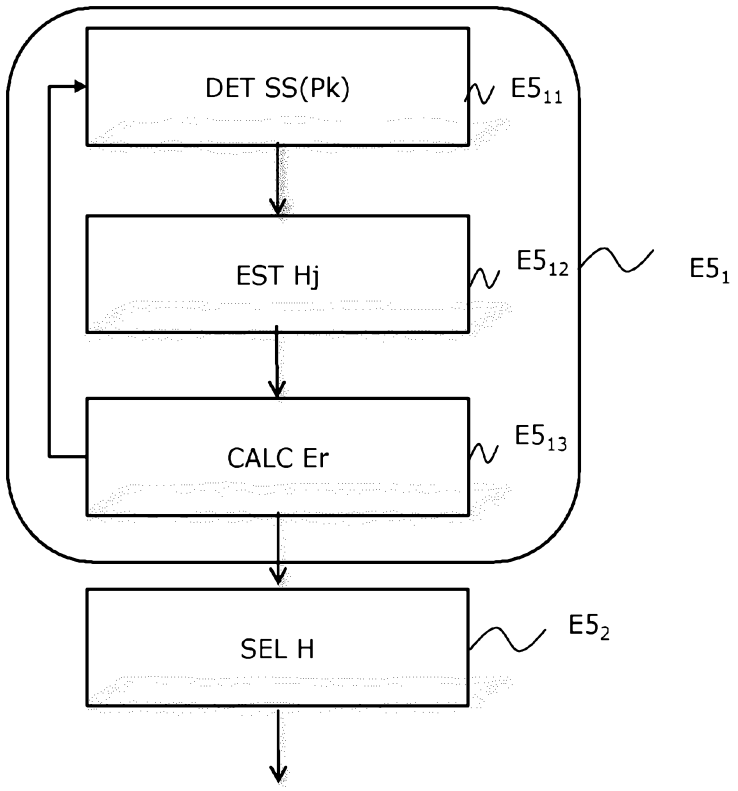

DET SS(Pk)
E5$_{11}$
EST Hj
E5$_{12}$
E5$_1$
CALC Er
E5$_{13}$
SEL H
E5$_2$

FIG. 10

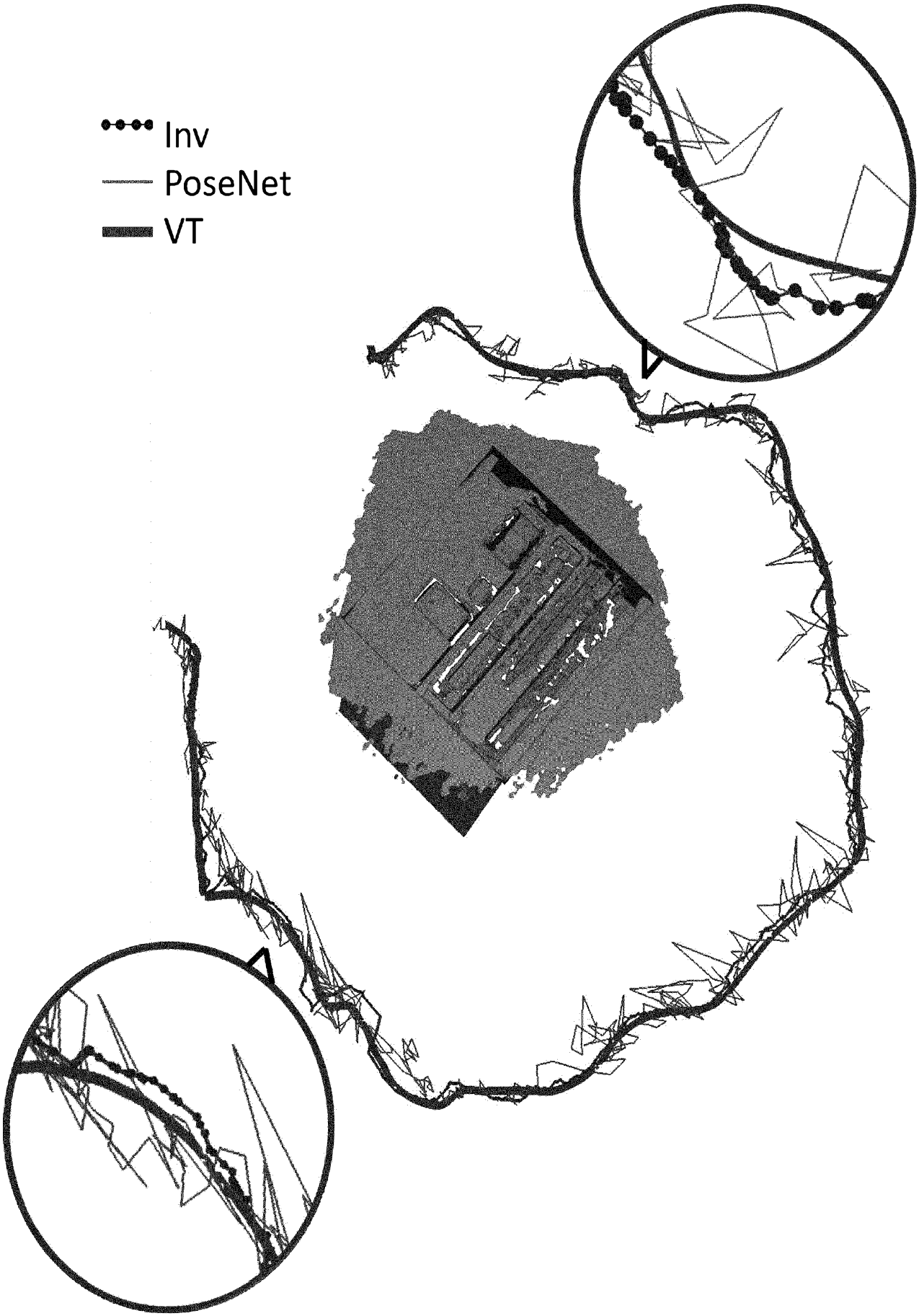


FIG. 11

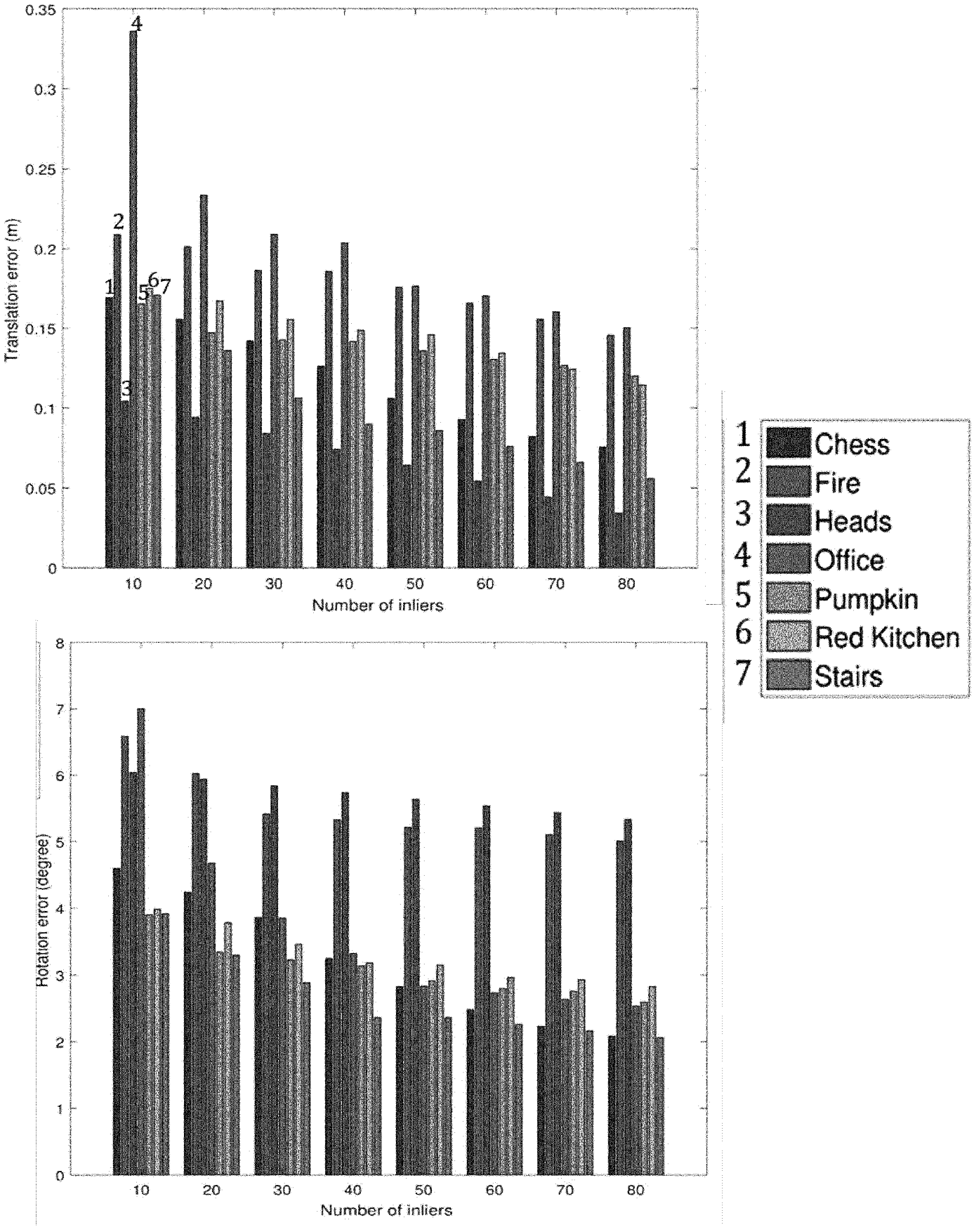
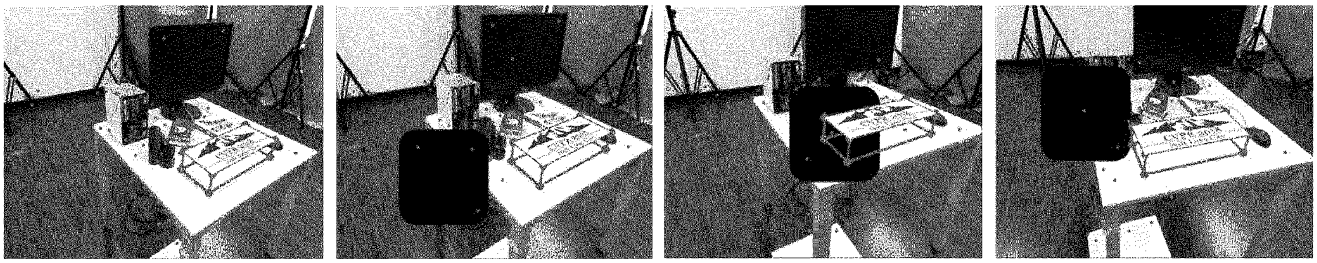

Translation error (m)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10
20
30
40
50
60
70
80
Number of inliers
1
2
3
4
5
6
7
Rotation error (degree)
8
7
6
5
4
3
2
1
0
10
20
30
40
50
60
70
80
Number of inliers
1 Chess
2 Fire
3 Heads
4 Office
5 Pumpkin
6 Red Kitchen
7 Stairs

FIG. 12



a) Frame #001 b) Frame #005 c) Frame #050 d) Frame #100

FIG. 13

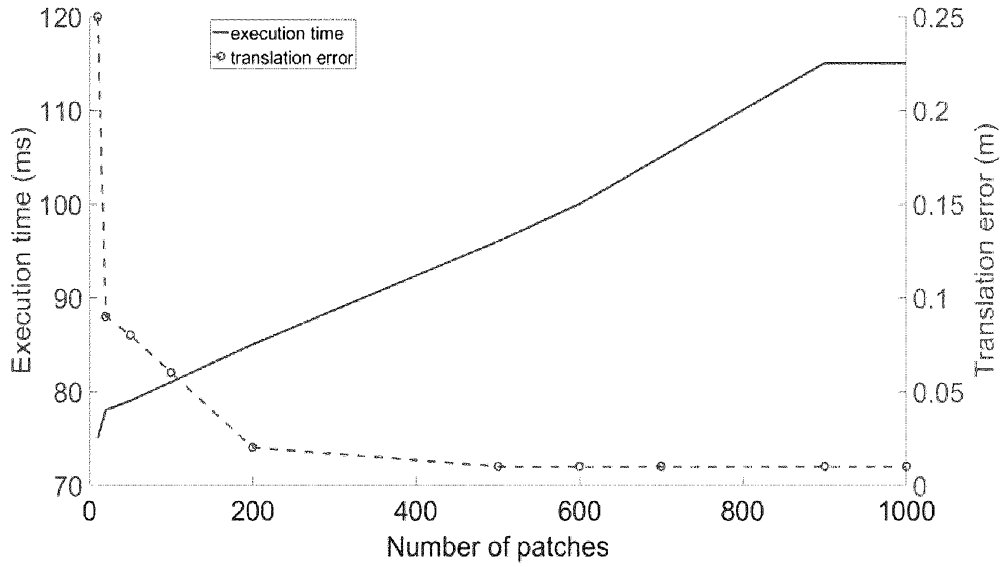


FIG. 14

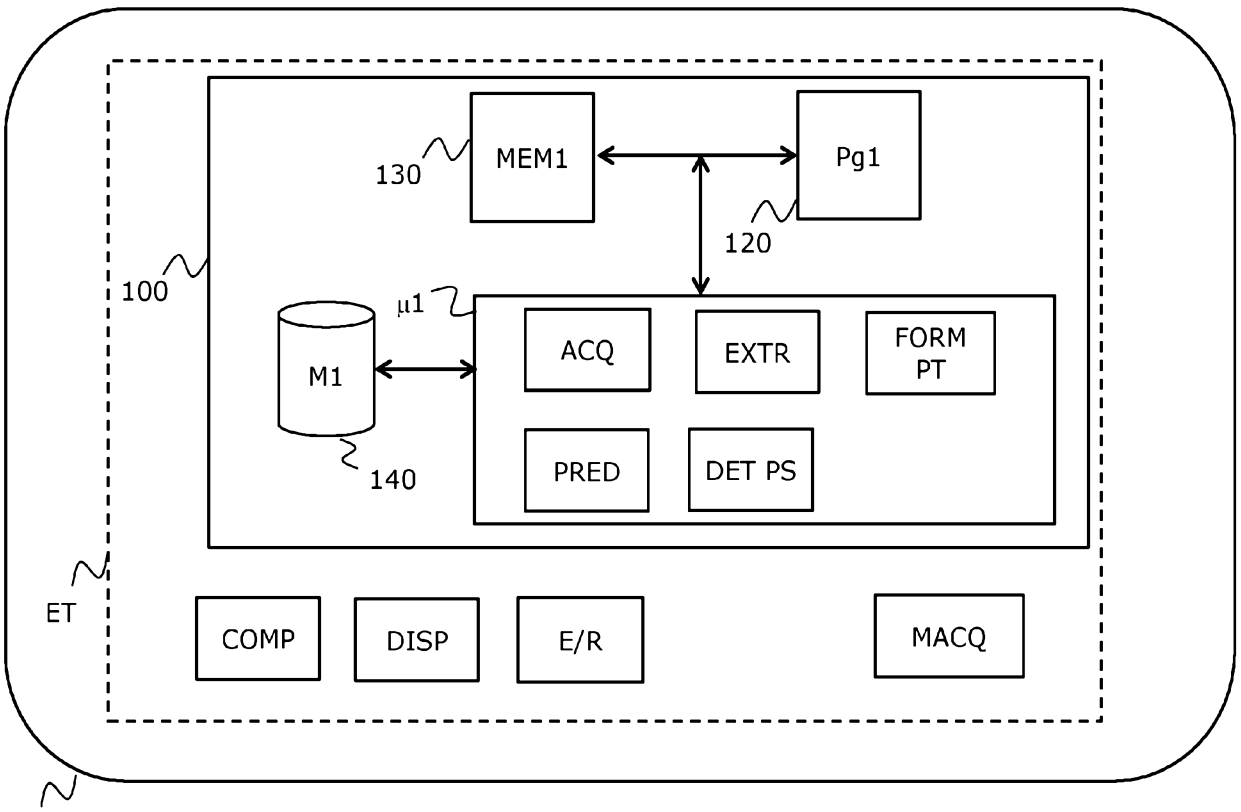


FIG. 15

RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications déposées avant le commencement de la recherche

N° d'enregistrement national

FA 846486
FR 1760541

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
X	TORSTEN SATTLER ET AL: "Fast image-based localization using direct 2D-to-3D matching", 2013 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 1 novembre 2011 (2011-11-01), pages 667-674, XP055298542, ISSN: 1550-5499, DOI: 10.1109/ICCV.2011.6126302	1-6,8-11	G06T7/80 G06K9/46 G06N3/08
Y	* abrégé * * figures 1,2 * * sections 1-3 *	7	

X	MENG LILI ET AL: "Backtracking regression forests for accurate camera relocalization", 2017 IEEE/RSJ INTERNATIONAL CONFERENCE ON INTELLIGENT ROBOTS AND SYSTEMS (IROS), IEEE, 24 septembre 2017 (2017-09-24), pages 6886-6893, XP033266762, DOI: 10.1109/IROS.2017.8206611 [extrait le 2017-12-13]	1-6,8-11	DOMAINES TECHNIQUES RECHERCHÉS (IPC) G06T
Y	* abrégé * * sections I, II, III, IV-B *	7	

Y	Eric Brachmann ET AL: "DSAC - Differentiable RANSAC for Camera Localization", arXiv, 7 août 2017 (2017-08-07), pages 1-11, XP055456666, Extrait de l'Internet: URL:https://arxiv.org/pdf/1611.05705 [extrait le 2018-03-06]	7	
	* abrégé * * sections 1, 3, 4.1 * * figure 2 *		

	-/--		

Date d'achèvement de la recherche	Examinateur
9 mars 2018	Eveno, Nicolas

CATÉGORIE DES DOCUMENTS CITÉS

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure.
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

1

EPO FORM 1503 12.99 (P04C14)

RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications déposées avant le commencement de la recherche

N° d'enregistrement national

FA 846486
FR 1760541

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
A	DAVID G. LOWE: "Distinctive Image Features from Scale-Invariant Keypoints", INTERNATIONAL JOURNAL OF COMPUTER VISION, vol. 60, no. 2, 5 janvier 2004 (2004-01-05), pages 91-110, XP055203065, ISSN: 0920-5691, DOI: 10.1023/B:VISI.0000029664.99615.94 * abrégé * * sections 4, 6 * -----	1-11	
A	IRSCHARA A ET AL: "From structure-from-motion point clouds to fast location recognition", 2009 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION : CVPR 2009 ; MIAMI [BEACH], FLORIDA, USA, 20 - 25 JUNE 2009, IEEE, PISCATAWAY, NJ, 20 juin 2009 (2009-06-20), pages 2599-2606, XP031607114, ISBN: 978-1-4244-3992-8 * le document en entier * -----	1-11	
			DOMAINES TECHNIQUES RECHERCHÉS (IPC)

Date d'achèvement de la recherche	Examinateur
9 mars 2018	Eveno, Nicolas

CATÉGORIE DES DOCUMENTS CITÉS

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure.
D : cité dans la demande
L : cité pour d'autres raisons
.....
& : membre de la même famille, document correspondant

1

EPO FORM 1503 12.99 (P04C14)