

12 DEMANDE DE BREVET D'INVENTION

A1

22 Date de dépôt : 07.03.23.

30 Priorité :

43 Date de mise à la disposition du public de la
demande : 13.09.24 Bulletin 24/37.

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) : SAFRAN SA — FR.

72 Inventeur(s) : MAALOUF, Aldo.

73 Titulaire(s) : SAFRAN SA.

74 Mandataire(s) : GEVERS & ORES.

54 PROCÉDE D'ESTIMATION D'UNE CARTE DE PROFONDEUR D'UNE IMAGE MONOCULAIRE ET DISPOSITIFS ASSOCIES.

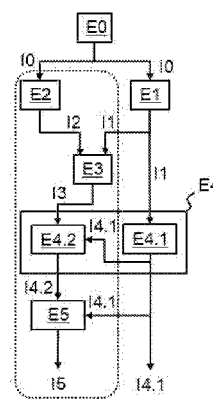
57 Procédé (E100) d'estimation d'une carte de profondeur d'une image monoculaire (I0), comprenant les étapes suivantes:

a) fournir (E0) une image monoculaire en couleurs (I0) acquise par un dispositif de captation optique monoculaire (0) dans des conditions d'observations réelles;

b) réaliser une correction colorimétrique (E1) de l'image monoculaire (I0) à partir d'un modèle d'apparence couleur inversible (1a), de manière à fournir une image corrigée en couleurs (I1) dans des conditions d'observation de référence, de sorte que la perception des couleurs dans l'image corrigée en couleurs (I1) soit identique à la perception des couleurs de l'image monoculaire en couleurs (I0) acquise dans les conditions d'observations réelles; et

c) traiter (E4; E4.1) l'image corrigée en couleurs (I1) en appliquant un réseau de neurones convolutif (4; 4.1; 4.1, 4.2) de manière à générer une carte de profondeur (I4.1; I5) de l'image monoculaire en couleurs (I0).

Figure pour l'abrégé : figure 1



E100



Description

Titre de l'invention : PROCEDE D'ESTIMATION D'UNE CARTE DE PROFONDEUR D'UNE IMAGE MONOCULAIRE ET DISPOSITIFS ASSOCIES

Domaine technique

- [0001] La présente invention concerne l'estimation de la profondeur d'une image monoculaire, i.e. une image acquise par un dispositif de captation optique ne disposant que d'un seul oculaire, telle qu'une caméra ou une lunette de vision monoculaire.
- [0002] L'invention concerne plus particulièrement un procédé d'estimation d'une carte de profondeur d'une image monoculaire en couleurs et un dispositif pour la mise en œuvre du procédé d'estimation.
- [0003] L'invention trouve une application privilégiée dans le domaine de la vision par ordinateur dite « vision artificielle », notamment dans le cadre de la navigation d'engins de mobilité autonomes, tels que des aéronefs, drones, véhicules ou robots.
- [0004] L'invention trouve également une application pour la reconstruction d'une scène en trois dimensions, à partir d'images monoculaires en deux dimensions.

Arrière-plan technologique

- [0005] L'estimation de la profondeur d'une image monoculaire, à partir d'une seule image en couleurs, est une tâche difficile de vision par ordinateur pour permettre aux ordinateurs de comprendre des scènes réelles.
- [0006] Pour chaque image monoculaire captée, la profondeur désigne la distance séparant le dispositif de captation (e.g. une caméra) d'un objet identifié sur cette image. Pour simplifier, la profondeur d'une image monoculaire sera désignée ci-après par l'expression « profondeur monoculaire » (de l'anglais « monocular depth »). Une carte de profondeur de l'image monoculaire représente les différents niveaux de profondeur présents dans l'image.
- [0007] De nombreuses techniques d'intelligence artificielle, notamment par apprentissage profond (de l'anglais « deep learning ») mis en œuvre par un réseau de neurones convolutif (CNN : « Convolutional Neural Network » en anglais) ont été proposées pour estimer automatiquement des cartes de profondeur monoculaire. De manière générale, ces techniques peuvent être regroupées en deux catégories principales : celles basées sur de l'apprentissage supervisé et celles basées sur de l'apprentissage auto-supervisé (i.e. non supervisé par un être humain).
- [0008] Pour l'apprentissage supervisé, des données de profondeur dites « vérité terrain » (de l'anglais « ground truth ») sont requises en entrée du réseau de neurones pour l'entraîner afin qu'il soit configuré pour estimer automatiquement la profondeur mo-

noculaire. Ces données d'entraînement sont fournies par des capteurs optoélectroniques coûteux, par exemple de type LiDAR (« Laser imaging Detection And Ranging » en anglais). Par conséquent, l'obtention de données significatives par de tels capteurs est une opération généralement coûteuse et difficile. Par ailleurs, ces données sont souvent creuses comparées à celles de l'image d'origine au niveau pixelique. Par exemple, dans les zones à faible luminosité, la capacité du LiDAR à fournir des données exploitables est largement affectée. Par conséquent, les performances obtenues par le réseau de neurones entraîné avec ces données sont insuffisantes. Enfin, lorsque le modèle du réseau de neurones ainsi entraîné est appliqué à des images provenant d'une autre caméra monoculaire ou lorsque ce modèle est utilisé pour prédire la profondeur d'une nouvelle scène dans des conditions d'observations différentes, la fiabilité de l'estimation de la profondeur fournie par ce réseau de neurones n'est pas garantie et peut être fortement réduite.

- [0009] Un exemple de méthode d'apprentissage supervisé pour prédire la profondeur monoculaire est décrite dans l'article de D. Eigen, C. Puhrsch, et R. Fergus, intitulé « Depth map prediction from a single image using a multi-scale deep network » et publié dans *Proceedings of the Advances in Neural Information Processing Systems* 27, 2014, pp. 2366-2374. Cet article décrit l'utilisation d'un réseau de neurones convolutif comprenant un sous-réseau à grande échelle configuré pour effectuer une première estimation de la profondeur de la scène à un niveau global en utilisant l'intégralité de l'image et un sous-réseau à petite échelle configuré pour effectuer une deuxième estimation permettant d'affiner l'estimation globale de la profondeur en utilisant des informations locales en se limitant à des régions de l'image. Toutefois, en tant que méthode supervisée, ce réseau de neurones requiert pour son entraînement des jeux de données annotées, ce qui est coûteux et contraignant.
- [0010] Pour s'affranchir de ces inconvénients, les méthodes d'apprentissage auto-supervisé ont été proposées pour estimer la profondeur monoculaire. Ces méthodes utilisent généralement des contraintes géométriques de paires d'images stéréoscopiques (ou de séquences d'images monoculaires) comme unique source de supervision des méthodes d'apprentissage. De telles méthodes d'apprentissage nécessitent l'obtention de paires d'images stéréoscopiques, par exemple par deux caméras observant la scène selon des angles de prise différents ou une caméra stéréoscopique (e.g. caméra RGB-D). En cas d'utilisation d'une seule caméra monoculaire, l'algorithme d'apprentissage auto-supervisé est configuré pour générer, à partir d'une seule image monoculaire acquise par la caméra, une image sous un autre angle de vue que l'image d'entrée, par un procédé de synthèse de vue, de manière à fournir une paire d'images stéréoscopiques.
- [0011] Un exemple de méthode auto-supervisée est décrit dans l'article de R. Garg, B.G.V. Kumar, G. Carneiro et I.D. Reid, intitulé « Unsupervised CNN for Single View Depth

Estimation: Geometry to the Rescue », publié dans *Proceedings of the 14th European Conference Computer Vision–ECCV 2016*, Volume 9912, pp. 740–756. Cette méthode utilise un réseau de neurones auto-encodeur configuré pour prédire la profondeur de l'image gauche d'une paire d'images stéréoscopiques.

- [0012] L'apprentissage auto-supervisé du réseau pour l'estimation de la profondeur se fait sur la base de paires d'images stéréoscopiques en minimisant une fonction d'erreur. Une image gauche reconstruite est synthétisée, par exemple par une contrainte géométrique épipolaire (de l'anglais « epipolar geometry constraint »), en utilisant l'image droite de la paire d'images stéréoscopiques. L'erreur photométrique entre l'image gauche et l'image gauche synthétisée est utilisée comme une fonction d'erreur pour entraîner le réseau auto-encodeur. Ainsi, l'estimation de la profondeur monoculaire revient à un problème d'appariement stéréoscopique de vue gauche-droite avec une perte photométrique gauche-droite à minimiser.
- [0013] Pour améliorer l'estimation de la profondeur monoculaire, la méthode de Godard *et al.* présentée ci-dessus a été modifiée en introduisant une fonction d'erreur de cohérence de profondeur entre les images gauche et droite de la paire d'images stéréoscopiques, ainsi qu'une fonction de perte de lissage des disparités entre les images gauche et droite, en tant que contraintes pour effectuer la reconstruction de l'image gauche. Cette amélioration est décrite dans l'article de C. Godard, O.M. Aodha et G.J. Brostow, intitulé « Unsupervised Monocular Depth Estimation with Left-Right Consistency », publié dans *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017*; pp. 6602–6611.
- [0014] Pour optimiser la qualité de l'estimation de la profondeur monoculaire, d'autres approches ont été développées, telle que la génération de vues virtuelles à trois dimensions à partir de paires d'images stéréoscopiques, comme décrite dans l'article de M. Poggi, F. Tosi et S. Mattoccia, intitulé « Learning monocular depth estimation with unsupervised trinocular assumptions », publié dans *Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018*; pp. 324–333. *Appl. Sci.* 2021, 11, 5383 15 of 15.
- [0015] Cette approche peut être combinée à l'utilisation de différents types de réseaux de neurones tels que :
 - des réseaux de contradiction générative (de l'anglais « Generative Adversarial Networks ») tels que décrits dans l'article de A. Pilzer, D. Xu, M. Puscas, E. Ricci, N. Sebe, intitulé « Unsupervised adversarial depth estimation using cycled generative networks », publié dans *Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018*; pp. 587–595 ;

- des réseaux sensibles au contexte, tels que décrit dans l'article de M. Zhai, X. Xiang, N. Lv, X. Kong, A. El Saddik, intitulé « An object context integrated network for joint learning of depth and optical flow », publié dans *IEEE Trans. Image Process.* 2020, 29, 7807–7818.
- [0016] Certes, les méthodes d'apprentissage auto-supervisé sont plus simples et moins coûteuses à mettre en œuvre que les méthodes d'apprentissage supervisé étant donné qu'elle ne nécessitent pas de marquage de données (« data labelling » en anglais). Elles présentent l'avantage de bénéficier d'une plus forte capacité de généralisation par rapport à différentes caméras et/ou différentes conditions d'observations (e.g. angle de prise de vue, luminosité ambiante, etc).
- [0017] Toutefois, en dépit des améliorations décrites ci-dessus, la précision et la fiabilité de la profondeur monoculaire estimée par les méthodes d'apprentissage auto-supervisé restent réduites par rapport à celles fournies par les méthodes d'apprentissage supervisé.
- [0018] Il est connu que cette perte de fiabilité et/ou précision réside dans une faible supervision de la cohérence photométrique (« photometric consistency » en anglais) des pixels entre l'image monoculaire d'entrée, acquise sous un angle de prise de vue donné, et l'image synthétisée à partir de l'image d'entrée par le procédé de synthèse sous un autre angle de prise de vue, de manière à générer la paire d'images stéréoscopiques en tant que signal de supervision.
- [0019] La plupart des méthodes d'estimation de profondeur monoculaire par apprentissage auto-supervisé reposent sur l'hypothèse de la constance des couleurs, selon laquelle les variations des couleurs perçues restent inchangées quelles soient les conditions d'observations, telles que l'angle de prise de vue de la caméra.
- [0020] Une telle hypothèse suppose que l'image d'entrée et l'image synthétisée (construite à partir de l'image d'entrée) sont identiques et ont les mêmes valeurs colorimétriques (e.g. RGB ou XYZ) à chaque pixel, si bien que l'erreur photométrique entre ces images est nulle. Or, il est connu qu'en réalité des fluctuations de couleurs de l'image peuvent être induites en raison de nombreux facteurs, tels que l'intensité lumineuse ambiante, le bruit et la réflexion de la lumière, ce qui peut provoquer des erreurs dans l'estimation automatique de la profondeur monoculaire. Sachant que ces facteurs ne sont pas nécessairement les mêmes pour deux angles de prise de vue différents, les couleurs perçues dans une même région de l'image peuvent fluctuer en fonction de la prise de vue.
- [0021] Par conséquent, la disparité des couleurs perçues qui peut exister dans une même zone d'image pour une paire d'images stéréoscopiques (ou de manière équivalente pour deux images monoculaires prises dans des conditions d'observations différentes) est source d'erreur dans l'estimation de la profondeur monoculaire.

- [0022] Pour tenir compte de l'incohérence photométrique due à des changements de luminosité, il a été proposé d'intégrer dans la fonction de perte, le paramètre SAD (de l'anglais « Sum of Absolute Difference ») ou le paramètre SSIM (de l'anglais « Structural Similarity Index Measure »). En général, le paramètre SSIM est privilégié dans la mesure où il prend en compte une plus grande quantité d'informations locales de l'image que le paramètre SAD, ce qui permet des performances de prédiction de la profondeur monoculaire accrues.
- [0023] Dans l'article de E. Guo, Z. Chen, Y. Zhou, D. O. Wu, intitulé « Unsupervised Learning of Depth and Camera Pose with Feature Map Warping », publié dans *Sensors* 2021, 21(3), 923, une fonction de perte pyramidale dite FPML (de l'anglais « Feature Pyramid Matching Loss ») est proposée pour répondre au problème d'incohérence photométrique causé par des changements de luminosité. La fonction FPML est plus robuste que les fonctions d'erreurs basées sur les paramètres SAD ou SSIM, dans la mesure où elle exploite des informations locales et globales des images. Toutefois, le problème principal de cette méthode est que la perte est calculée en combinant un réseau d'estimation de profondeur avec un réseau d'estimation de pose de la caméra, ce qui conduit à une perte de l'information sur l'échelle absolue dans les prédictions.
- [0024] Une autre raison de dégradation des performances de l'estimation monoculaire effectuée par les réseaux de neurones convolutifs à apprentissage auto-supervisé provient du fait que les images traitées ont une faible résolution et/ou une faible qualité (e.g. image bruitée et/ou faiblement texturée).
- [0025] Les caméras RGB monoculaires commercialement disponibles peuvent générer du bruit de compression ayant pour effet d'augmenter le rapport signal sur bruit et par conséquent de réduire la qualité de l'image fournie. Par ailleurs, pour des raisons d'optimisation des ressources et/ou de coût, la résolution de la caméra peut être volontairement réduite.
- [0026] Certaines régions de l'image acquise par la caméra peuvent être faiblement texturées. Ceci peut être dû à une faible résolution de la caméra et/ou aux conditions d'observations. Par exemple, dans des conditions de luminosité particulières (e.g. faible ou fort ensoleillement), la texture de certaines régions de l'image peut être très affaiblie voire quasiment invisible, même en utilisant des caméras de haute résolution. Dans ce cas, il est quasiment impossible pour les réseaux de neurones convolutifs auto-supervisés de l'art antérieur de prédire correctement la valeur de la disparité des pixels (e.g. estimation de la fonction d'erreur) en raison du nombre réduit de caractéristiques qu'ils peuvent extraire à partir de ces régions.

Document de l'art antérieur le plus proche

- [0027] L'article de N. Yang, L.V. Stumberg, R. Wang, D. Cremers, intitulé "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry", publié dans

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1281-1292, décrit une procédé d'estimation automatique d'une profondeur d'une image monoculaire en niveaux de gris.

- [0028] Ce document décrit un ajustement de la luminosité globale de l'image dû à des changements d'exposition de la caméra pour permettre un apprentissage auto-supervisé basé sur l'hypothèse de la constance de luminosité décrite ci-avant.
- [0029] Cet ajustement est effectué par un modèle de transformation linéaire, selon lequel la luminosité ambiante I de l'image est ajustée selon une fonction affine L^{ab} , telle que $L^{ab}(I) = a.I + b$. Toutefois, cette transformation porte uniquement sur la luminosité ambiante de l'image, i.e. une intensité lumineuse globale qui ne tient pas compte de variations locales de l'intensité lumineuse, ce qui a pour effet de limiter les performance de prédiction de la profondeur monoculaire.
- [0030] Ce procédé n'est pas adapté pour estimer la profondeur d'une image monoculaire en couleurs.
- [0031] Il peut ainsi être souhaité de prévoir un procédé d'estimation d'une profondeur d'une image monoculaire en couleurs qui permette de s'affranchir d'au moins une partie des problèmes et contraintes précités.

Résumé de l'invention

- [0032] Il est donc proposé un procédé d'estimation d'une carte de profondeur d'une image monoculaire, ledit procédé comprend les étapes suivantes :
 - fournir une image monoculaire en couleurs acquise par un dispositif de captation optique monoculaire dans des conditions d'observations réelles ;
 - réaliser une correction colorimétrique de l'image monoculaire à partir d'un modèle d'apparence couleur inversible, de manière à fournir une image corrigée en couleurs dans des conditions d'observation de référence, de sorte que la perception des couleurs dans l'image corrigée en couleurs soit identique à la perception des couleurs de l'image monoculaire en couleurs acquise dans les conditions d'observations réelles; et
 - traiter l'image corrigée en couleurs en appliquant un réseau de neurones convolutif de manière à générer une carte de profondeur de l'image monoculaire en couleurs.
- [0033] La correction colorimétrique de l'image monoculaire en couleurs a pour effet que la perception des couleurs de l'image reste inchangée quelles que soient les conditions d'observations réelles, telles que les conditions d'éclairage (e.g luminosité ambiante de la scène, nature lumineuse de l'environnement, intensité de fond), les caractéristiques photométriques de la caméra (e.g. coordonnées chromatiques du système RGB de la caméra) et les paramètres de prise de vue, tels que l'angle de prise de vue.

- [0034] Il s'ensuit que la fiabilité de la carte de profondeur est accrue par rapport à celles des méthodes auto-supervisées de l'art antérieur qui ne procèdent à aucune correction colorimétrique de l'image monoculaire.
- [0035] Ainsi, le procédé d'estimation de la profondeur monoculaire selon l'invention est plus robuste que ceux de l'art antérieur, dans le sens où la carte de profondeur n'est pas affectée par les conditions particulières dans lesquelles la scène est observée.
- [0036] L'invention peut en outre comporter l'une ou plusieurs des caractéristiques optionnelles suivantes, selon toute combinaison techniquement possible.
- [0037] L'image monoculaire en couleurs comprend une pluralité de pixels et la correction colorimétrique comprenant les étapes suivantes :
- appliquer le modèle d'apparence couleur inversible à chaque pixel de l'image monoculaire en couleurs, de manière à déterminer au moins un attribut perceptuel du pixel sélectionné parmi : clarté, chromaticité, angle de teinte, brillance, saturation, composition de teinte ;
 - inverser le modèle d'apparence couleur inversible de manière à fournir un modèle d'apparence couleur inverse ; et
 - pour chaque pixel, appliquer ledit au moins un attribut perceptuel au modèle d'apparence couleur inverse configuré avec les conditions d'observation de référence, de manière à fournir des valeurs trichromatiques corrigées dudit pixel.
- [0038] La correction colorimétrique comprend en outre une estimation fréquentielle, lors de laquelle une fréquence spatiale de chaque pixel de l'image monoculaire en couleurs est estimée, la fréquence estimée étant fournie en entrée du modèle d'apparence couleur inverse pour obtenir les valeurs trichromatiques corrigées dudit pixel.
- [0039] L'estimation fréquentielle comprend une décomposition de l'image monoculaire en couleurs en une pluralité de domaines fréquents et une décomposition de l'image monoculaire en couleurs selon plusieurs orientations spatiales, la fréquence étant estimée en tenant compte de l'influence des domaines fréquents adjacents au domaine fréquentiel auquel appartient ledit pixel de manière à obtenir une carte spatio-fréquentielle des pixels.
- [0040] Le modèle d'apparence couleur inversible est le modèle CIECAM2 configuré avec selon au moins l'un des paramètres suivants : coordonnées chromatiques d'un blanc de référence, luminosité ambiante, intensité lumineuse de fond, indicateur d'environnement.
- [0041] Le procédé comprend en outre les étapes suivantes :
- filtrer l'image monoculaire en couleurs en appliquant un filtre quaternionique linéaire guidé par une image de référence, de manière à fournir une image quaternionique filtrée comprenant des données de textures de l'image mo-

- nculaire; et
 - soustraire l'image filtrée de l'image corrigée en couleurs, de manière à obtenir une carte de textures.
- [0042] Le procédé comprend en outre les étapes suivantes :
- sur-échantillonner la carte de profondeur, de résolution réduite, par application d'un réseau hiérarchique de manière à fournir des caractéristiques de profondeur sur-échantillonnées ; et
- [0043] concaténer la carte des textures et les caractéristiques de profondeur sur-échantillonnées pour fournir une carte de profondeur résiduelle ; additionner la carte de profondeur résiduelle et la carte de profondeur réduite, de manière à obtenir une carte de profondeur finale d'une résolution au moins égale à celle de l'image monoculaire. Une originalité de la présente invention est l'utilisation d'un modèle colorimétrique de référence, de préférence CIECAM2, pour assurer une invariance des couleurs perçues, couplée au filtrage quaternionique pour renforcer les textures de l'image monoculaire, permettant d'extraire à la fois la couleur illuminant chaque pixel et les textures dans les régions faiblement texturées ou bruitées.
- [0044] Le filtrage quaternionique a pour avantage d'améliorer la précision sur les couleurs de l'image filtrée, tout en optimisant le traitement d'image réalisé lors du filtrage en termes de ressources de mémoire et de capacité de calcul.
- [0045] Le procédé selon l'invention permet d'estimer la profondeur des images compressées de faible qualité (i.e. avec des niveaux de bruit élevée) et peu texturées. La carte de texture obtenue en soustrayant l'image corrigée en couleurs de l'image filtrée contient des hautes fréquences spatiales permettant de faciliter la localisation des limites de profondeur dans l'image monoculaire.
- [0046] L'invention vise aussi, dans au moins un mode de réalisation, un dispositif pour mettre en œuvre le procédé d'estimation tel que décrit ci-avant.
- [0047] Ce dispositif comprenant :
- un dispositif de captation optique monoculaire pour obtenir une image monoculaire en couleurs dans des conditions d'observations réelles ;
 - un module de correction colorimétrique de l'image monoculaire en couleurs basé sur un modèle d'apparence couleur inversible, de manière à fournir une image corrigée en couleurs dans des conditions d'observation de référence, de sorte que la perception des couleurs dans l'image corrigée en couleurs soit identique à la perception des couleurs dans l'image monoculaire en couleurs acquise dans les conditions d'observations réelles; et
 - un module de génération d'une carte de profondeur configuré pour appliquer un réseau de neurone convolutif à l'image corrigée en couleurs.
- [0048] Le dispositif selon l'invention peut en outre comporter l'une ou plusieurs des caracté-

ristiques optionnelles suivantes, selon toute combinaison techniquement possible :

- un module de filtrage quaternionique linéaire guidé par une image de référence configuré pour filtrer l'image monoculaire en couleurs de manière à obtenir une image filtrée comprenant des données de texture de l'image monoculaire en couleurs ; et
- un module de soustraction configuré pour soustraire l'image filtrée de l'image corrigée en couleurs.

[0049] L'invention vise aussi, dans au moins un mode de réalisation, un engin de mobilité autonome comprenant un dispositif selon l'invention tel que décrit ci-avant.

[0050] L'invention vise aussi un programme d'ordinateur téléchargeable depuis un réseau de communication et/ou enregistré sur un support lisible par ordinateur, caractérisé en ce qu'il comprend des instructions pour l'exécution des étapes d'un procédé d'estimation de profondeur d'une image monoculaire tel que décrit ci-avant, lorsque ledit programme est exécuté sur un ordinateur.

[0051] L'invention vise également un moyen de stockage d'informations, amovible ou non, partiellement ou totalement lisible par un ordinateur ou un microprocesseur comportant des instructions de code d'un programme d'ordinateur pour l'exécution des étapes du procédé d'estimation de profondeur d'une image monoculaire selon l'invention comme décrit ci-dessus.

Brève description des figures

[0052] L'invention sera mieux comprise à l'aide de la description qui va suivre, donnée uniquement à titre d'exemple et faite en se référant aux dessins annexés dans lesquels :

- [Fig.1] illustre schématiquement un procédé d'estimation selon un mode de réalisation de l'invention ;
- [Fig.2] illustre schématiquement un exemple de dispositif pour mettre en œuvre le procédé selon la [Fig.1] ;
- [Fig.3] illustre schématiquement un exemple de réalisation de l'étape de correction colorimétrique mise en œuvre dans le procédé d'estimation selon la [Fig.1] ;
- [Fig.4] illustre schématiquement un exemple de réalisation de l'étape de filtrage mise en œuvre dans le procédé d'estimation selon la [Fig.1] ;
- [Fig.5] illustre schématiquement un exemple de réalisation d'un dispositif pour mettre en œuvre le procédé selon l'invention ;
- [Fig.6] illustre un exemple d'architecture d'un dispositif adapté à mettre en œuvre le procédé selon un mode de réalisation de l'invention ;
- [Fig.7] illustre schématiquement un robot terrestre dans lequel le dispositif selon l'invention est mis en œuvre ; et

- [Fig.8] illustre schématiquement un drone dans lequel le dispositif selon l'invention est mis en œuvre.

Description détaillée de l'invention

- [0053] En référence à la [Fig.1], un procédé E100 d'estimation de la profondeur d'une image monoculaire va être à présent décrit selon un mode de réalisation de l'invention.
- [0054] Le procédé E100 comprend une étape d'obtention E0 d'une image monoculaire I0 en couleurs, par exemple selon le format RGB (de l'anglais « Red Green Blue »). Cette image monoculaire I0 est acquise dans des conditions d'observations dites réelles, par exemple au moyen d'une caméra monoculaire ou de tout autre dispositif optique monoculaire, tel qu'une lunette de vision monoculaire. Dans d'autres modes de réalisation (non illustrés), la caméra 0 sera configurée pour fournir une séquences d'images monoculaires.
- [0055] Le procédé E100 comprend en outre une étape de correction colorimétrique E1, lors de laquelle l'image monoculaire I0 est corrigée de manière à fournir une image corrigée en couleurs I1. Cette correction est telle que la perception des couleurs dans l'image corrigée I1 dans des conditions d'observations de référence soit identique à la perception des couleurs de l'image monoculaire I0 dans les conditions d'observations réelles. Cette étape de correction colorimétrique E1 sera décrite plus en détails en référence à la [Fig.3].
- [0056] Lors d'une étape de génération E4 et plus particulièrement lors d'une sous-étape E4.1, une carte de profondeur réduite I4.1 est générée à partir de l'image corrigée en couleurs I1. Cette carte de profondeur est qualifiée de réduite étant donné qu'elle dispose d'une résolution inférieure à celle de l'image monoculaire I0 ou de l'image corrigée en couleurs I1.
- [0057] Grâce à l'ajustement de colorimétrie opéré lors de l'étape de correction colorimétrique E1, la profondeur monoculaire estimée dans la carte de profondeur réduite I4.1 ne dépend pas des conditions d'observations réelles. Cette correction est particulièrement avantageuse pour assurer la cohérence des couleurs d'une paire d'images stéréoscopiques obtenue à partir de l'image monoculaire I0 pour l'apprentissage auto-supervisé de réseaux neuronaux.
- [0058] Le procédé E100 peut être complété par les étapes optionnelles suivantes regroupées sur la [Fig.1] par un cadre en traits pontillés.
- [0059] Lors d'une étape de filtrage E2, l'image monoculaire I0 est filtrée de manière à fournir une image filtrée I2 qui correspond à l'image monoculaire I0. Un des objectifs du filtrage est d'obtenir une image I2 en haute fréquence, dite image « passe haut », dans laquelle les textures et/ou les bordures de l'objet sont bien définies, i.e. mieux définies que dans l'image monoculaire d'origine I0. Une fois que les textures et éven-

tuellement les bordures des différents objets constitutifs de l'image sont obtenues, une meilleure estimation de la profondeur monoculaire peut être obtenue pour chaque objet dans l'image. Cette étape de filtrage E2 sera décrite plus en détails en référence à la [Fig.4].

- [0060] Lors d'une étape de soustraction E3, l'image corrigée en couleurs I1 est soustraite à l'image filtrée I2, de manière à obtenir une carte de textures I3 de l'image monoculaire I0. La carte de textures I3 est une image corrigée en couleurs dans laquelle les textures de l'image monoculaire I0 sont mises en évidence.
- [0061] Le procédé E100 comprend en outre une étape de génération E4 lors de laquelle au moins une carte de profondeur améliorée est déterminée. Plus particulièrement, cette étape E4 comprend une sous-étape E4.2 lors de laquelle une carte de profondeur résiduelle I4.2 est générée sur la base de la carte de profondeur réduite I4.1 et de la carte de textures I3. L'étape de génération E4 sera décrite plus en détails en référence à la [Fig.5].
- [0062] Le procédé E100 comprend en outre une étape d'addition E5, lors de laquelle la carte de profondeur réduite I4.1 est ajoutée à la carte de profondeur résiduelle I4.2, de manière à fournir une carte de profondeur finale I5 d'une résolution au moins égale à celle de l'image monoculaire I0 fournie lors de l'étape E0.
- [0063] En référence à la [Fig.2], un exemple de dispositif 100 pour mettre en œuvre le procédé selon la [Fig.1] tel que décrit ci-avant va être à présent décrit.
- [0064] Dans le présent exemple, le dispositif 100 selon l'invention comprend une caméra monoculaire 0 configurée pour acquérir l'image monoculaire I0. Comme indiqué précédemment, tout autre dispositif de captation monoculaire adapté à capter des images monoculaires pourra être utilisé à la place de la caméra. Dans le présent exemple, la caméra 0 fait partie du dispositif 100. Toutefois, dans d'autres variantes de réalisation (non illustrées), la caméra 0 ou tout autre dispositif de captation monoculaire pourra être extérieur au dispositif 100, auquel cas le dispositif 100 comprendra un moyen d'obtention de l'image monoculaire I0 préalablement acquise par la caméra 0.
- [0065] Le dispositif 100 comprend en outre un module de correction colorimétrique 1 de l'image monoculaire I0. Ce module 1 est configuré pour fournir une image corrigée en couleurs I1, de sorte que la perception des couleurs dans l'image corrigée en couleurs I1 dans des conditions d'observation de référence soit identique à la perception des couleurs de l'image monoculaire I0 dans les conditions d'observations réelles. Ce module de correction colorimétrique 1 sera décrit plus en détails en référence à la [Fig.3].
- [0066] Le dispositif 100 comprend en outre un module de filtrage 2 pour filtrer l'image monoculaire I0, de manière à fournir l'image filtrée I2. De préférence, le module de filtrage 2 est configuré pour mettre en œuvre un filtre quaternionique guidé qui sera

décrit plus en détails en référence à la [Fig.4].

- [0067] Le module de correction colorimétrique 1 et le module de filtrage 2 sont reliés à une même sortie de la caméra monoculaire 0, de manière à recevoir et traiter simultanément la même image monoculaire I0.
- [0068] Le dispositif 100 comprend en outre un module de soustraction 3 relié en sortie du module de correction colorimétrique 1 et du module de filtrage 2. Le module de soustraction 3 est configuré pour soustraire l'image corrigée en couleurs I1 et l'image filtrée I2, de manière à fournir la carte de textures I3, telle que $I3 = I2 - I1$. Un exemple de carte de textures I3 est illustré sur la [Fig.2], dans laquelle des textures et éventuellement des bordures de l'image monoculaire I0 ont été mises en évidence.
- [0069] La carte de textures I3 ainsi obtenue est une image filtrée, telle que les basses fréquences ont été supprimées, si bien que la carte de textures I3 comprend uniquement des composantes à hautes fréquences utiles pour estimer la profondeur de champ de l'image.
- [0070] Le dispositif 100 comprend en outre un réseau de neurones convolutif 4 comprenant un premier sous-réseau 4.1 pour estimer la profondeur de l'image monoculaire I0 et un deuxième sous-réseau 4.2 pour améliorer la profondeur estimée par le premier sous-réseau 4.1. Le premier sous-réseau 4.1 est configuré pour fournir la carte de profondeur réduite I4.1. Le deuxième sous-réseau 4.2 est configuré pour fournir la carte de profondeur résiduelle I4.2.
- [0071] Le premier sous-réseau 4.1 est relié en sortie du module de correction colorimétrique 1 de manière à recevoir en entrée l'image corrigée en couleurs I1. Le deuxième sous-réseau 4.2 est relié en sortie du module de soustraction 3, de manière à recevoir en entrée la carte de textures I3. Le deuxième sous-réseau 4.2 est relié au premier sous-réseau 4.1 de sorte que le premier sous-réseau 4.1 fournisse la carte de profondeur réduite I4.1 au deuxième sous-réseau 4.2.
- [0072] La mise en œuvre du réseau de neurones convolutif 4, et plus particulièrement du premier sous-réseau 4.1 et du deuxième sous-réseau 4.2 sera décrite plus en détails en référence à la [Fig.5].
- [0073] Le dispositif 100 comprend en outre un module d'addition 5 relié en sortie des premier et deuxième sous-réseau 4.1, 4.2. Le module d'addition 5 est configuré pour obtenir la carte de profondeur finale I5 en ajoutant la carte de profondeur réduite I4.1 à la carte de profondeur résiduelle I4.2.
- [0074] En référence à la [Fig.3], un exemple de mise en œuvre de l'étape de correction colorimétrique E1 par le module de correction colorimétrique 1 va être à présent décrit.
- [0075] Dans cet exemple, on suppose que l'image monoculaire I0 fournie par la caméra monoculaire 0 est encodée dans l'espace colorimétrique RGB.
- [0076] Lors d'une sous-étape de conversion E10, l'image monoculaire I0 est convertie dans

l'espace colorimétrique XYZ, de manière à fournir une image convertie I1.1 comprenant une pluralité de pixels P_{ij} en coordonnées XYZ.

- [0077] Cette conversion est effectuée à partir des coordonnées chromatiques du système RGB de la caméra 0 respectivement pour les composantes R, G et B : (x_R, y_R) , (x_G, y_G) , (x_B, y_B) et à partir des coordonnées chromatiques du niveau de blanc de référence (X_w, Y_w, Z_w) , selon les normes de la Commission Internationale de l'Eclairage (CIE). Cette conversion n'est bien évidemment pas requise si l'image monoculaire I0 est déjà fournie dans l'espace colorimétrique XYZ.
- [0078] Lors d'une sous-étape E11, des attributs perceptuels, tels que la clarté J, la chromaticité C, l'angle de teinte h, la brillance Q, la saturation s, la composition de teinte H sont déterminés en appliquant à l'image convertie I1.1, un modèle d'apparence couleur réversible (CAM : « Color Appearance Modelling » en anglais).
- [0079] De préférence, ce modèle est le modèle CIECAM02, tel que publié en 2022 par le comité technique 8-01 (« Color Appearance Modelling for Color Management Systems ») de la commission internationale de l'éclairage (CIE). Le modèle CIECAM02 est particulièrement adapté pour prendre en compte de manière fiable l'environnement d'un objet coloré afin de compenser l'influence de son environnement sur la perception de la couleur. Parmi les avantages du modèle CIECAM02 par rapport au modèle CIECAM97, on peut noter une transformée d'adaptation chromatique linéaire qui permet d'inverser le modèle plus facilement, et un facteur d'adaptation D simplifié.
- [0080] Les attributs perceptuels précités peuvent être calculés selon les standards CIE, e.g. selon le modèle CIECAM2, mais ne seront pas détaillés ici par souci de concision.
- [0081] Dans le présent exemple, le modèle CIECAM2 est configuré en utilisant au moins l'un des paramètres de configuration suivants : les coordonnées chromatiques du niveau de blanc de référence X_w, Y_w, Z_w , la luminosité (ou intensité lumineuse) ambiante L_a de l'image exprimée en lux ou cd/m^2 , une intensité lumineuse de l'arrière-plan Y_b de l'image, un indicateur d'éclairage moyen de l'environnement d'observation E (e.g. clair, sombre, moyen, lumière de jour standard).
- [0082] De préférence, tous ces paramètres sont pris en compte pour configurer le modèle dans la mesure où ils influencent la vision de la couleur (perception colorimétrique). Certains de ces paramètres de configuration sont fixés par l'utilisateur, tels que ceux relatifs à l'environnement d'observation E, la luminosité ambiante L_a , le blanc de référence X_w, Y_w, Z_w . Pour fixer les paramètres d'observation E, on pourra définir trois paramètres supplémentaires, tels que l'impact de l'environnement c, le facteur d'induction chromatique N_c , le facteur pour le degré d'adaptation F, en suivant les recommandations classique des modèles CIECAM97 et/ou CIECAM02. Tous ces paramètres peuvent être préalablement fixés lors de la sous-étape E11.

- [0083] D'autres paramètres de configuration du modèle peuvent être déterminés lors de la sous-étape E11 par le module de correction de colorimétrie 1, tels que la luminosité de l'arrière-plan de l'image Y_b . Pour chaque pixel P_{ij} , la valeur Y_b de ce paramètre est déterminée, en sélectionnant une fenêtre W_{ij} formée par un ensemble de pixels au voisinage du pixel P_{ij} , et en calculant E110 la moyenne des intensités des pixels appartenant à la fenêtre W_{ij} . De manière alternative, l'intensité Y_b est déterminée en utilisant une moyenne pondérée par une gaussienne. De préférence, la valeur de Y_b calculée est inférieure à 100, e.g. $Y_b=2$ pour le noir, $Y_b=20$ pour un gris moyen, $Y_b=90$ pour un blanc quasiment pur.
- [0084] Selon une particularité de l'invention, le modèle CIECAM2 référencé 1a est appliqué à chaque pixel P_{ij} de l'image I1.1. Ainsi, pour chaque pixel P_{ij} , les attributs perceptuels sont déterminés et fournis à l'issue de l'étape E11.
- [0085] De manière générale, les attributs perceptuels peuvent être sélectionnés parmi : la clarté J , la chromaticité C , la composition de teinte H (ou teinte en quadrature), l'angle de teinte ou teinte angulaire h (*hue* en anglais), la brillance G , la saturation s .
- [0086] Dans le présent exemple, les attributs perceptuels considérés pour chaque pixel P_{ij} sont la clarté J_{ij} , la chromaticité C_{ij} , l'angle de teinte h_{ij} . Dans d'autres variantes de réalisation, d'autres attributs perceptuels pourront être considérés en substitution ou en complément de ces paramètres.
- [0087] Lors d'une étape inverse E12, le modèle CIECAM2 1a est inversé. Cette opération est facilitée compte-tenu que le modèle utilise une transformation d'adaptation linéaire qui est elle-même facile à inverser. Le modèle inverse 1b est noté CIECAM2⁻¹.
- [0088] Lors de l'étape inverse E12, les attributs perceptuels J_{ij} , C_{ij} , h_{ij} précédemment calculés lors de l'étape E11 sont ensuite fournis en entrée du modèle inverse 1b, noté CIECAM2⁻¹, de manière à déterminer des valeurs trichromatiques X'_{ij} , Y'_{ij} , Z'_{ij} corrigées du pixel P_{ij} . Le modèle inverse 1b est configuré en appliquant des paramètres d'environnement E' , de luminosité ambiante L_a' , de niveau de blanc X'_w , Y'_w , Z'_w . Ces paramètres sont ajustés selon des conditions d'observations de référence.
- [0089] Lors d'une étape de conversion inverse E14, les valeurs trichromatiques X'_{ij} , Y'_{ij} , Z'_{ij} fournies par le modèle inverse CIECAM2⁻¹ sont converties dans l'espace colorimétrique RGB, de manière à fournir l'image corrigée I1 au format RGB, i.e. dans le même domaine colorimétrique que celui de la caméra 0.
- [0090] De manière optionnelle, l'étape de correction colorimétrique E1 comprend en outre une estimation fréquentielle E13, lors de laquelle une fréquence spatiale F_{ij} de chaque pixel P_{ij} de l'image monoculaire I0, ici l'image monoculaire convertie I1.1 au format RGB, est estimée. La fréquence estimée F_{ij} à l'issue de l'étape E13 est fournie en entrée du modèle d'apparence couleur inverse CIECAM2⁻¹ référencé 1b pour obtenir les valeurs trichromatiques corrigées X'_{ij} , Y'_{ij} , Z'_{ij} du pixel P_{ij} .

- [0091] Un exemple d'estimation fréquentielle E13 va être à présent décrit plus en détails toujours en référence à la [Fig.3].
- [0092] Lors d'une décomposition spectrale E13.1, l'image monoculaire I1.1 au format colorimétrique XYZ est transformée par application d'une transformée de Fourier en deux dimensions (*2D-Fourier Transform* en anglais) de manière à obtenir un spectre I1.2 de l'image monoculaire I1.1. Le spectre I1.2 comprend les composantes fréquentielles spatiales de l'image I1.1 dans l'espace de Fourier.
- [0093] Lors d'une étape de répartition spatio-fréquentielle E13.2, ces composantes fréquentielles sont réparties dans différents domaines spectraux et selon différentes orientations spatiales, de manière à former une pluralité de secteurs, fournissant ainsi un spectre ordonné en fréquences et dans l'espace I1.3.
- [0094] Dans le présent exemple, le spectre ordonné I1.3 comprend trois domaines spectraux : un domaine à fréquence nulle D_0 contenant les composantes continues (i.e. de fréquence nulle), un domaine à basses fréquences D_{BF} contenant des composantes fréquentielles inférieures à un seuil prédéterminé, et un domaine à hautes fréquences D_{HF} contenant des composantes fréquentielles supérieures audit seuil. Par exemple, le seuil est fixé comme étant égal à la fréquence moyenne de l'ensemble de fréquences (i.e. fréquence médiane d'un histogramme des fréquences du spectre I1.3).
- [0095] Comme illustré, le domaine à fréquence nulle D_0 est délimité par un cercle central C_0 . Le domaine à basses fréquences D_{BF} est compris entre le cercle central C_0 et un cercle intermédiaire C_i concentrique avec le cercle central C_0 . Le domaine à hautes fréquences D_{HF} est compris entre les bordures de l'image I1.3 et le cercle intermédiaire C_i . Pour chaque domaine de fréquences D_{BF} , D_{HF} , une fréquence moyenne est calculée et notée F_{BF} , F_{HF} respectivement pour les domaines à basses fréquences et à hautes fréquences. Pour chaque domaine, la fréquence moyenne F_{BF} , F_{HF} est calculée en fonction de la taille de l'image monoculaire originale I_0 .
- [0096] Dans des variantes de réalisation non décrites, plusieurs seuils fréquents pourront être définis de manière à répartir les composantes fréquentielles à travers un plus grand nombre de domaines fréquents. Par exemple, en définissant deux seuils fréquents, trois domaines fréquents (en dehors du domaine de fréquences nulles) pourront être définis (e.g. basses fréquences, moyennes fréquences, hautes fréquences) avec une fréquence moyenne associée à chaque domaine.
- [0097] Lors de l'étape E13.2, les fréquences spatiales sont réparties dans l'espace géométrique de l'image en deux dimensions (X,Y).
- [0098] Pour cela, l'image est découpée en plusieurs secteurs centrés suivant différentes orientations principales, de manière à former une pluralité de secteurs $\{s_k\}$. Ainsi, chaque secteur s_k (k étant un entier naturel, variant par exemple de 0 à 16) regroupe des fréquences spatiales dans l'espace de Fourier qui correspondent à une zone de l'image

originale dans l'espace géométrique, le passage entre un secteur s_k et la zone d'image associée s'effectuant par une transformée de Fourier inverse. Dans le présent exemple, le nombre de secteurs de fréquences non nulle est égal à 16 (avec k variant de 1 à 16). Toutefois, dans d'autres modes de réalisation, ce nombre pourra être ajusté, par exemple à des valeurs supérieures à 16 pour un échantillonnage plus fin permettant d'améliorer encore les résultats d'estimation de la profondeur monoculaire.

- [0099] La décomposition spatiale de l'image en plusieurs secteurs est relativement simple à mettre en œuvre et ne nécessite pas d'autres paramètres que la fréquence spatiale, contrairement à d'autres modèles de l'art antérieur (e.g. modèle de Daly), qui nécessitent une pluralité de paramètres, tels que la fréquence radiale, l'orientation, la luminance, la surface, la distance d'observation (i.e. distance entre la caméra et les objets de l'image), l'excentricité. Cette simplicité est particulièrement avantageuse pour permettre une détermination de la profondeur de champ en temps réel, notamment dans le cadre de la navigation d'un véhicule autonome.
- [0100] Dans le présent exemple, on considère les orientations principales suivantes : 0° , $+45^\circ$, -45° , $+90^\circ$, -90° , 135° , -135° , 180° . Ainsi, les fréquences nulles se répartissent dans le secteur s_0 . Les fréquences basses se répartissent dans les secteurs numérotés s_1 , s_2 , s_3 , s_4 , s_5 , s_6 , s_7 , s_8 . Les fréquences hautes se répartissent dans les secteurs numérotés s_9 , s_{10} , s_{11} , s_{12} , s_{13} , s_{14} , s_{15} , s_{16} .
- [0101] Dès lors que le spectre ordonné I1.3 est obtenu à l'issue de l'étape E13.2, la fréquence de chaque pixel P_{ij} est déterminée lors de l'étape E13.3. Pour cela, il s'agit de déterminer, pour chaque secteur s_k , la valeur $V_{x(i,j)}^k$ que possède le pixel P_{ij} considéré (sachant qu'un pixel peut contribuer fréquemment à plusieurs secteurs simultanément).
- [0102] La fréquence de la composante colorimétrique X du pixel P_{ij} , notée $F_x(i,j)$, est obtenue en moyennant les contributions du pixel P_{ij} sur la zone d'image associée à chacun des secteurs s_k . Chaque pixel P_{ij} de l'image originale se retrouve dans une des domaines fréquentiels prédéfinis et il s'agit de déterminer précisément la fréquence de ce pixel.
- [0103] Par exemple, la fréquence $F_x(i,j)$ est calculée en moyennant les influences de chaque zone associée au secteur s_k sur le pixel P_{ij} , selon l'équation 1 suivante : [Math 1]
- $$F_{x(i,j)} = \frac{\sum_{k=1}^{k=8} V_{x(i,j)}^k \times F_{BF} + \sum_{k=9}^{k=16} V_{x(i,j)}^k \times F_{HF}}{\sum_{k=1}^{k=16} V_{x(i,j)}^k} \text{ où } V_{x(i,j)}^k \text{ désigne la valeur colorimétrique } X$$
- du pixel P_{ij} sur la zone d'image associée au secteur s_k .
- [0104] La fréquence $F_x(i,j)$ ainsi obtenue correspond à la somme de la moyenne des domaines basses fréquences et hautes fréquences respectivement pondérées par la sommes des valeurs colorimétriques des pixels dans les secteurs associés. Cette

fréquence est normalisée dans le sens où elle comprend au dénominateur la somme des $V_{x(i,j)}^k$ sur l'ensemble des secteurs s_k .

[0105] Par exemple, si le pixel P_{ij} est fortement représenté dans le secteur s_1 , il présentera une forte corrélation avec les composantes de basses fréquences horizontales, i.e. selon une direction sensiblement horizontale selon l'axe X.

[0106] Les mêmes opérations sont réalisées pour les composantes colorimétrique Y et Z, de manière à obtenir les composantes fréquentielles de chaque pixel P_{ij} selon les équations suivantes : [Math 2]

$$F_{y(i,j)} = \frac{\sum_{k=1}^{k=8} V_{y(i,j)}^k \times F_{BF} + \sum_{k=9}^{k=16} V_{y(i,j)}^k \times F_{HF}}{\sum_{k=1}^{k=16} V_{y(i,j)}^k} ; \text{ [Math 3]}$$

$$F_{z(i,j)} = \frac{\sum_{k=1}^{k=8} V_{z(i,j)}^k \times F_{BF} + \sum_{k=9}^{k=16} V_{z(i,j)}^k \times F_{HF}}{\sum_{k=1}^{k=16} V_{z(i,j)}^k} \text{ où } V_{y(i,j)}^k, V_{z(i,j)}^k \text{ désignent respectivement les}$$

valeurs colorimétriques Y, Z du pixel P_{ij} sur la zone d'image associée au secteur s_k .

[0107] Finalement, la fréquence $F(i,j)$ du pixel P_{ij} est déterminée en calculant la moyenne des fréquences précédemment calculées pour chacune des trois composantes colorimétriques X, Y, Z, selon l'équation suivante : [Math 4] $F_{(i,j)} = \frac{F_{x(i,j)} + F_{y(i,j)} + F_{z(i,j)}}{3}$

[0108] Ainsi, la répartition des fréquences spatiales en différents secteurs spatiaux constitue une décomposition grossière de l'espace de Fourier permettant d'extraire par transformée de Fourier inverse les zones de l'image originale. Une telle décomposition permet de calculer de manière fiable la fréquence de chaque pixel de l'image originale tout en optimisant le temps de calcul et les ressources de mémoire. En effet, le regroupement des composantes spectrales en différents domaines fréquentiels est plus avantageux qu'une approche conventionnelle consistant à considérer l'influence de chacun des points du spectre I1.3 sur le pixel considéré P_{ij} . Cette astuce de calcul est d'autant plus avantageuse qu'il est nécessaire de réitérer les calculs pour chacune des trois composantes colorimétriques X, Y, Z du pixel P_{ij} .

[0109] La fréquence F_{ij} calculée pour chaque pixel P_{ij} est appliquée en entrée du modèle inverse CIECAM₂⁻¹ lors de l'étape inverse E12 pour déterminer les attributs perceptuels du pixel P_{ij} , en tenant compte de sa fréquence F_{ij} . Ainsi, à l'issue de l'étape de conversion E14, les coordonnées colorimétriques corrigées X'_{ij} , Y'_{ij} , Z'_{ij} du pixel P_{ij} converties au format RGB tiennent compte de la fréquence de chaque pixel P_{ij} .

[0110] La prise en compte de la fréquence F_{ij} pour chaque pixel P_{ij} permet avantageusement d'ajuster la correction de couleurs d'une façon adaptative qui préserve au mieux les détails de couleurs de l'image.

[0111] Les étapes E10, E11, E12, E13, E14 sont effectuées pour chacun des pixels P_{ij} de l'image d'origine I0 ou I1.1.

[0112] En référence à la [Fig.4], un exemple de réalisation de l'étape de filtrage E2 mise en

œuvre dans le procédé d'estimation selon la [Fig.1] va être à présent décrit en détails.

[0113] L'étape de filtrage E2 est configurée pour filtrer l'image monoculaire I0 de manière à obtenir une image filtrée I2, comprenant au moins des données de textures et éventuellement des données de bordures de l'image monoculaire I0.

[0114] Lors d'une étape de codage quaternionique E20, chaque pixel P_{ij} de l'image originale I0 en couleurs au format RGB est converti en un quaternion pur codant les trois composantes colorimétriques R, G, B selon l'équation suivante : [Math 5]

$P_{ij} = iP_R(i, j) + jP_G(i, j) + kP_B(i, j)$, où $P_R(i, j)$, $P_G(i, j)$ et $P_B(i, j)$ sont les composantes rouge, verte et bleue respectivement.

[0115] Par définition, un quaternion q est un nombre hypercomplexe comprenant une partie réelle et trois parties imaginaires, comme exprimé selon l'expression suivante : [Math 6] $q = q_r + iq_i + jq_j + kq_k$, où q_r est la partie réelle, q_i , q_j et q_k sont les trois parties imaginaires et i , j , k sont trois opérateurs complexes suivant les règles suivantes :

$$[\text{Math 7}] \quad i^2 = j^2 = k^2 = ijk = -1 \text{ et } \begin{cases} ij = -ji = k \\ jk = -kj = i \\ ki = -ik = j \end{cases}$$

[0116] Si le quaternion q a une norme unitaire, i.e. $\|q\| = 1$, on dit que q est un quaternion unitaire. Si la partie réelle de q est nulle, i.e. $q_r = 0$, on dit que q est un quaternion pur. Le module et le conjugué du quaternion q sont définis respectivement selon les expressions suivantes : [Math 8] $\|q\| = \sqrt{q\bar{q}} = \sqrt{q_r^2 + q_i^2 + q_j^2 + q_k^2}$; [Math 9] $\bar{q} = q_r - iq_i - jq_j - kq_k$.

[0117] L'utilisation d'une représentation quaternionique de l'image monoculaire pour effectuer le filtrage est avantageuse car le traitement quaternionique est particulièrement efficace pour optimiser les ressources en mémoire et en capacité de calcul, et avec une meilleure précision sur les informations de couleurs que les approches conventionnelles, car chaque pixel P_{ij} est présenté par un vecteur dans un espace à trois dimensions (RGB) qui est traité comme une seule entité.

[0118] De préférence, le filtrage est un filtrage quaternionique guidé.

[0119] Ce filtre est configuré lors d'une étape de configuration E02 préalable selon un modèle prédéfini. De préférence, le filtre quaternionique guidé est un modèle linéaire et local entre une image guide I et la réponse du filtre q qui est une transformée linéaire dans une fenêtre de pixels W_k centrée autour d'un pixel k selon l'équation suivante : [Math 10] $q_i = a_k I_i + b_k$; $\forall i \in W_k$; où (a_k, b_k) sont des coefficients linéaires et des quaternions purs dans la fenêtre W_k .

[0120] Dans le présent exemple, pour chaque pixel p_i à filtrer, lors d'une étape E21, le filtre quaternionique est appliqué localement à une pluralité de fenêtres de pixels

comprenant le pixel p_i . L'image guide I dite image de référence pour guider le filtrage est l'image I1 obtenue après application de la correction couleur.

[0121] Les coefficients linéaires (a_k, b_k) sont déterminés en minimisant la différence entre la réponse du filtre $q_i = a_k p_i + b_k$ et son entrée p_i .

[0122] Lors d'une étape E22, une fonction de coût quaternionique E_Q exprimant cette différence est calculée dans W_k , par exemple, selon l'équation suivante : [Math 11]

$$E_Q(a_k, b_k) = \sum_{i \in W_k} \left(\| \text{Im}(a_k I_i + b_k) - p_i \|^2 + \lambda \| \text{Re}(a_k I_i + b_k) \|^2 + \varepsilon \| a_k \|^2 \right)$$

, où les coefficients a_k, b_k du filtre quaternionique guidé, les pixels I_i de l'image de guide et le pixel p_i de l'image originale à filtrer sont exprimés respectivement sous une forme quaternionique comme suit : [Math 12] $a_k = ia_k^1 + ja_k^2 + ka_k^3$; [Math 13] $b_k = ib_k^1 + jb_k^2 + kb_k^3$; [Math 14] $I_i = iI_i^1 + jI_i^2 + kI_i^3$; [Math 15] $p_i = ip_i^1 + jp_i^2 + kp_i^3$; où $\text{Im}()$, $\text{Re}()$ désignent respectivement la partie imaginaire et la partie réelle ; et λ et ε sont des paramètres permettant d'ajuster le résultat du filtrage de telle sorte à ce qu'il soit constant dans la fenêtre W_k ; ε étant en outre un paramètre de régularisation pour éviter que a_k soit trop grand.

[0123] Lors d'une étape de minimisation E23, la fonction de coût quaternionique E_Q est minimisée par rapport aux composantes quaternioniques des coefficients (a_k, b_k) , i.e. $a_k^1, a_k^2, a_k^3, b_k^1, b_k^2, b_k^3$ en résolvant l'ensemble des équations suivantes : [Math 16]

$$\frac{\partial E_Q}{\partial a_k^1} = 0, \frac{\partial E_Q}{\partial a_k^2} = 0, \frac{\partial E_Q}{\partial a_k^3} = 0, \frac{\partial E_Q}{\partial b_k^1} = 0, \frac{\partial E_Q}{\partial b_k^2} = 0, \frac{\partial E_Q}{\partial b_k^3} = 0$$

[0124] Ainsi, les coefficients a_k, b_k minimisant la fonction de coût quaternionique E_Q sont déterminés par les coefficients quaternioniques $a_k^1, a_k^2, a_k^3, b_k^1, b_k^2, b_k^3$.

[0125] Lors d'une étape E24, le pixel filtré q_i en sortie du filtre quaternionique guidé est déterminé à partir des coefficients a_k, b_k préalablement obtenus lors de l'étape E23, tel que $q_i = a_k \cdot I_i + b_k$. En tant qu'élément d'image quaternionique, ce pixel filtré q_i s'exprime sous la forme quaternionique suivante [Math 17] : $q_i = iq_i^1 + jq_i^2 + kq_i^3$.

[0126] Lors d'une étape E25, les coefficients quaternioniques a_k, b_k et le pixel filtré q_i sont stockés en association avec la fenêtre W_k .

[0127] Les étapes E21, E22, E23, E24, E25 sont répétées pour chaque fenêtre W_k .

[0128] Cependant, pour les différentes fenêtres W_k couvrant un même pixel P_{ij} , noté p_i pour simplifier, les valeurs de q_i calculées précédemment ne sont pas identiques. Par conséquent, le procédé comprend en outre une étape E26 configurée pour calculer la moyenne de ces différentes valeurs obtenues de q_i obtenues pour différentes fenêtres*

[0129] Par exemple, l'étape E26 fournit pour chaque pixel p_i de l'image originale I0, un pixel filtré et moyenné selon l'expression suivante : [Math 18]

$$q_i = \frac{1}{|W|} \sum_{k \in W_k} (a_k I_i + b_k) = \frac{1}{|W|} \sum_{k \in W_k} \left((ia_k^1 + ja_k^2 + ka_k^3)(I_i^1 i + I_i^2 j + I_i^3 k) + (ib_k^1 + jb_k^2 + kb_k^3) \right) = -(\hat{a}_i I_i^1 + \hat{a}_i^2 I_i^2 + \hat{a}_i^3 I_i^3) + i(\hat{a}_i^1 I_i^1 - \hat{a}_i^2 I_i^2 + \hat{b}_i^1) + j(\hat{a}_i^1 I_i^1 - \hat{a}_i^2 I_i^2 + \hat{b}_i^2) + k(\hat{a}_i^1 I_i^1 - \hat{a}_i^2 I_i^2 + \hat{b}_i^3)$$

- [0130] Lors du calcul de la fonction de coût quaternionique E_Q lors de l'étape E22, le paramètre λ peut être ajusté de sorte que la partie réelle de $a_k I_i + b_k$ se rapproche de 0 sans être nulle. Par exemple, lorsque le coefficient λ est relativement grand, de préférence compris entre 0,1 et 0,5, les détails de l'image dans la partie réelle de $a_k I_i + b_k$ sont quasiment nuls. Par conséquent, l'énergie de l'image en couleurs est en grande partie portée par la partie imaginaire des pixels q_i après filtrage. Dans ce cas, le pixel filtré peut être considéré comme un quaternion pur de forme suivante :
- $$q_i = iq_i^1 + jq_i^2 + kq_i^3, \text{ où : [Math 19] } q_i^1 = \bar{a}_i^2 I_i^3 - \bar{a}_i^3 I_i^2 + b_i^1; q_i^2 = \bar{a}_i^3 I_i^1 - \bar{a}_i^1 I_i^3 + b_i^2;$$
- $$q_i^3 = \bar{a}_i^1 I_i^2 - \bar{a}_i^2 I_i^1 + b_i^3 \text{ avec } \bar{a}_i^1 = \frac{1}{|W|} \sum_{k \in W_k} a_k^1; \bar{a}_i^2 = \frac{1}{|W|} \sum_{k \in W_k} a_k^2; \bar{a}_i^3 = \frac{1}{|W|} \sum_{k \in W_k} a_k^3;$$
- et $\bar{b}_i^1 = \frac{1}{|W|} \sum_{k \in W_k} b_k^1; \bar{b}_i^2 = \frac{1}{|W|} \sum_{k \in W_k} b_k^2; \bar{b}_i^3 = \frac{1}{|W|} \sum_{k \in W_k} b_k^3$
- [0131] Ainsi, le filtre quaternionique permet de fournir une image passe-bas dans laquelle les détails sont préservés.
- [0132] En référence à la [Fig.5], un exemple de réalisation détaillé d'un dispositif pour mettre en œuvre le procédé selon l'invention va être à présent décrit en détails.
- [0133] Ce dispositif est un exemple de mise en œuvre du dispositif 100 préalablement décrit de manière générique en référence à la [Fig.2]. L'architecture générale et le principe de fonctionnement restent identiques. En référence à la [Fig.5], on se concentre principalement sur la mise en œuvre du réseau de neurones 4, plus particulièrement du premier sous-réseau 4.1 et du deuxième sous-réseau 4.2 dans le même contexte que précédemment.
- [0134] Le premier sous-réseau 4.1 comprend une première couche de convolution 4.10 prenant en entrée l'image corrigée en couleurs I1 issue du module de correction chromatique 1. Par exemple, la première couche de convolution 4.10 a un noyau de convolution de taille 7x7 pixels permettant de fournir un bloc de données I410 de taille réduite par un facteur 2 par rapport à celle de l'image I1.
- [0135] En sortie de la première couche de convolution 4.10, le premier sous-réseau 4.1 comprend une couche de « pooling » 4.11 configurée pour réduire la taille des données de l'image d'un facteur par exemple égale à 2. Ainsi, la couche de « pooling » 4.11 produit un bloc de données I411 d'une taille réduite d'un facteur égale à 4 par rapport à celle de l'image I1.
- [0136] En sortie de la couche de pooling 4.11, le premier sous-réseau 4.1 comprend une deuxième couche de convolution 4.12. Par exemple, cette couche est configurée avec un noyau de convolution d'une taille égale à 3x3 pixels, pour réduire le volume de données d'un facteur égal à 2. Ainsi, la deuxième couche de convolution 4.12 produit un bloc de données I412 d'une taille réduite d'un facteur 8 par rapport à celle de l'image I1.

- [0137] En sortie de la deuxième couche de convolution 4.12, le premier sous-réseau 4.1 comprend deux couches de convolution dilatée 4.13, 4.14, fournies en série à l'une à la suite de l'autre. Ces convolutions dilatées permettent de fournir deux dernières couches de sous-échantillonnages, en s'affranchissant des opérateurs de « stride » utilisés pour traduire le noyau de convolution des couches de convolution conventionnelles.
- [0138] Par définition, une convolution dilatée est une technique de convolution selon laquelle le noyau de convolution (« kernel » en anglais) est élargi en insérant des trous entre ses éléments consécutifs. Autrement dit, il s'agit d'une convolution impliquant un saut de pixels, de manière à couvrir une plus grande surface de l'image d'entrée.
- [0139] Les deux couches de convolution dilatée 4.13, 4.14 ont pour effet de maintenir constante la taille des blocs de données en sortie de ces couches I413, I414 respectivement, ce qui permet avantageusement de limiter la réduction de la résolution de la carte de profondeur réduite I4.1 fournie en sortie du premier sous-réseau 4.1. Dans le présent exemple, la résolution de l'image en sortie des deux couches 4.13, 4.14 est maintenue à 1/8 de celle de l'image I1 fournie en entrée du réseau neuronal 4. En sortie des deux couches de convolution dilatée 4.13, 4.14, le premier sous-réseau 4.1 comprend en parallèle une pluralité, de préférence quatre, couches de convolution dilatée 4.15a, 4.15b, 4.15c, 4.15d. De manière avantageuse, les couches de convolution dilatée ont des facteurs de dilatation fixés à des valeurs distinctes. Par exemple, les quatre couches de convolution dilatée 4.15a, 4.15b, 4.15c, 4.15d ont des facteurs de dilatation égaux à 6, 12, 18 et 24 respectivement.
- [0140] Les couches de convolution dilatées 4.15a, 4.15b, 4.15c, 4.15d forment une structure pyramidale permettant d'extraire quatre cartes de caractéristiques de profondeur sous-échantillonnées I415a, I415b, I415c, I415d respectivement, avec différents champs réceptifs (« receptive field » en anglais). Par champ « réceptif », on comprend un champ de capteur destiné à capturer des caractéristiques d'image plus ou moins globales (ou locales). Par exemple, on peut considérer des champs réceptifs correspondant respectivement à différents niveaux de détails d'images, i.e. plus ou moins fins ou grossiers.
- [0141] L'ensemble de ces couches 4.15a, 4.15b, 4.15c, 4.15d constitue un schéma « multi-échelle » qui permet avantageusement de garantir que les régions d'image à une échelle arbitraire peuvent être récupérées avec précision. En effet, la prédiction d'une carte de profondeurs à partir d'une caméra fixe ne peut pas exploiter des objets captés à différentes échelles, contrairement à une caméra mobile. Pour surmonter cet inconvénient, une solution consiste à utiliser les mêmes architectures de réseaux de neurones convolutifs sur différentes versions redimensionnées des mêmes ensembles de données, puis à concaténer les cartes de caractéristiques de différents réseaux de neurones convolutifs pour obtenir un résultat final.

- [0142] Cet ensemble de couches permet en outre de s'affranchir de couches entièrement connectées (*fully-connected* en anglais) qui constituent les couches terminales d'un réseau de neurones convolutif classique. Ainsi, cet ensemble de couches 4.15a, 4.15b, 4.15c, 4.15d permet avantageusement de simplifier la structure du réseau neuronal et par conséquent de faciliter sa mise en œuvre, notamment dans le cadre d'applications embarquées, telles que la navigation de véhicules autonomes avec la détection d'obstacles en temps réel.
- [0143] En sortie des couches de convolution dilatée, le premier sous-réseau 4.1 comprend un module de fusion 4.16 configuré pour fusionner les quatre cartes de profondeur I415a, I415b, I415c, I415d, de manière à fournir la carte de profondeur réduite I4.1.
- [0144] En raison de l'agrégation des caractéristiques de profondeur sous-échantillonnées obtenue par le module de fusion 4.16, la résolution de la carte de profondeur réduite I4.1 est égale à seulement 1/8 de la résolution de l'image I1 fournie en entrée du premier sous-réseau 4.1.
- [0145] Pour augmenter la résolution de la carte de profondeur réduite I4.1, le premier sous-réseau 4.1 est relié en sortie du module de fusion 4.16 au deuxième sous-réseau 4.2. De préférence, le deuxième sous-réseau 4.2 est un réseau hiérarchique plutôt qu'une interpolation bicubique pour reconstruire progressivement une carte de profondeur à la même taille que celle de l'image couleurs d'entrée I1.
- [0146] Le terme « hiérarchique » fait référence à la structure hiérarchique ou pyramidale du sous-réseau 4.2, i.e. présentant plusieurs niveaux de résolution augmentant progressivement avec les couches neuronales constitutives du sous-réseau de manière à atteindre une résolution au moins égale à celle de l'image corrigée en couleurs I1.
- [0147] Par exemple, le deuxième sous-réseau 4.2 comprend successivement trois couches de sur-échantillonnage 4.20, 4.21, 4.22 pour sur-échantillonner progressivement en trois niveaux la carte de profondeur réduite I4.1 issue du premier sous-réseau 4.1. De préférence, chacune de ces trois couches de sur-échantillonnage comprend une couche de convolution suivie d'une couche de convolution transposée. Chaque couche de sur-échantillonnage 4.20, 4.21, 4.22 est configurée de telle sorte qu'elle agrandisse la taille de la carte de profondeur ou de manière équivalente, la résolution d'un facteur, de préférence égal à 2.
- [0148] Ainsi, la carte de profondeur réduite I4.1 fournie en entrée du deuxième sous-réseau 4.2 est constituée par un bloc de données I420 d'une résolution égale à 1/8 de celle de l'image I1 fournie en entrée du premier sous-réseau 4.1. Le bloc de données I421 fourni en sortie de la couche 4.20 a une taille (ou résolution) égale à 1/4 de celle de l'image d'entrée I1. Le bloc de données I422 fourni en sortie de la couche 4.21 a une taille (ou résolution) égale à 1/2 de celle de l'image d'entrée I1. Ainsi, le deuxième sous-réseau 4.2 permet de sur-échantillonner progressivement la carte de profondeur

intermédiaire I4.1 fournie par le premier sous-réseau 4.1 pour accroître la taille, et par conséquent la résolution de la carte de profondeur.

- [0149] Le deuxième sous-réseau 4.2 comprend en outre une couche de convolution 4.23 reliée à la sortie du module de soustraction 3, de manière à fournir une carte intermédiaire I'3. Par exemple, cette couche de convolution 4.23 présente un noyau de convolution de taille égale à 3x3.
- [0150] Le deuxième sous-réseau 4.2 comprend en outre un module de concaténation 4.24 relié au module de convolution 4.23 et à la couche de sur-échantillonnage 4.22. A partir des caractéristiques sur-échantillonnées I424 fournies par la couche 4.22, le module de concaténation 4.24 est configuré pour augmenter la résolution de la carte de profondeur, de sorte qu'elle ait une résolution égale à celle de la carte de l'image corrigée en couleurs I1 (ou de l'image monoculaire d'origine I0). Le module de concaténation 4.24 fournit en sortie un bloc de données I425.
- [0151] En sortie du module de concaténation 4.24, le deuxième sous-réseau 4.2 comprend en outre, plusieurs (de préférence deux) couches de convolution 4.25, 4.26, de manière à fournir la carte de profondeur résiduelle I4.2. Par exemple, chacune de ces couches 4.25, 4.26 a un noyau de convolution de taille égale à 3x3 pixels.
- [0152] En sortie du premier sous-réseau 4.1 et du deuxième sous-réseau 4.2 est relié le module d'addition 5, de manière à fournir la carte de profondeur finale I5 en ajoutant la carte de profondeur résiduelle I4.2 à la carte de profondeur réduite I4.1.
- [0153] Ainsi, le deuxième sous-réseau 4.2 permet d'améliorer l'estimation de la profondeur réalisée par le premier sous-réseau 4.1. En particulier, le deuxième sous-réseau 4.2 fournit un apprentissage résiduel permettant un entraînement rapide et stable qui contribue à obtenir des résultats d'estimation de la profondeur améliorés.
- [0154] Entraînement
- [0155] Le réseau neuronal 4 dans son ensemble est entraîné en utilisant un outil d'apprentissage automatique, tel que *TensorFlow* développé par *Google*.
- [0156] Le premier sous-réseau 4.1 est initialisé en utilisant un réseau de neurones profond résiduel (ResNet), par exemple le réseau ResNet-101. Ce réseau est ensuite entraîné en utilisant une première fonction de perte E1 de norme L2 appliquée entre des images de profondeur sous-échantillonnées et les sorties de profondeurs sous-échantillonnées par un facteur égal à 8 correspondantes. Par exemple, la base de données d'entraînement KITTI est utilisée pour fournir des images et les cartes de profondeurs correspondantes qui ont été sous-échantillonnées.
- [0157] Le deuxième sous-réseau 4.2 est initialisé au hasard et entraîné en utilisant une deuxième fonction de perte E2 de norme L2 appliquée entre les paires de profondeur de basse résolution (1/8) correspondant à la carte de profondeur réduite I4.1 et de profondeur de haute résolution correspondant à la carte de profondeur résiduelle I4.2.

Comme pour le premier sous-réseau 4.1, la base de données d'entraînement KITTI est utilisée pour obtenir ces paires de profondeur à haute résolution (HR) et à basse résolution (LR).

- [0158] Les premier 4.1 et deuxième 4.2 sous-réseaux sont ajustés conjointement en utilisant une seule fonction de perte, commune aux deux sous-réseaux 4.1, 4.2. De préférence, cette fonction de perte, désignée L , est définie comme suit [Math 20] :

$$L(\hat{d}_s, \hat{d}, d_s, d) = \sum_{i=1}^N \left(\left\| \hat{d}_s^{(i)} - d_s^{(i)} \right\|_2^2 + \lambda \left\| \hat{d}^{(i)} - d^{(i)} \right\|_2^2 \right) \text{ où } d_s \text{ désigne la}$$

carte de profondeur de vérité terrain sous-échantillonnée d'un facteur (par exemple à 8) ; $\hat{d}_s$ désigne la carte de profondeur en sortie du réseau ; d désigne la carte de profondeur de la vérité terrain ; $\hat{d} = g(c, \theta)$ désigne la carte de profondeur obtenue par une fonction de mappage (de l'anglais « mapping function ») g , telle que $\hat{d} = g(c, \theta)$ ressemble à la carte de profondeur de la vérité terrain d en fonction de l'image d'entrée c (au format RGB) et de l'ensemble des paramètres θ du réseau à optimiser ; λ est un paramètre de balance entre les deux termes de perte

$$\left\| \hat{d}^{(i)} - d^{(i)} \right\|_2^2 \text{ et } \left\| \hat{d}_s^{(i)} - d_s^{(i)} \right\|_2^2 ; N \text{ est le nombre d'échantillons}$$

d'apprentissage.

- [0159] Le but de l'entraînement est d'apprendre la fonction de mappage g pour générer la carte de profondeur $\hat{d} = g(c, \theta)$ avec un degré de ressemblance optimum à la carte de profondeur d de la vérité terrain.
- [0160] Un des avantages de la stratégie dite « à double perte » telle que proposée selon l'équation [Math 20] ci-dessus est qu'elle permet une supervision de couches intermédiaires, c'est-à-dire une supervision simultanément des deux sous-réseaux 4.1, 4.2. En plus, tous les filtres convolutifs utilisés pour l'extraction de caractéristiques, l'échantillonnage et le schéma multi-échelle sont entraînés conjointement d'une façon de bout-en-bout (de l'anglais « end-to-end »).
- [0161] L'architecture complète est entraînée en utilisant par exemple la plateforme (ou framework) *TensorFlow*. Chaque sous-réseau 4.1, 4.2 est tout d'abord entraîné séparément.
- [0162] Concernant le premier sous-réseau 4.1 de prédiction de la profondeur, le réseau « ResNet 101 » est utilisé pour initialiser ce sous-réseau réseau 4.1. Le modèle de ce sous-réseau est ensuite affiné avec une perte de norme L2 calculée entre les images de profondeur sous-échantillonnées et les sorties de profondeur 1/8 correspondantes.
- [0163] Concernant le deuxième sous-réseau 4.2 d'amélioration de la profondeur, le modèle de ce sous-réseau 4.2 est initialisé aléatoirement. Ce sous-réseau 4.2 est ensuite

entraîné avec une fonction de perte de norme L2 calculée entre les paires de profondeur dite LR-HR, i.e. à basse-résolution (LR) - haute résolution (HR).

[0164] Ensuite, les deux sous-réseaux 4.1, 4.2 sont ajustés conjointement à l'aide de la fonction de perte L proposée avec une double supervision.

[0165] De préférence, le paramètre d'inertie (*momentum* en anglais) est fixé à 0,9 et le taux d'apprentissage est initialisé à 10^{-4} pour toutes les couches des deux sous-réseaux 4.1, 4.2 puis diminué de 0,90 à chaque époque ou itération (*epoch* en anglais).

[0166] Résultats et performances

[0167] Le réseau neuronal 4, plus particulièrement les deux sous-réseaux 4.1, 4.2 sont entraînés en utilisant une base de données d'entraînement connue, telle que la base KITTI.

[0168] Les résultats du réseau neuronal 4 selon l'invention sont comparés avec ceux de trois méthodes de l'état de l'art désignées ci-après A, B et C et respectivement décrites dans les articles suivants :

- [A] : H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/ CVF Conf. Computer Vision Pattern Recognition*, Jun. 2018, pp. 2002–2011.
- [B] : J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation" 2019, arXiv:1907.10326.
- [C] : Z. Hao, Y. Li, S. You, and F. Lu, "Detail preserving depth estimation from a single image using attention guided networks," in *Proc. Int. Conf. 3D Vis. (DV)*, Sep. 2018, pp. 304–313.

[0169] Six métriques d'évaluation telles que décrites ci-dessous sont utilisées pour comparer les résultats de la présente invention avec ceux obtenus par les trois méthodes de l'état de l'art A, B, C :

- L'erreur relative absolue (« Absolute Relative Error » en anglais) :

$$Abs\ Rel = \frac{1}{|T|} \sum_{y \in T} |y - y^*| / y^*,$$
où T désigne un ensemble de pixels de référence pour lesquels des valeurs de vérité terrain de profondeur sont connues, |T| désigne le nombre des pixels valides de la « vérité terrain » et y^* désigne la « vérité terrain » de profondeur.
- L'erreur relative absolue au carré (« Squared Absolute Relative Error » en anglais) : $Sq\ Rel = \frac{1}{|T|} \sum_{y \in T} \|y - y^*\|^2 / y^*.$
- La racine de l'erreur quadratique moyenne (« Root Mean Square Error » en anglais) : $RMSE = \sqrt{\frac{1}{|T|} \sum_{y \in T} \|y - y^*\|^2}.$
-

$$RMSE_{log} = \sqrt{\frac{1}{|T|} \sum_{y \in T} \|\log y - \log y^*\|^2}.$$

- L'écart logarithmique moyen : $\log 10 = \frac{1}{|T|} \sum_{y \in T} |\log y - \log y^*|$.
- La précision (*accuracy* en anglais) δ est un rapport, pouvant être exprimé par un pourcentage de y sous la condition $\max\left(\frac{y}{y^*}, \frac{y^*}{y}\right) = \delta < Th$, où Th désigne un seuil prédéterminé. Typiquement, le seuil Th est fixé à 1,25, ce qui correspond à une précision de prédiction de $\{\pm 0,25, \pm 0,5625, \pm 0,9531\}$ comparé à la vérité terrain.

[0170] Les performances de l'invention sont comparées à celles des trois méthodes A, B, C pour les paramètres δ , Abs Rel, Log10 et RMSE comme indiqué dans le tableau 1 ci-dessous. En particulier, le paramètre de précision δ est estimé pour trois seuils distincts : $Th=1,25$; $Th=1,25^2=1,5623$; $Th=1,25^3=1,95$.

Méthode	$\delta < 1,25$	$\delta < 1,25^2$	$\delta < 1,25^3$	Abs Rel	log10	RMSE
A	0,828	0,965	0,992	0,115	0,051	0,509
B	0,882	0,979	0,995	0,112	0,047	0,352
C	0,841	0,966	0,991	0,127	0,053	0,555
Invention	0,885	0,979	0,995	0,110	0,047	0,393

[Table 1] Tableau 1 : Performances

- [0171] Au vu de ces résultats, la présente invention fournit de meilleures performances que celles obtenues par les méthodes de l'état de l'art A, B et C. En particulier, les valeurs de précision δ sont maximales pour chacune des trois conditions testées, i.e. $\delta < Th$; $\delta < Th^2$; $\delta < Th^3$. Par exemple, l'invention permet d'obtenir des valeurs égales à celles obtenues avec les méthode B et C respectivement pour $\delta < Th^2$ et $\delta < Th^3$. Plus les valeurs de δ sont proches de 1, plus la méthode est précise. En ce qui concerne les trois autres critères, plus leurs valeurs sont faibles, plus la méthode est fiable dans le sens où les estimations présentent peu de dispersion par rapport aux données de « vérité terrain ». Ainsi, les valeurs 0,110 ; 0,047 ; 0,393 obtenues respectivement pour les critères Abs Rel, Log10 et tout particulièrement RMSE sont inférieures à celles obtenues avec les méthodes A, B et C, montrant ainsi une fiabilité et une précision accrue de la profondeur monoculaire de l'invention par rapport à l'art antérieur.
- [0172] Selon un mode de réalisation de l'invention, le dispositif 100 tel que décrit en référence à la [Fig.2], est au moins en partie mise en œuvre par un système informatique ou ordinateur 6 qui va à présent être décrit plus en détails en référence à la [Fig.6].

- [0173] Ce système informatique 600 comporte une unité de traitement 6.1 de données (e.g. un microprocesseur noté CPU, de l'anglais « Central Processing Unit ») et une mémoire principale 6.2 (e.g. RAM de l'anglais « Random Access Memory ») et une mémoire morte 6.3 (e.g. ROM de l'anglais « Read Only Memory »), accessibles par l'unité de traitement 6.1. Le système informatique 6 comporte en outre, par exemple une interface réseau 6.4, et/ou un support lisible par ordinateur, comme par exemple un support local, tel qu'un disque dur local 6.6 ou bien un support distant (tel qu'un disque dur distant et accessible via l'interface réseau 6.4 au travers d'un réseau de communication) ou bien encore un support amovible (tel qu'une clé USB, de l'anglais « Universal Serial Bus », ou bien un CD, de l'anglais « Compact Disc » ou bien un DVD, de l'anglais « Digital Versatile Disc ») lisible au moyen d'un lecteur approprié du système informatique 6 (tel qu'un port USB ou bien un lecteur de disque CD et/ou DVD).
- [0174] Un programme d'ordinateur Q contenant des instructions pour l'unité de traitement 6.1 est enregistré sur le support 6.6 et/ou téléchargeable via l'interface réseau 6.4. Ce programme d'ordinateur Q est, par exemple, destiné à être chargé dans la mémoire principale 6.2, afin que l'unité de traitement 6.1 exécute ses instructions. Ces instructions peuvent être organisées en modules logiciels, tels que décrits précédemment en référence aux figures 2 et 5, en particulier, le module de correction colorimétrique 1, le module de filtrage 2, le réseau de neurones 4 y compris le premier sous-réseau 4.1 et le deuxième sous-réseau 4.2, le module de soustraction 3, le module d'addition 5, pour mettre en œuvre les étapes E1, E2, E3, E4, E5 du procédé E100 tel que décrit en référence à la [Fig.1]. Ainsi, les étapes E1, E2, E3, E4, E5 du procédé E100 sont mises en œuvre par ordinateur.
- [0175] Cependant, cette présentation ne préjuge pas de la forme du programme d'ordinateur, qui peut être quelconque. Alternativement, tout ou partie de ces modules pourrait être implémenté sous forme de modules matériels, c'est-à-dire sous forme d'un circuit électronique, par exemple micro-câblé, ne faisant pas intervenir de programme d'ordinateur.
- [0176] En référence à la [Fig.7], un robot terrestre 700 dans lequel l'invention est mise en œuvre va être à présent décrit.
- [0177] Le robot terrestre 700 est adapté à se déplacer sur le sol, de manière autonome.
- [0178] Par exemple, il comprend le dispositif 100 selon la [Fig.2] pour déterminer la profondeur de champ des images captées par la caméra 0. Il est adapté à détecter des obstacles, estimer sa distance par rapport aux différents obstacles détectés et ajuster sa trajectoire en conséquence de manière à éviter toute collision avec un des obstacles détectés.
- [0179] En référence à la [Fig.8], un drone 800 dans lequel le dispositif selon l'invention est

mis en œuvre va être à présent décrit.

[0180] Le drone 800 est adapté à se déplacer dans l'air, de manière autonome.

[0181] Par exemple, il comprend le dispositif 100 selon la [Fig.2] pour déterminer la profondeur de champ des images captées par la caméra 0. Il est en outre adapté à détecter des obstacles fixes ou mobiles, estimer sa distance par rapport aux différents obstacles détectés et ajuster sa trajectoire en conséquence de manière à éviter toute collision avec un obstacle.

[0182] On notera par ailleurs que l'invention n'est pas limitée aux modes de réalisation décrits précédemment. Il apparaîtra en effet à l'homme de l'art que diverses modifications peuvent être apportées aux modes de réalisation décrits ci-dessus, à la lumière de l'enseignement qui vient de lui être divulgué.

Applications industrielles

[0183] Selon une première application industrielle, le procédé d'estimation de la profondeur monoculaire selon l'invention pourra être mis en œuvre pour la navigation d'un engin de mobilité en trois dimensions, par exemple, basée sur la réception de signaux satellitaires (e.g. GPS : « Global Positioning System » en anglais).

[0184] Le dispositif selon l'invention pourra être intégré dans un système de vision artificielle embarqué dans l'engin de mobilité et configuré pour détecter la présence d'obstacles à partir d'une image monoculaire en couleurs captée par une caméra monoculaire et estimer la distance entre le dispositif et l'obstacle détecté en fonction de la profondeur monoculaire estimée selon le procédé de l'invention.

[0185] Le dispositif et le procédé selon l'invention trouvent une application privilégiée pour les engins de mobilité (e.g. aéronefs, drones, robots) dits autonomes, i.e. adaptés à se déplacer de manière autonome.

[0186] Dès lors qu'un obstacle potentiel est identifié, l'engin de mobilité autonome doit être capable d'estimer avec une fiabilité accrue la profondeur monoculaire pour en déduire la distance séparant l'obstacle de l'engin de mobilité, afin d'ajuster sa trajectoire en temps réel et éviter toute collision.

[0187] Selon une deuxième application industrielle, le procédé d'estimation de la profondeur monoculaire selon l'invention pourra être mis en œuvre pour la reconstruction automatique d'une scène en trois dimensions à partir d'une ou plusieurs images en deux dimensions captées par une caméra monoculaire (i.e. un flux d'images monoculaires).

[0188] Dans la présentation détaillée de l'invention qui est faite précédemment, les termes utilisés ne doivent pas être interprétés comme limitant l'invention aux modes de réalisation exposés dans la présente description, mais doivent être interprétés pour y inclure tous les équivalents dont la prévision est à la portée de l'homme de l'art en appliquant ses connaissances générales à la mise en œuvre de l'enseignement qui vient de

lui être divulgué.

Revendications

[Revendication 1] Procédé (E100) d'estimation d'une carte de profondeur d'une image monoculaire (I0), ledit procédé étant caractérisé en ce qu'il comprenant les étapes suivantes :

- fournir (E0) une image monoculaire en couleurs (I0) acquise par un dispositif de captation optique monoculaire (0) dans des conditions d'observations réelles ;
- réaliser une correction colorimétrique (E1) de l'image monoculaire (I0) à partir d'un modèle d'apparence couleur inversible (1a), de manière à fournir une image corrigée en couleurs (I1) dans des conditions d'observation de référence, de sorte que la perception des couleurs dans l'image corrigée en couleurs (I1) soit identique à la perception des couleurs de l'image monoculaire en couleurs (I0) acquise dans les conditions d'observations réelles; et
- traiter (E4 ; E4.1) l'image corrigée en couleurs (I1) en appliquant un réseau de neurones convolutif (4 ; 4.1 ; 4.1, 4.2) de manière à générer une carte de profondeur (I4.1 ; I5) de l'image monoculaire en couleurs (I0).

[Revendication 2] Procédé selon la revendication 1, dans lequel l'image monoculaire en couleurs (I0) comprend une pluralité de pixels (P_{ij}) et la correction colorimétrique (E1) comprenant les étapes suivantes :

- appliquer (E11) le modèle d'apparence couleur inversible (1a) à chaque pixel (P_{ij}) de l'image monoculaire en couleurs (I0 ; I1.1), de manière à déterminer au moins un attribut perceptuel du pixel (J_{ij} , C_{ij} , h_{ij}) sélectionné parmi : clarté (J), chromaticité (C), angle de teinte (h), brillance (Q), saturation (s), composition de teinte (H) ;
- inverser (E12) le modèle d'apparence couleur inversible (1a) de manière à fournir un modèle d'apparence couleur inverse (1b) ; et
- pour chaque pixel (P_{ij}), appliquer (E12) ledit au moins un attribut perceptuel (J_{ij} , C_{ij} , h_{ij}) au modèle d'apparence couleur inverse (1b) configuré avec les conditions d'observation de référence, de manière à fournir des valeurs trichromatiques

corrigées (X'_{ij} , Y'_{ij} , Z'_{ij}) dudit pixel (P_{ij}).

- [Revendication 3] Procédé selon la revendication 2, dans lequel la correction colorimétrique (E1) comprend en outre une estimation fréquentielle (E13), lors de laquelle une fréquence spatiale (F_{ij}) de chaque pixel (P_{ij}) de l'image monoculaire en couleurs ($I0$; I1.1) est estimée, la fréquence estimée (F_{ij}) étant fournie en entrée du modèle d'apparence couleur inverse (1b) pour obtenir les valeurs trichromatiques corrigées (X'_{ij} , Y'_{ij} , Z'_{ij}) dudit pixel (P_{ij}).
- [Revendication 4] Procédé selon la revendication 3, dans lequel l'estimation fréquentielle (E13 ; E13.1, E13.2) comprend une décomposition (E13.1) de l'image monoculaire en couleurs ($I0$; I1.1) en une pluralité de domaines fréquentiels (s_0 - s_{16} ; s_k) et une décomposition (E13.2) de l'image monoculaire en couleurs ($I0$; I1.1) selon plusieurs orientations spatiales (E13.2), la fréquence (F_{ij}) étant estimée en tenant compte de l'influence des domaines fréquentiels adjacents au domaine fréquentiel auquel appartient ledit pixel (P_{ij}) de manière à obtenir une carte spatio-fréquentielle des pixels (I1.3).
- [Revendication 5] Procédé selon l'une quelconque des revendications 1 à 4, dans lequel le modèle d'apparence couleur inversible (1a) est le modèle CIECAM2 configuré avec selon au moins l'un des paramètres suivants : coordonnées chromatiques d'un blanc de référence (X_w , Y_w , Z_w), luminosité ambiante (L_a), intensité lumineuse de fond (Y_b), indicateur d'environnement (E).
- [Revendication 6] Procédé selon l'une quelconque des revendications 1 à 5, comprenant en outre les étapes suivantes :
- filtrer (E2) l'image monoculaire en couleurs ($I0$) en appliquant un filtre quaternionique linéaire guidé par une image de référence (I), de manière à fournir une image quaternionique filtrée (q ; I2) comprenant des données de textures de l'image monoculaire ($I0$) ; et
 - soustraire (E3) l'image filtrée (I2) de l'image corrigée en couleurs (I1), de manière à obtenir une carte de textures (I3).
- [Revendication 7] Procédé selon la revendication 6, comprenant en outre les étapes suivantes :

- sur-échantillonner (E ; E4.2) la carte de profondeur, de résolution réduite, (I4.1) par application d'un réseau hiérarchique (4.2) de manière à fournir des caractéristiques de profondeur sur-échantillonnées (I424) ;
- concaténer (E4.24) la carte des textures (I3 ; I'3) et les caractéristiques de profondeur sur-échantillonnées (I413) pour fournir une carte de profondeur résiduelle (I4.2) ; et
- additionner (E5) la carte de profondeur résiduelle (I4.2) et la carte de profondeur réduite (I4.1), de manière à obtenir une carte de profondeur finale (I5) d'une résolution au moins égale à celle de l'image monoculaire (I0).

[Revendication 8]

Dispositif (100 ; 600) pour la mise en œuvre du procédé d'estimation selon l'une quelconque des revendications 1 à 7, comprenant :

- un dispositif de captation optique monoculaire (0) pour obtenir une image monoculaire en couleurs (I0) dans des conditions d'observations réelles ;
- un module de correction colorimétrique (1) de l'image monoculaire en couleurs (I0) basé sur un modèle d'apparence couleur inversible (1a), de manière à fournir une image corrigée en couleurs (I1) dans des conditions d'observation de référence, de sorte que la perception des couleurs dans l'image corrigée en couleurs (I1) soit identique à la perception des couleurs dans l'image monoculaire en couleurs (I0) acquise dans les conditions d'observations réelles ; et
- un module de génération (4) d'une carte de profondeur (I4.1 ; I5) configuré pour appliquer un réseau de neurone convolutif (4 ; 4.1, 4.2) à l'image corrigée en couleurs (I1).

[Revendication 9]

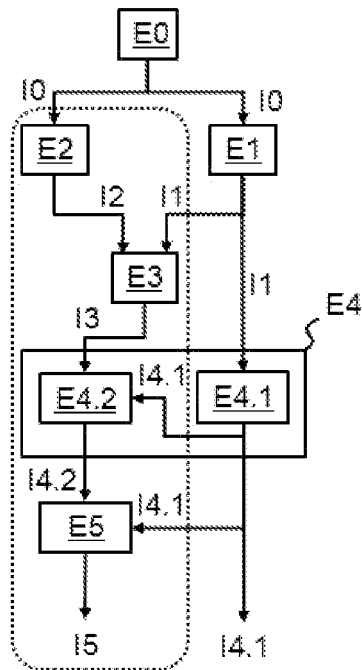
Dispositif (100 ; 600) selon la revendication 8, comprenant en outre :

- un module de filtrage (2) quaternionique linéaire guidé par une image de référence (I) configuré pour filtrer l'image monoculaire en couleurs (I0) de manière à obtenir une image filtrée (q ; I2) comprenant des données de texture de l'image monoculaire en couleurs (I0) ; et
- un module de soustraction (3) configuré pour soustraire

l'image filtrée (I2) de l'image corrigée en couleurs (I1).

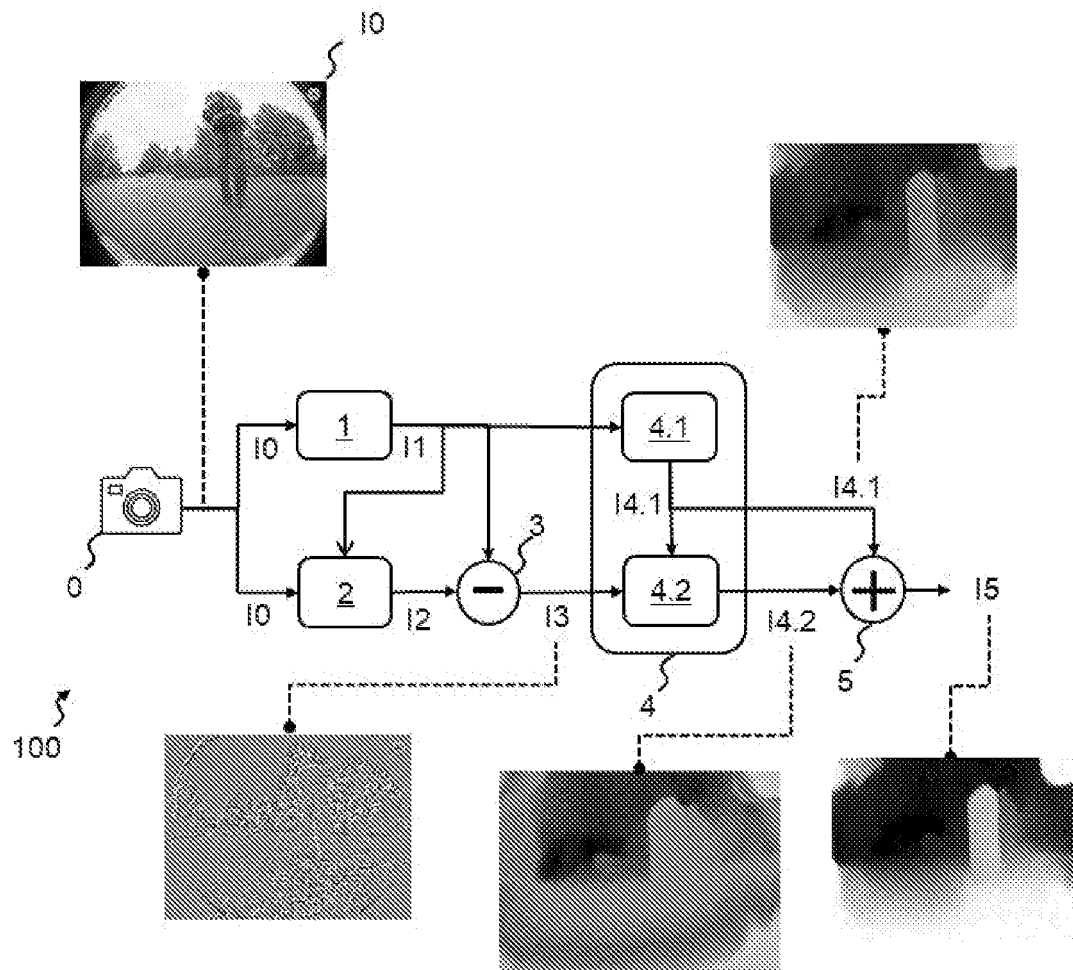
[Revendication 10] Engin de mobilité autonome (700 ; 800) comprenant un dispositif (100 ; 600) selon la revendication 8 ou 9.

[Fig. 1]

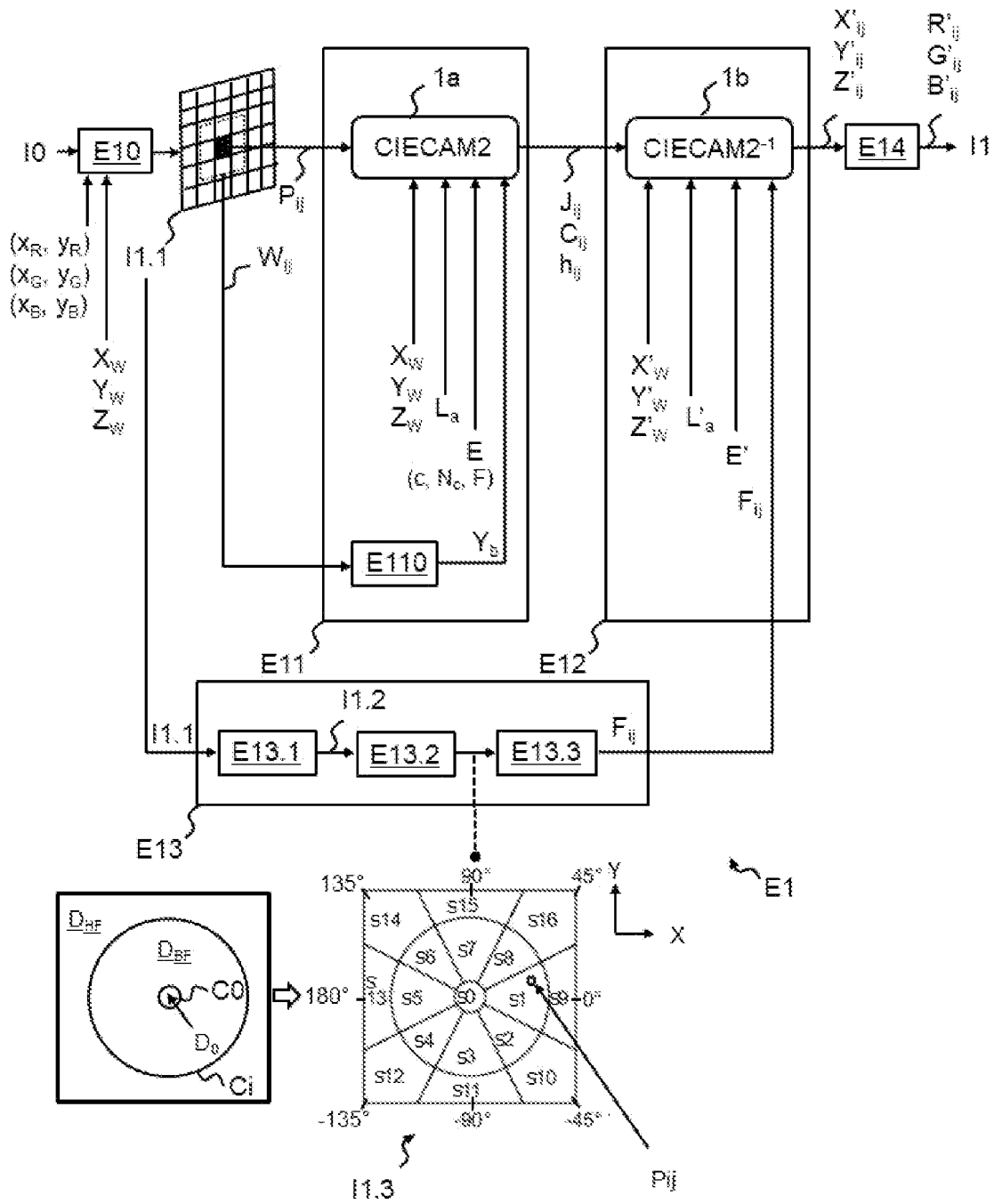


§
E100

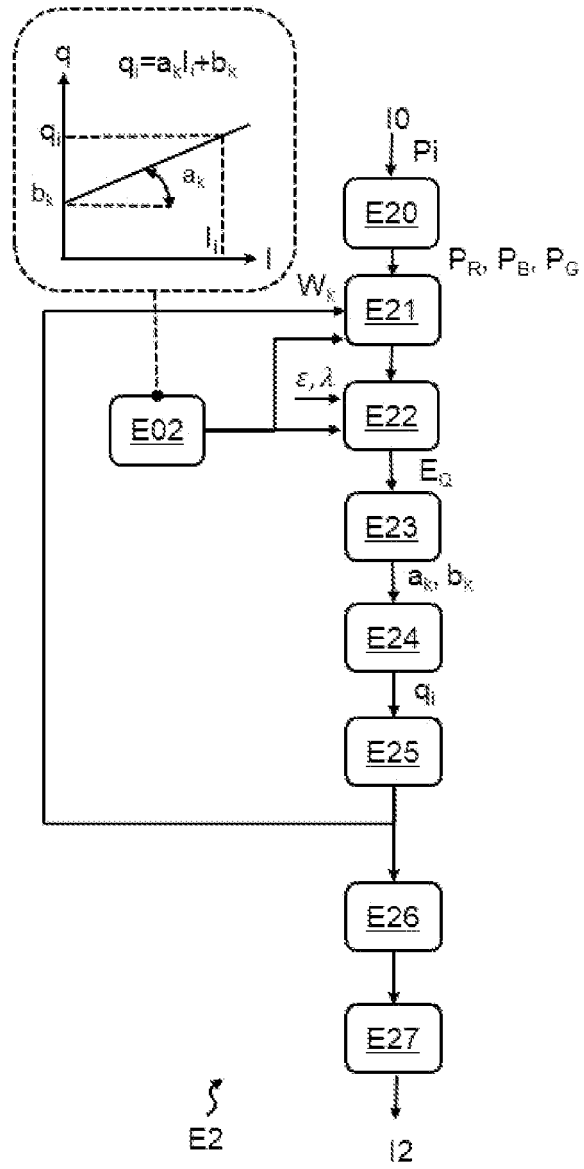
[Fig. 2]



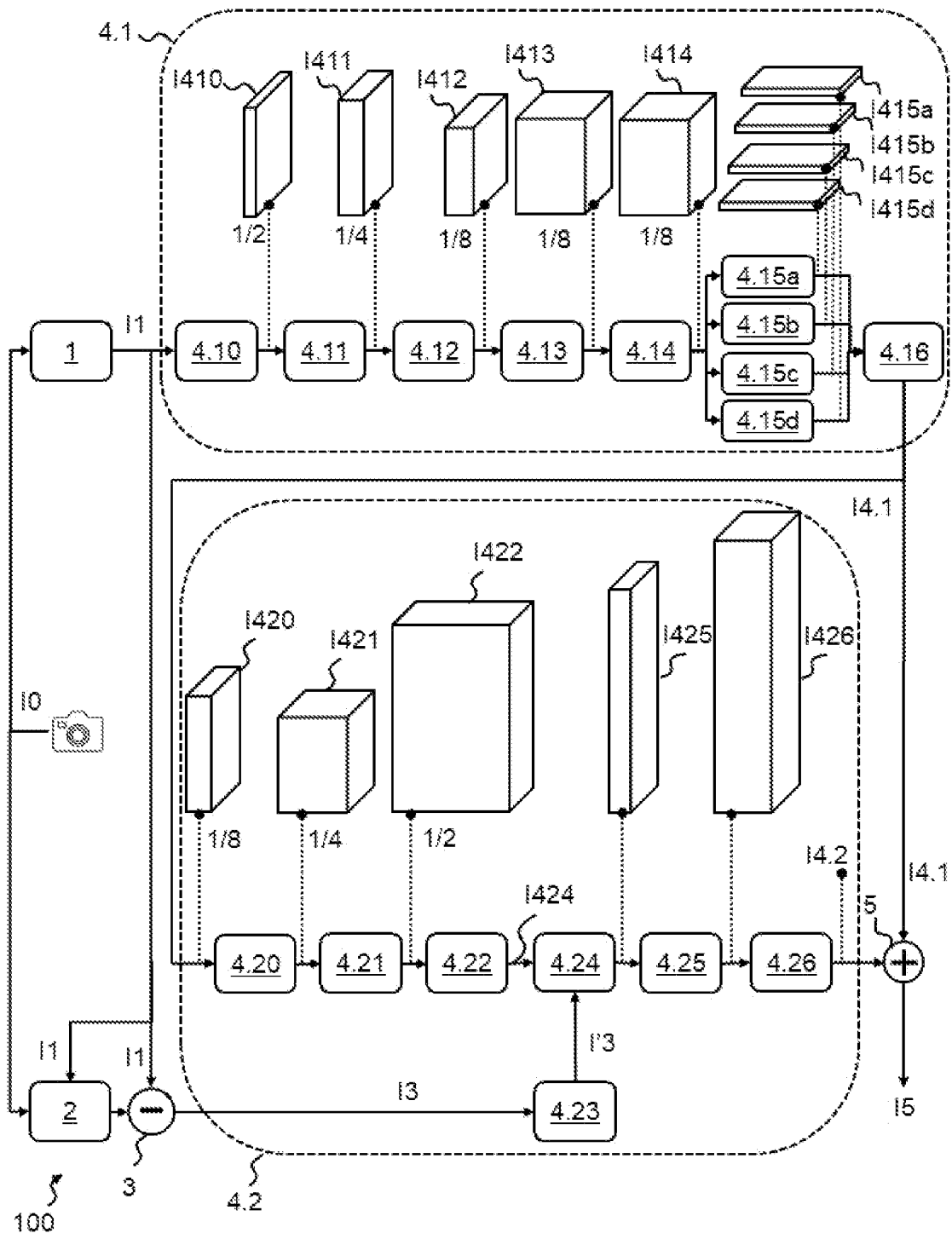
[Fig. 3]



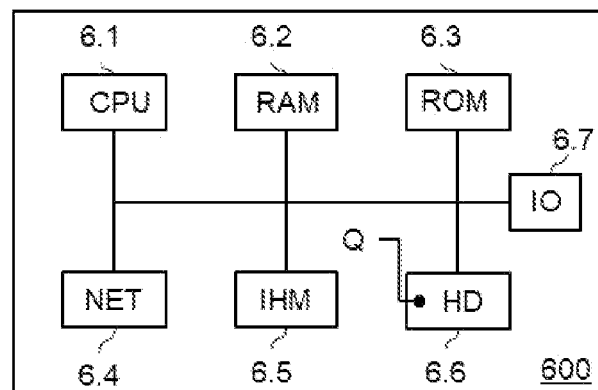
[Fig. 4]



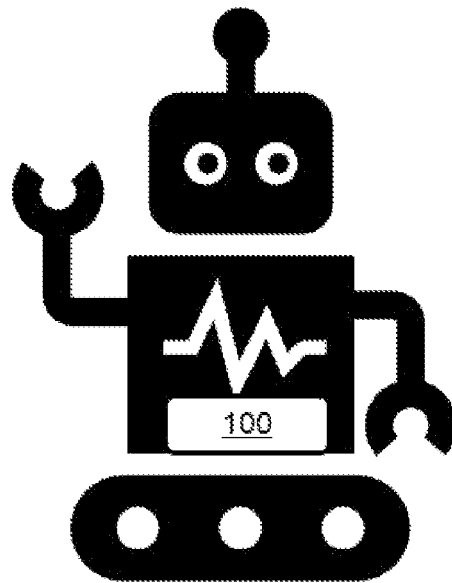
[Fig. 5]



[Fig. 6]

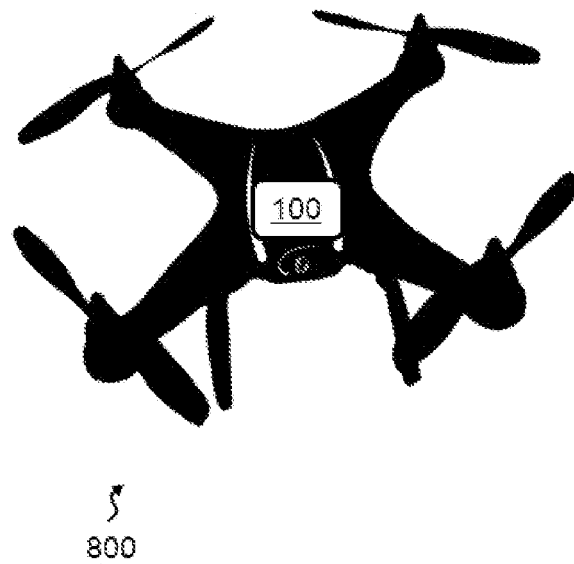


[Fig. 7]



700

[Fig. 8]



RAPPORT DE RECHERCHE PRÉLIMINAIRE

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

N° d'enregistrement
national

FA 917892
FR 2302116

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, D	NAN YANG ET AL: "D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 2 mars 2020 (2020-03-02), XP081612497,	1, 8, 10	G06N 3/02 G06T 5/00 G06T 7/50
Y	* abrégé *	2, 3, 5	
A	* page 1 – page 2, colonne 1 * * page 4, colonne 1 * * page 6, colonne 1 *	4, 6, 7, 9	
Y	YIN LING ET AL: "Stereoscopic chroma key matting using statistical analysis in CIECAM02 color space", 2015 IEEE INTERNATIONAL SYMPOSIUM ON HAPTIC, AUDIO AND VISUAL ENVIRONMENTS AND GAMES (HAVE), IEEE, 11 octobre 2015 (2015-10-11), pages 1-5, XP032835295, DOI: 10.1109/HAVE.2015.7359452 [extrait le 2015-12-16]	2, 3, 5	
A	* abrégé * * page 2, colonne 2 *	1, 4, 6-10	DOMAINES TECHNIQUES RECHERCHÉS (IPC) G06T
A	DENIS ET AL: "Spatial and spectral quaternionic approaches for colour images", COMPUTER VISION AND IMAGE UNDERSTANDING, ACADEMIC PRESS, US, vol. 107, no. 1-2, 1 juin 2007 (2007-06-01), pages 74-87, XP022103081, ISSN: 1077-3142, DOI: 10.1016/J.CVIU.2006.11.019 * abrégé *	1-10	
Date d'achèvement de la recherche		Examineur	
18 octobre 2023		Millet, Christophe	
CATÉGORIE DES DOCUMENTS CITÉS		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	
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			