Méthodes de Reconstruction Denses pour la Vision Active

Dense Reconstruction Methods for Active Vision

par Emmanuelle CLERGUE et Thierry VIÉVILLE

INRIA, Unité de Recherche de Sophia-Antipolis 2004, route des Lucioles 06902 Sophia-Antipolis Cedex
e-mail : clergue @eurecom.fr

résumé et mots clés

Ce papier cherche à analyser comment introduire des données tridimensionnelles au sein d'un système de vision active. En effet, nous nous sommes proposés de réaliser une reconstruction dense 3D à partir de l'analyse d'une séquence monoculaire, dans un paradigme de vision active. Nous présentons tout d'abord les différents algorithmes déjà existants dans le domaine de la reconstruction dense, en faisant ressortir leurs avantages et leurs inconvénients. Nous décrivons, ensuite, l'algorithme choisi pour pallier à certains de ces inconvénients. Enfin, nous montrons quelques résultats à partir d'images synthétiques ou de vues réelles acquises par la tête artificielle.

Vision active, Reconstruction dense en 3D rapide, Segmentation région, Régularisation.

abstract and key words

This paper aims to analyse how to introduce 3D information into an active vision system. In order to do so, we propose to realize a dense reconstruction from a monocular sequence in an active vision application. We first present existing algorithms for dense reconstruction. After having compared their drawbacks and advantages, we describe the chosen algorithm. Finally, we show the results obtained from synthetic images and images acquired when using an artificial robotic head.

Active vision, Fast 3D reconstruction, Region segmentation, Regularization.

1. introduction

La vision active connaît depuis ces dix dernières années un essor considérable. Les recherches s'orientent vers des tâches de plus en plus complexes et les systèmes visuels doivent se montrer de plus en plus autonomes. De nos jours, on veut non seulement obtenir des informations précises par le traitement d'images, mais aussi que ces informations puissent commander et contrôler un système robotique tel qu'un robot mobile, un bras manipulateur ou une tête artificielle. Nous entrons alors dans le domaine de la vision active. De nombreuses tâches nécessitent une représentation du monde dans lequel évolue le complexe visuel-robotique.

On peut remarquer que les mesures obtenues directement dans une image sont des primitives 2D (points, droites, segments, courbes ...). Il est donc a priori plus simple d'appliquer des algorithmes 2D, utilisant tels quels ces attributs. Le passage à une véritable représentation 3D introduit de nouveaux paramètres et rend donc le système d'état plus complexe. Par conséquent le temps d'exécution va augmenter et l'observabilité et la stabilité du système risquent d'être affectées elles aussi. Ainsi, pour conserver la robustesse et l'efficacité, on préfère souvent travailler avec des algorithmes 2D.

Par contre, si pour plusieurs tâches visuelles cette représentation simplifiée est suffisante, pour d'autres, qui requièrent des données plus précises, on a besoin explicitement de la troisième dimension. Par exemple cela devient nécessaire pour des applications telles que le positionnement et la manoeuvre d'un véhicule, l'évitement d'obstacles fixes ou en mouvement, la mise en correspondance 3D pour la surveillance ...

La vision active requiert donc de nouveaux algorithmes de plus grande efficacité, capables de donner des représentations 3D de l'environnement et de les traiter en temps réel pour le contrôle du système visuel-robotique associé.

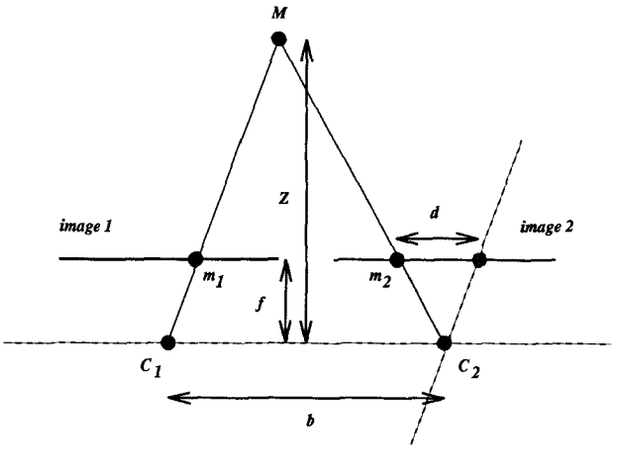

Figure 1. – La disparité d dans le cadre d'une géométrie standard.

2. méthodes de reconstruction dense

C'est dans ce cadre que nous nous sommes proposés de réaliser une reconstruction dense 3D à partir de l'analyse d'une séquence monoculaire, sur un système de vision active. Tout d'abord, nous présentons les différents algorithmes déjà existants dans le domaine de la reconstruction dense, en faisant ressortir leurs avantages et leurs inconvénients. Nous expliquerons ensuite le choix de l'algorithme utilisé dans la prochaine section.

En se plaçant dans le cadre de la géométrie standard [8], nous avons tout d'abord la relation de base liant en chaque pixel (x, y), la disparité $d(x, y)$ à la profondeur $z(x, y)$:

$$z(x, y) = \frac{b.f}{d(x, y)} \qquad (1)$$

où b représente la distance entre les centres optiques des deux caméras et f la distance focale voir figure ().

Retrouver la profondeur revient alors à calculer la disparité connaissant ces paramètres b et f.

A partir de cette relation de base, toutes les méthodes générales de reconstruction dense 3D, qu'elles utilisent une approche contours ou régions, se décomposent en 2 étapes :

– mise en correspondance des attributs entre deux vues qui permet d'obtenir une carte de disparités comportant éventuellement des "trous" et/ou des erreurs,

– complétion et affinage de la carte de disparités obtenue lors de l'étape précédente.

Nous allons faire une brève présentation des principales méthodes actuellement disponibles.

La "stéréoscopie par corrélation" est basée sur le fait que la disparité entre deux points corrélés est liée à la profondeur de ce même point en 3D. Par exemple la relation (1) représente le lien direct entre ces deux entités dans le cadre de la géométrie standard. Connaissant certaines mises en correspondance entre les deux vues, on en déduit alors une carte des disparités pour les points corrélés, ce qui permet de remonter à la profondeur par exemple à l'aide de cette relation.

La recherche des correspondances entre les deux images peut se faire de différentes manières. On peut par exemple utiliser des **fenêtres de corrélation** [9] : on calcule un score de corrélation à l'intérieur de la fenêtre considérée, basé sur la similitude entre les deux vues des niveaux de gris, des gradients ... ; la mise en correspondance est validée, soit pour le maximum de ce score, soit dès que ce score est supérieur à un certain seuil d'acceptabilité, ce qui est moins robuste mais plus rapide. Souvent, si on est dans le cas de l'analyse d'une séquence d'images où la disparité entre deux vues n'est pas très importante (par exemple en vision active) on peut alors restreindre la recherche du correspondant dans une région proche [30]. Le critère utilisé dans ces fenêtres de corrélation est basé sur la ressemblance d'une vue à l'autre, puisque le critère tient compte des informations contenues dans le voisinage défini par la fenêtre sélectionnée. Cette hypothèse entraîne évidemment de nombreux problèmes au niveau des zones d'**occultation**, puisque ce sont des parties qui ne sont visibles que dans une des vues. Si l'on cherche absolument à les apparier dans les deux vues, alors de fausses mises en correspondance apparaîssent. Une solution consiste à se donner un seuil sur le critère de corrélation : le nombre de fausses mises en correspondance diminue mais certaines régions restent alors sans correspondant [9]. Enfin, il existe aussi des travaux sur l'utilisation de fenêtres de corrélations adaptatives [18] : la taille et la forme de ces fenêtres tiennent compte des fortes variations d'intensité dans l'image; ces variations sont susceptibles de correspondre aux contours des objets dans la scène où se situent précisément les zones d'occultation et les **discontinuités de profondeur**.

Pour limiter le nombre de correspondants possibles, on peut utiliser la **contrainte épipolaire**. En fait, dans le cas de deux vues, le correspondant est à chercher sur une droite prédéterminée appelée **droite épipolaire**, qui est en fait l'image dans la deuxième vue de la droite passant par le centre optique de la première caméra et le point dont on cherche le correspondant. L'ensemble des correspondants possibles est alors réduit à une droite. On peut ainsi faire glisser le long de cette épipolaire une fenêtre de corrélation. Des attributs photométriques permettent de choisir le correspondant sur cette épipolaire.
Dans le cas de trois vues, la double contrainte épipolaire donne la possibilité de vérifier si le correspondant choisi se trouve bien à l'intersection des deux épipolaires dans la troisième image, c'est à dire si cette intersection correspond elle aussi à un pic de corrélation [8].
On peut aussi chercher les mises en correspondances à partir des **paramètres de calibration**. Lorsque l'on connaît les matrices de projection ainsi que le déplacement rigide entre deux vues, on peut mettre en correspondance les points des images, comme dans le cas précédent.

On suppose donc que l'on connaît les matrices de projection ainsi que le mouvement entre les deux vues, décomposé ici en une rotation et une translation.
Cette mise en correspondance donne de bons résultats pour les parties communes des images considérées. En cas d'occultation les correspondances sont erronées et nécessitent, comme pour la plupart des méthodes, un traitement à part. De plus quelque soit la disparité entre les deux vues, certaines bandes qui bordent les images, n'étant pas communes, toutes les images, ne pourront être mises en correspondance.
Dans le cas où l'on ne connaît aucun paramètre de calibration, on peut utiliser cette mise en correspondance dans un schéma itératif avec une régularisation portant sur les paramètres à estimer. Par cette méthode, on obtient non seulement une carte de mises en correspondance qui va nous permettre de remonter à l'information 3D, mais aussi l'estimation des paramètres de calibration [31].

On peut aussi utiliser une **approche régions**, qui nous donnera une carte de disparités des régions et non pas en chaque point.
Une segmentation en régions est donc appliquée aux deux images, ce qui les transforme en "bandes dessinées". Ensuite, pour établir les correspondances, on calcule une matrice de scores de corrélation. On se donne un critère de corrélation qui peut par exemple prendre en compte la position des barycentres des régions sous l'hypothèse de disparité faible entre les deux vues, l'intensité moyenne, la taille .... A partir de là, on peut corréler en utilisant différentes méthodes.
Une première méthode consiste à évaluer ce critère pour toutes les corrélations possibles puis mettre en correspondances les paires de régions pour lesquelles le score est maximum, et ce tant que ce score est supérieur à un seuil de tolérance, en sachant qu'une région ne peut avoir qu'un seul correspondant.
Un deuxième méthode, moins robuste mais plus rapide, consiste à évaluer ce critère pour une région de la première image et chacune des régions de la deuxième en validant la mise en correspondance dès que la valeur du critère est supérieure à un seuil. Mais cette dernière méthode est très sensible aux ambiguïtés.
Enfin, une troisième méthode correspond à un schéma de propagation-vérification d'hypothèses [4, 8] qui permet de revenir en arrière si une correspondance s'avère erronée. Elle reste cependant difficile à mettre en oeuvre en temps réel [24].
Par ces méthodes de mise en correspondances, on peut remarquer que dans la plupart des cas, la carte des disparités obtenue n'est pas complète. Il faudra alors les compléter par des méthodes d'interpolation, qui peuvent tenir compte de contraintes sur le monde (continu, continu par morceaux ...). Certains travaux font interagir ces deux étapes en ne validant que les corrélations pour lesquelles la surface obtenue après interpolation est considérée comme la "meilleure" vis-à-vis des contraintes données [13].

Au niveau de la mise en correspondance de ses régions, d'importants travaux ont permis de mettre en oeuvre des techniques qui limitent une combinatoire coûteuse lors de l'appariement et diminue les faux appariements potentiels. Dans [16] une structure de donnée complexe basée sur une technique de tables relationelles permet de faire coopérer des attributs géométriques et des relations structurelles entre les différents attributs de différentes primitives afin d'accélérer l'appariement. La méthode est néanmoins très lourde à mettre en oeuvre. De manière plus simple, différentes approches basées sur des primitives de type région ont été proposées, soit en coopération avec des primitives ponctuelles [33], soit de manière hiérarchique [5], en appariant de grosses régions puis les sous-régions qui la composent, soit sous l'hypothèse que les régions correspondent à des objets plans [25] et des surfaces lambertiennes [20], comme c'est le cas lorsque ces techniques cherchent à introduire des caractéristiques photogramétriques dans les critères d'appariement proposés [10].

Dans toutes ses approches dédiées à la stéréoscopie, il n'y a pas utilisation du fait que les disparités entre deux vues consécutives sont faibles ce qui a pour effet de réduire considérablement le champ de recherche. On peut alors comme dans [35] calculer de manière très rapide des candidats à l'appariement à partir d'un critère raisonnable et raffiner cet appariement par relaxation, c'est à dire en appariant d'abord les amers les plus probablement en relation et de se servir de ces premières estimations pour aider à résoudre les ambiguïtés posées par des amers moins faciles à apparier. C'est une technique de ce type qui sera mise en oeuvre ici.

La **"méthode de régularisation"** est une **méthode variationnelle** accompagnée d'un terme de **régularisation** qui permettra de remonter à l'information de profondeur. Elle se présente sous forme d'une **énergie à minimiser** [34, 26], comme introduit par [32]. Cette énergie est exprimée comme la somme de deux termes : le premier est un terme de vraisemblance sur la corrélation et le second un terme de contrainte sur le monde. De la même façon que précédemment, on part du fait que d'une image à l'autre, le point observé conserve certaines propriétés (intensité, focalisation, ...). En revanche, la méthode étant itérative, on a besoin d'une carte de disparités de départ.

Le terme de vraisemblance doit caractériser la carte de disparité en entier et non seulement en chacun des pixels de l'image. Il sera donc de la forme :

$$\varepsilon(d) = \int_u \int_v \sum_k (I_1^k(u,v) - I_2^k(u+d_u, v+d_v))^2 du dv$$

u et v recouvrant le domaine pour lequel on veut obtenir une carte de disparités [26].

Le terme de contrainte permet, par exemple, de rendre compte, au sein du modèle, de la continuité ou de l'élasticité de la courbe représentant le champ de disparité. Voici deux termes de régularisation d'ordre 1 et 2 :

$$\Psi_1(d) = \int_u \int_v \| \nabla d \|^2 \, du dv$$

qui entraîne uniquement une contrainte de continuité de la disparité, et :

$$\Psi_2(d) = \Psi_1(d) + \mu \int_u \int_v ((\frac{\partial^2 d}{\partial u^2})^2 + 2(\frac{\partial^2 d}{\partial u \partial v})^2 + (\frac{\partial^2 d}{\partial v^2})^2) du dv$$

qui impose aussi que la direction de la normale à la surface varie de façon continue.

Ces contraintes vont permettre de diminuer le nombre des solutions possibles pour le terme de vraisemblance. On doit minimiser une énergie du type $\varepsilon(\mathbf{d})$, qui est en général non convexe. Rajouter ces contraintes va accroître les chances de tomber sur le minimum global au lieu d'un minimum local lors de la convergence de cette méthode itérative (convexification). On voit donc à présent tout l'intérêt d'avoir une bonne initialisation de la carte de disparités afin d'augmenter là aussi les chances de converger vers le minimum global. On peut alors utiliser une approche pyramidale ou multi-résolution comme dans les travaux de [26, 34, 19]. Le principe est de considérer différentes résolutions de l'image de départ pour le schéma itératif de la régularisation. On part d'une résolution grossière de l'image et d'une carte de disparités modélisée par exemple par un plan fronto-parallèle. A ce niveau de résolution, le nombre de minima locaux est bien moindre et par conséquent le risque de converger vers l'un d'eux est fortement réduit. Par la régularisation, on va raffiner la carte de disparités pour cette résolution et se servir de cette nouvelle estimation pour la régularisation à un niveau de résolution plus fin ... Ainsi nous obtenons avec plus de sûreté, la carte de disparités cherchée.

L'**"approche par les champs de Markov"** [22, 17, 12, 1] utilise une modélisation par **Champs de Markov** ainsi qu'un estimateur Bayesien.

Pour cette méthode on va aussi devoir minimiser une énergie. On doit là aussi caractériser au mieux la carte de disparités que l'on cherche.
Comme précédemment, l'énergie est en général constituée de deux termes :

– terme de ressemblance (vraisemblance)

– terme de contrainte sur le modèle (continu, lisse, continu par morceaux ...)

Le choix de l'énergie est donc fait d'une manière similaire au cas de la régularisation. Le premier terme de l'énergie portera sur la ressemblance entre le label choisi, correspondant ici à une disparité, pour le site et la donnée réelle correspondante et ce pour tous les sites de l'image. Dans ce terme on tient donc compte des données et des étiquettes qui caractérisent les régions.
Le deuxième terme représente l'énergie dans un voisinage donné. Cette énergie tient compte des relations de voisinage et n'est fonction que des étiquettes de régions.
Le minimum d'énergie correspond à un maximum de probabilité pour la segmentation.
Le calcul de la carte de disparités revient à trouver l'étiquette en chacun des sites correspondant à un maximum de cette probabilité.

Pour minimiser cette énergie on a, à notre disposition, des méthodes déterministes ou non déterministes.

Des méthodes déterministes sont, par exemple, **ICM** et **HCM** [22, 12] qui sont rapides (ICM converge au bout de cinq à dix itérations et HCM au bout de quatre à cinq itérations). Cependant, elles nécessitent une bonne initialisation afin de ne pas converger vers un minimum local de l'énergie.

Une méthode non déterministe est le **recuit simulé**. Elle converge beaucoup plus lentement (de 30 à 1000 itérations), mais permet de trouver le minimum global, ce qui correspond mieux à la volonté de modéliser précisément une carte de disparités. On se sert alors d'un estimateur Bayesien pour le calcul des probabilités correspondant au choix d'une étiquette en chaque voisinage. On part en fait d'une température élevée, c'est à dire d'un critère fortement convexifié, on visite les différents sites pour les mettre à jour. Puis on diminue la température et on recommence.
Le fait de commencer avec une température élevée permet aux sites de pouvoir changer d'étiquette facilement donc de pouvoir passer les minima locaux de l'énergie plus aisément.
La remise à jour des sites peut se faire soit par la méthode **Metropolis**, soit par **l'échantillonneur de Gibbs**. La différence entre ces deux méthodes réside au niveau du choix du nouveau label.
Les sites peuvent être "visités" de différents façon :

– aléatoirement, ce qui implique un nombre important de visites pour toucher tous les sites ou au moins un maximum afin que la nouvelle estimation soit cohérente, – séquentiellement,

– parallèlement, ce qui implique un choix des ensembles de sites à voir afin de ne pas remettre à jour des sites appartenant au voisinage d'un autre site.

On peut facilement tenir compte des discontinuités en rajoutant des processus de lignes qui tiennent compte de la présence de ces discontinuités entre un site et ses voisins.
Tous les détails de ces méthodes appliquées à la reconstruction surfacique sont dans les travaux [22, 17, 1].

L'**"approche différentielle"** [28] utilise des **contraintes différentielles** pour la reconstruction. En détectant certains points particuliers comme ceux appartenant aux contours d'occultation, on obtient des contraintes géométriques en ces points, ce qui permet de reconstruire le volume 3D au moins partiellement.
Cette méthode, contrairement aux autres, ne va pas donner une carte dense de l'environnement 3D en une seule étape. En effet, elle se base sur les caractéristiques géométriques de certains points particuliers de l'images. Ces points représentent des points spécifiques d'une surface. On a donc une approche de reconstruction 3D partielle sur certaines régions particulières des objets dans la scène. Même si ces reconstructions se trouvent finalement dans une même carte de l'environnement, il s'agit là plutôt d'une approche sur les différents objets d'une même scène.
Cette méthode se base sur la détection de ces points particuliers qui permettent de remonter à des informations 3D localement sur la surface de l'objet étudié. Par exemple dans les travaux [28], on utilise les contours d'occultation pour la reconstruction 3D. Ces

contours d'occultation sont les limites des surfaces régulières, comme les limites d'un cylindre. Ils ne correspondent pas à des arrêtes, mais sont tout de même des contours de l'objet, bien qu'ils ne soient pas fixes sur l'objet. Au niveau de ces contours spécifiques, d'une vue à l'autre, il n'y a pas discontinuité de la normale à la surface, mais il y a discontinuité de la distance à la caméra : les points appartenant à un contour d'occultation dans une vue ne seront plus sur ce même contour dans une autre vue. Ces contours font donc échouer les méthodes stéréoscopiques classiques de reconstruction. Un moyen de les détecter est la violation par ces contours de la contrainte épipolaire présentée précédemment. Une fois qu'ils sont localisés, on peut alors les utiliser pour remonter à des informations géométriques différentielles qui permettent de reconstruire localement la surface de l'objet considéré. Si on arrive à suivre les contours d'occultation d'un même objet tout au long d'une séquence d'images, on arrivera à reconstruire le profil de l'objet.

Les **"méthodes optiques"** [3, 2, 23, 14, 15, 29] utilisent les informations obtenues à partir de la **mise au point** (focus), de la **vergence** et de la **calibration** : la mise en correspondance, puis la comparaison d'images obtenues avec différents paramètres optiques et mécaniques permettent d'estimer la disparité ou la profondeur.

Par ces méthodes, on peut, par exemple, reconstruire la surface des objets dans une scène grâce aux informations obtenues à l'aide du focus et de la vergence [2, 30].

Regardons d'abord comment obtenir une information 3D à partir de la vergence [3, 2]. On travaille sur des paires stéréoscopiques. Connaissant la distance entre les deux centres optiques des caméras, ainsi que les angles de rotation pour chacune d'elle, on peut alors remonter facilement à la profondeur du point d'intersection des deux lignes de vues par triangulation. Ce point est appelé point de fixation. Le problème est alors de savoir si les deux caméras fixent bien le même point 3D de la scène. On peut alors évaluer la disparité au centre de l'image et la réduire à zéro. L'estimation de cette disparité peut être faite à l'aide des méthodes précédemment décrites.

On peut aussi obtenir la profondeur d'un point dans l'image grâce au focus [3, 2, 23, 27]. C'est un processus monoculaire. On a alors différents moyens de remonter à l'information 3D. Dans [27] par exemple, c'est la dégradation de la netteté entre deux images entre lesquelles on fait uniquement varier le focus qui va permettre de remonter à la profondeur cherchée. C'est une mesure relative qui est utilisée et non une mesure exacte. Dans [23], on a la même approche par dégradation de la netteté, accompagnée d'un critère de netteté. Une fois le point de netteté trouvé, en connaissant les paramètres de calibration, on peut alors remonter à l'information 3D cherchée.

C'est en principe la dégradation de la netteté qui est principalement utilisée, car elle est plus simple et plus rapide.

On pourra voir dans la partie suivante comment de telles informations obtenues à partir des paramètres optiques peuvent être très utiles pour la détection des discontinuités.

3. choix de la méthode de reconstruction proposée

Les principaux problèmes de la reconstruction 3D à partir de la stéréo sont la détection et le traitement des Occultations (**O**) et des Discontinuités de Profondeur (**DP**). On fait en général l'hypothèse que la profondeur est constante par morceaux ou continue. Ces contraintes ne sont pas adaptées pour la reconstruction 3D car elles correspondent à une représentation trop simplifiée du monde réel.

La profondeur dans une scène réelle se rapproche plus d'un modèle continu par morceaux, permettant d'admettre des **DP** aux bords des régions continues détectées.

De plus les **O** interviennent lorsque deux objets dans la scène se recouvrent partiellement. Les problèmes entraînés par les **DP** ainsi que les **O** se localisent aux frontières apparentes des objets, qui nécessiteront un traitement spécifique. De nombreux travaux ont été menés dans ce sens, que ce soit pour la localisation ou le traitement des **O** et des **DP** [11, 13, 18, 22, 26, 27, 34].

Lors de la reconstruction, le fait de ne pas admettre de **DP** a pour effet de lisser la carte de l'environnement.

De part et d'autre d'une **DP**, on se trouve dans deux régions totalement différentes. Si la mise en correspondance n'est pas suffisamment précise, des pixels de chaque côté d'une **DP** peuvent être corrélés. De telles erreurs qui se jouent à quelques pixels, peuvent entraîner dans un schéma de régularisation, des rectifications importantes et fausses. On voit donc que la connaissance de la localisation des **DP** est très importante et permettrait aux méthodes de reconstruction d'adapter leur traitement.

Les **O** étant des parties cachées dans une des images, suivant les techniques de mise en correspondance, on va soit corréler des pixels ne représentent pas la même chose, soit ne trouver aucun correspondant dans l'autre image et être incapable d'estimer la profondeur. C'est donc aux frontières des objets occultants, où sont localisées les **O**, que l'on obtiendra des estimées aberrantes pouvant se répercuter par les relations de voisinage utilisées.

Il est donc nécessaire de localiser ces pixels spécifiques afin de les traiter séparemment.

Dans les travaux de [34], pour détecter les **O**, on part de la remarque suivante : aux endroits où un point dans l'image de droite n'a pas de correspondant dans l'image de gauche, les dérivées partielles de la fonction de la disparité dans la direction de l'épipolaire doivent être négatives et importantes en valeur absolue. Un critère retrace cette remarque et il suffit alors de voir quels sont les pixels qui le vérifient.

De même, on peut remarquer qu'aux discontinuités de profondeur, le gradient de la fonction de disparité aura un maximum local dans la direction de la normale à l'élément de contour. A partir de cette idée, deux critères simples décrivent cette configuration dans le cas d'une discontinuité horizontale et deux autres dans le cas d'une discontinuité verticale.

Une fois les **O** et le **DP** détectées, on utilise des fonctions de visibilité qui permettent de régulariser de façon adaptée : la propagation des contraintes de continuité est évitée en présence des **O** et des **DP**.

On peut raffiner encore cette méthode par la détection les **DP** horizontales et verticales. On peut alors séparer le terme de contrainte sur le modèle afin de ne pénaliser que la partie correspondante.

Le problème majeur de cette méthode vient du fait que toutes les décisions sont prises grâce à des seuils, souvent très difficiles à ajuster. On peut alors essayer de les rendre adaptatifs.

D'autres travaux, comme ceux de [11], se sont portés sur ce genre de fonctions de visibilité adaptative. On ne parle donc ici que de traitement des **O** et des **DP**. Ces fonctions de visibilité sont moins brutales. Elles tiennent en fait compte des informations locales. Une fois la carte de disparités approximée par la minimisation d'un critère de la forme :

$$C = \int s(I_1(u,v) - I_2(u,v,d(u,v))^2 + \lambda_u((\frac{\partial d(u,v)}{\partial u})^2 + \lambda_v(\frac{\partial d(u,v)}{\partial v})^2)$$

on peut alors réajuster les paramètres λ_u et λ_v en tenant compte des gradients de l'intensité et de la disparité. Ainsi on peut retrouver un profil moins lissé.

Dans la cas de l'approche par régions, on a directement une notion de frontières apparentes des objets, ce qui permet de localiser plus simplement les **O** et les **DP**.

Une fois les contours localisés par la segmentation par régions, il faut alors vérifier si ces pixels peuvent poser des problèmes lors de la reconstruction. Pour cela on peut utiliser plusieurs méthodes :

– vérifier s'il y a une cassure d'intensité au voisinage de ces points dans le cas des **DP**. En effet, si on a une transition importante du niveau de gris de part et d'autre du bord de la région détectée, on a des chances d'être en présence de deux objets différents.

– rajouter une contrainte du type épipolaire pour confirmer ou non la mise en correspondance de points de part et d'autre du contour, dans le cas des **O**, comme dans les travaux de [21].

– vérifier la taille des régions mises en correspondance. Si cette différence est trop importante, c'est qu'une partie de la région a été recouverte entre les deux vues . En fait, on peut appliquer ce même raisonnement en utilisant les informations obtenues à partir de la vergence [27]. C'est un processus monoculaire qui peut s'intégrer facilement lorsqu'on requiert plus d'informations. Il suffit alors de fixer la partie centrale de l'image sous un autre point de vue. On peut alors faire les mêmes vérifications que précédemment.

– utiliser les informations apportées par le focus ou le defocus afin d'estimer la profondeur par un autre moyen. En comparant ces résultats cela permet de discerner l'objet caché de l'occultant par estimation de la profondeur des part et d'autre du point considéré [27]. Ce processus est monoculaire mais très simple. Comme nous l'avons décrit précédemment, en étudiant la dégradation de la netteté entre deux images acquises avec le même point de vue et avec des focus différents, on peut alors estimer les profondeurs correspondantes, donc les positions relatives des régions. Les régions représentant les différents objets de la scène, on pourra savoir à l'avance où sont suceptibles de se produire les problèmes dûs aux **O**.

Pour ces raisons, nous avons choisi une approche par segmentation en régions en un premier temps, car elle est particulièrement bien adaptée à la détection et au traitement des **O** et des **DP**. Après la mise en correspondance des régions des deux vues, nous obtiendrons une carte des disparités. Ceci nous permettra d'avoir une carte initiale des proximités (inverse de la profondeur), sous forme de morceaux de plans fronto-parallèles. Cette estimation peut être approximative mais suffisamment représentative de la disposition des objets dans la scène. On appliquera alors un processus de régularisation afin d'affiner cette estimation 3D grossière de la scène.

Il nous faut donc un algorithme simple et rapide afin d'accomplir cette tâche.

4. description de la segmentation régions

Nous nous sommes basés sur les travaux de Fairfield [7].

On considère que les régions sont des parties de l'image où l'intensité est homogène. Les frontières des régions correspondent aux contours dans l'image et sont donc caractérisées par de

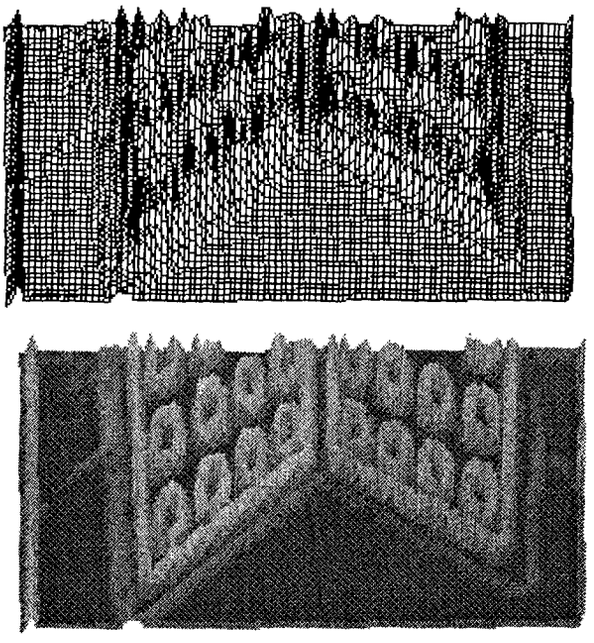

Figure 2. – Méthode toboggan.

forts gradients. Dans le cas où les régions se caractérisent par différentes textures, on suppose qu'un procédé de lissage a gommé les textures au profit d'une intensité moyenne dans ces régions. Donc à l'intérieur d'une région, les normes de gradients sont soit faibles ou en tout cas inférieur à un seuil fixant l'homogénéité, soit nulles. On va donc pouvoir relier les points d'une même région en reliant chaque point à son voisin de norme de gradient la plus faible. Cette méthode est dite du **"toboggan"** car on glisse automatiquement le long des forts gradients vers l'intérieur de la région dans laquelle on se trouve.

Elle est très simple et très rapide. Comme nous l'avons dit précédemment, nous voulons obtenir à terme une carte de proximité à partir des disparités sur les régions de l'image. Cette carte n'a pas besoin d'être très détaillée mais doit représenter grossièrement la disposition des principaux objets de la scène et permettre de prédire les **O** et les **DP**. On veut donc que la segmentation mettre en évidence les silhouettes les plus importantes sans avoir forcément les détails. Pour cela nous avons deux paramètres à notre disposition :

GradThres : le seuil sur la norme du gradient, qui intervient lors du calcul de l'image de la norme du gradient. Toutes les valeurs inférieures à ce seuil sont mises à zéro. Ce seuil permet de ne conserver que les endroits de forts contrastes.

SmoothFen : le facteur d'échelle d'un filtre de lissage qui nous permet de lisser plus ou moins l'image de départ.

Il faut donc trouver un bon équilibre entre ces deux paramètres, afin d'avoir en main une image bien conditionnée avant d'appliquer la segmentation.

Méthode automatique d'ajustement du lissage.

Il nous faut tout d'abord éliminer au maximum le bruit qui existe dans l'image au moyen d'un filtrage léger. Plus on va filtrer plus on va éliminer de détails puisqu'on atténue les contrastes. Mais si on lisse trop, on finit par aplanir complètement l'intensité et il n'y a alors plus vraiment de frontières. Entre ces deux extrêmes, on peut remarquer une phase que l'on peut qualifier de stationnaire. Plus précisément, pour une norme de gradient donnée, le nombre de régions en fonction d'un taux de lissage croissant décroît fortement en un premier temps, puis marque un palier avant de décroître à nouveau fortement. Ce sont les valeurs de lissage correspondant à ce palier, pour lesquelles le nombre de régions diffère peu, qui nous permettent d'obtenir un bon équilibre entre les deux paramètres de la régularisation que nous traitons ici. On obtient ainsi un optimum qui équilibre l'élimination de régions liées au bruit par lissage. La méthode et quelques résultats sont disponibles dans [30].

Comme à terme nous voulons reconstruire l'environnement à partir d'une séquence d'images, nous devons fixer ces paramètres pour un maximum de vues successives. En effet, si l'on fait varier un des deux paramètres, les régions ne vont pas être détectées dans les mêmes conditions et la mise en correspondance sera moins robuste. Par contre, le fait de travailler avec une séquence avec peu de disparité entre les vues, nous permet d'espérer que pour un bon nombre de vues successives, les régions importantes ne varieront pas beaucoup. L'ajustement des paramètres pourra alors être valable pour ces vues. Il semble normal au bout d'un certain temps de réajuster ces paramètres, mais il faudra alors, soit fusionner les cartes de proximités différentes obtenues avant et après ces ajustements, soit essayer de conserver toutes les informations de la carte précédente en y ajoutant simplement les informations obtenues pour les nouvelles parties de la scène.

Une fois cette partition faite, nous avons besoin de caractériser ces régions afin d'effectuer une mise en correspondance qui nous permettra d'estimer une carte de disparité correspondante.

5. mise en correspondance des régions

5.1. Représentation des régions

Nous allons donc, par un parcours de toute l'image, essayer de collecter un maximum d'informations pour la mise en correspondance. Tout d'abord, nous supposons que les disparités entre deux images consécutives d'une séquence sont faibles. Nous supposons aussi que les surfaces des objets dans la scène sont approximativement lambertiennes, c'est-à-dire que l'intensité ne dépend pratiquement que du point considéré sur la surface et non du point de vue. Ces hypothèses vont permettre de simplifier considérablement le critère de corrélation. De quoi avons nous besoin pour représenter une région dans ces conditions?

D'après la première hypothèse, le correspondant de la région considérée sera localisé à peu près au même endroit dans la deuxième image. Il nous faut caractériser simplement le déplacement d'une région puisque ce module doit fournir le plus rapidement possible une carte approximative de proximités de la scène. Nous avons choisi d'estimer le déplacement du barycentre de la région. Comme la disparité d'une vue à l'autre est peu importante cela veut aussi dire que la structure des régions va être peu modifiée (même en cas d'occultation, la surface recouverte entre deux vues sera faible ...). Nous allons donc récupérer les coordonnées des barycentres de chacune des régions, ainsi que leur taille en nombre de pixels.

D'après l'hypothèse de surfaces lambertiennes, les régions d'une vue à l'autre doivent conserver en moyenne le même niveau de gris. Nous allons donc récupérer la somme des intensités ainsi que le nombre d'éléments dans chaque région.

Pour affiner notre estimation, nous avons modélisé chaque région par un plan d'intensité. Ce modèle, dans le cas d'un éclairement lambertien, fournit une autre indication sur la disparité comme nous allons le calculer. Nous avons choisi une modélisation

dépendant d'un point (u_0, v_0) appartenant à la région $\Re$ et ce pour tout (u, v) appartenant à $\Re$:

$$I(u,v) = I_0 + I_u(u-u_0) + I_v(v-v_0) \quad (2)$$

avec $I_u = \frac{\partial}{\partial u} I(u_0, v_0)$, $I_v = \frac{\partial}{\partial v} I(u_0, v_0)$ et $I_0 = I(u_0, v_0)$.
On veut donc minimiser pour chaque région $\Re$ le critère de moindres carrés suivant :

$$J = \sum_{u,v \in \Re} [I - (I_0 + I_u(u-u_0) + I_v(v-v_0))]^2 \quad (3)$$

où **I** représente l'intensité dans l'image. Le critère étant quadratique et positif, il suffit de rechercher le minimum global qui correspond à l'annulation du gradient du critère.
On dérive alors J par rapport à chacune des inconnues (I_0, I_u, I_v) :

$$\begin{cases} 0 = \frac{1}{2}\frac{\partial J}{\partial I_0} \\ \quad = \sum_{u,v\in\Re}[I - (I_0 + I_u(u-u_0) + I_v(v-v_0))] \\ 0 = \frac{1}{2}\frac{\partial J}{\partial I_u} \\ \quad = \sum_{u,v\in\Re}[I - (I_0 + I_u(u-u_0) + I_v(v-v_0))](u-u_0) \\ 0 = \frac{1}{2}\frac{\partial J}{\partial I_v} \\ \quad = \sum_{u,v\in\Re}[I - (I_0 + I_u(u-u_0) + I_v(v-v_0))](v-v_0) \end{cases}$$

Grâce à ces trois dérivées, on obtient une système de trois équations linéaires qui nous permet d'estimer les quantités (I_0, I_u, I_v).

$$\begin{array}{l} \sum_{u,v\in\Re} I = \\ \quad I_0 \sum_{u,v\in\Re} 1 + I_u \sum_{u,v\in\Re}(u-u_0) + I_v \sum_{u,v\in\Re}(v-v_0) \\ \sum_{u,v\in\Re} I(u-u_0) = \\ \quad I_0 \sum_{u,v\in\Re}(u-u_0) + I_u \sum_{u,v\in\Re}(u-u_0)^2 + I_v S_{uv} \\ \sum_{u,v\in\Re} I(v-v_0) = \\ \quad I_0 \sum_{u,v\in\Re}(v-v_0) + I_u S_{uv} + I_v \sum_{u,v\in\Re}(v-v_0)^2 \end{array}$$

avec : $S_{uv} = \sum_{u,v\in\Re}(u-u_0)(v-v_0)$.
Notons pour simplifier $\sum$ au lieu de $\sum_{u,v\in\Re}$.
Si on choisit le barycentre de la région comme le point de coordonnées $(u_0, v_0)^T$, on doit alors résoudre le système simplifié mis sous forme matricielle suivant :

$$\begin{pmatrix} \Sigma 1 & 0 & 0 \\ 0 & \Sigma u^2 - 2u_0\Sigma u\Sigma 1 + u_0^2\Sigma 1 & S_{uv} \\ 0 & S_{uv} & \Sigma v^2 - 2v_0\Sigma v\Sigma 1 + v_0^2\Sigma 1 \end{pmatrix} \cdot \begin{pmatrix} I_0 \\ I_u \\ I_v \end{pmatrix} = \begin{pmatrix} \Sigma I \\ \Sigma I * u - u_0\Sigma I \\ \Sigma I * v - v_0\Sigma I \end{pmatrix}$$

avec : $S_{uv} = \Sigma uv - v_0\Sigma u - u_0\Sigma v + u_0 v_0 \Sigma 1$

Après avoir collecté toutes les régions et avoir modélisé chacune de ces régions par un plan, nous avons à notre disposition les informations suivantes pour chaque région r :

$\mathbf{u_0^r}, \mathbf{v_0^r}$: les coordonnées du barycentre,
$\mathbf{Size^r}$: le nombre de pixels appartenant à la région,
$(\mathbf{I_0^r}, \mathbf{I_u^r}, \mathbf{I_u^r})$: les paramètres de modélisation par un plan.

5.2. obtention d'une carte de disparités

Notre score de corrélation entre une région **I** d'une première vue et une région **J** d'une seconde sera alors de la forme :

$$\begin{array}{rl} S(I,J) = & \alpha[(u_0^I - u_0^J)^2 + (v_0^I - v_0^J)^2] \\ & +\beta\,[(Size^I - Size^J)^2] \\ & +\gamma\,[(I_0^I - I_0^J)^2 + (I_u^I - I_u^J)^2 + (I_v^I - I_v^J)^2] \end{array} \quad (4)$$

La première ligne de ce critère caractérise le fait que les disparités sont faibles entre les deux vues, donc le barycentre de la région correspondante se localise à peu près au même endroit que dans la première vue.

La deuxième ligne est aussi liée aux faibles disparités mais représente le fait que d'une vue à l'autre la région n'aura pas beaucoup changé et même en cas d'occultation, la surface recouverte n'est pas très importante. Elle permet d'éliminer quelques ambiguïtés.

La troisième ligne, quant à elle, fait référence à l'hypothèse de surfaces lambertiennes. Les modélisations obtenues par plans d'intensité sont donc proches elles aussi.

Ce critère est très simple et peut bien sûr être adapté à différentes situations en pondérant chacun des termes du score de corrélation.

Nous calculons ces scores de corrélation pour chacune des mises en correspondances possibles. Puis nous validons la mise en correspondance du score le plus faible, retirons tous les scores qui faisaient intervenir une de ces deux régions, et recommençons en cherchant à nouveau le score le plus faible ... Nous arrêtons ce processus dès que le score minimum dépasse un seuil de tolérance donné, qui décrit quel taux de différences entre les deux vues est accepté. Cette méthode est en $O(N^2 log(N))$ si N est le nombre de régions, puisqu'elle dépend du nombre de corrélations possibles entre toutes les régions des deux images. Elle est donc relativement coûteuse mais permet d'éliminer un bon nombre d'ambiguïtés. Nous ne cherchons pas à trouver un correspondant pour chacune des régions de la première vue, car il est évident que de nouvelles régions apparaissent dans la deuxième vue même en cas de disparités faibles.

D'autre part la mise en correspondance des régions de petites tailles est souvent peu robuste et donc inintéressante. Nous ne calculons donc pas les scores de corrélation faisant intervenir une région jugée de taille insuffisante. Comme on peut assigner aux régions non corrélées une disparité moyenne ou minimum selon les cas obtenue à partir des corrélations voisines, cela paraît plus juste de rectifier par la suite la proximité des petites régions basées

sur des valeurs voisines, que de rectifier une proximité obtenue à partir d'une mauvaise mise en correspondance.

5.3. obtention d'une carte de proximités

Nous avons donc à présent une carte de disparités correspondant à la paire d'images considérée. Ce que nous voulons ensuite, c'est une carte de proximités. Il nous faut donc passer de l'une à l'autre à présent, et sans se limiter au cas de la géométrie standard, qui est trop limitatif en vision active.

Grâce aux paramètres de calibration, on obtient une fonction de mise en correspondance. Nous montrerons qu'à partir de cette fonction, on peut établir une relation entre la disparité et la profondeur.

Développons les équations liées à la mise en correspondance utilisée. Ces équations connues, nous permettrons de clarifier nos notations.

On se place dans le cas d'une paire d'images stéréo. La configuration des caméras est représentée dans la figure 5.3.

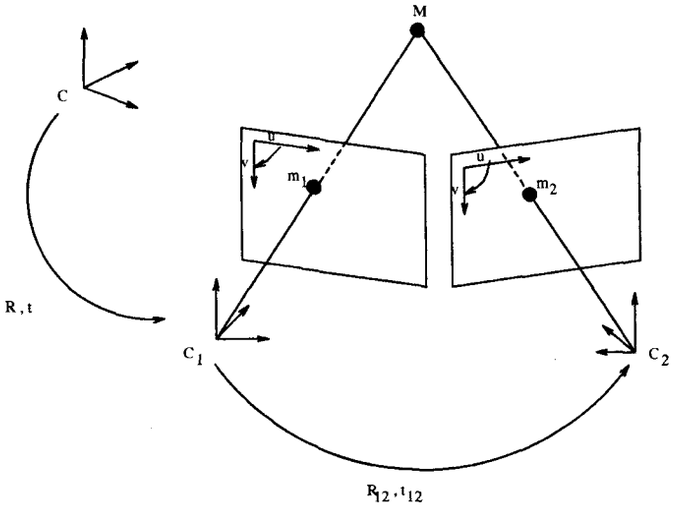

Figure 3. – Représentation d'une paire d'images quelconque.

Chaque caméra possède son repère propre lié à son centre optique respectivement C_1 et C_2, et les transformations pour passer du repère de la première caméra à la seconde sont décomposées ici en une rotation R_{12} et une translation t_{12}.

on a donc :

– dans le repère absolu :

$M = \begin{pmatrix} X & Y & Z \end{pmatrix}^T$ le point 3D

– dans les repères respectifs des caméras :

$M_1 = \begin{pmatrix} X_1 & Y_1 & Z_1 \end{pmatrix}^T$ et $M_2 = \begin{pmatrix} X_2 & Y_2 & Z_2 \end{pmatrix}^T$ qui représentent M

– Dans les repères liés aux plans images des caméras :

$m_1 = \begin{pmatrix} u_1 & v_1 & 1 \end{pmatrix}^T$ et $m_2 = \begin{pmatrix} u_2 & v_2 & 1 \end{pmatrix}^T$ qui sont les projections respectives du point M, en supposant la distance focale égale à 1.

Soit $\mathbf{A}_1$ et $\mathbf{A}_2$ les matrices des paramètres intrinsèques des caméras **1** et **2**. Ces matrices permettent de passer du repère de la caméra au repère de son plan image. Elles se présentent sous la forme suivante [8] :

$$A_i = \begin{pmatrix} \alpha_u^i & \gamma^i & u_0^i \\ 0 & \alpha_v^i & v_0^i \\ 0 & 0 & 1 \end{pmatrix}$$

Grâce à cette notation matricielle, nous avons donc les relations suivantes [8] :

$$\begin{aligned} Z_1 m_1 &= A_1 M_1 \\ Z_2 m_2 &= A_2 M_2 \end{aligned} \tag{5}$$

Nous connaissons aussi la transformation entre les deux repères caméras par la matrice 3x3 de rotation $\mathbf{R}_{12}$ et la matrice 1x3 de translation $\mathbf{t}_{12}$. Cette connaissance est liée à la calibration dans le cas de la stéréo et à l'odométrie dans le cas du mouvement sur un capteur actif. On a alors :

$$M_2 = R_{12} M_1 + t_{12} \tag{6}$$

En remplaçant $\mathbf{M}_2$ dans (5) par son expression donnée par (6) on a alors :

$$Z_2 m_2 = A_2 R_{12} A_1^{-1} Z_1 m_1 + A_2 t_{12} \tag{7}$$

On pose alors les notations suivantes :

$$Q = A_2 R_{12} A_1^{-1}$$

$\mathbf{Q}$ est l'homographie du plan à l'infini [31].

$$s = A_2 t_{12}$$

$\mathbf{s}$ est le foyer d'expansion ou épipole [31].

La relation (7) devient alors :

$$Z_2 m_2 = Q Z_1 m_1 + s \tag{8}$$

En faisant la remarque que :

$$z^T m = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = 1$$

et en utilisant la définition de la proximité $\Pi_1 = 1/Z_1$, c'est-à-dire $\Pi = 0$ pour la profondeur à l'infini on arrive alors, à partir de l'équation (7), à la formule suivante :

$$Z_2 m_2 = Z_1(Q + s\Pi_1 z^T)m_1 \quad (9)$$

On remarque que l'usage de la proximité permet de manipuler des quantités plus "linéaires" que si nous avions pris la valeur de la profondeur. De plus, par défaut, les points d'un système visuels sont vus à l'infini, i.e. la composante translationelle du mouvement qu'ils induisent est négligeable par rapport à la composante rotationelle [30]. Cela correspond à une disparité et proximité nulles et une profondeur infinie. Il est clair que l'usage de la proximité s'impose en ce cas.

Nous avons donc maintenant une relation simple qui permet à partir des paramètres de calibration et éventuellement de l'odométrie du système, de nous donner les couples de points en correspondance dans les deux images. On a alors facilement les coordonnées de m_2 en fonction de celles de m_1 par :

$$u_2 = \frac{Q_{00}u_1 + Q_{01}v_1 + Q_{02} + s_0\Pi_1}{Q_{20}u_1 + Q_{21}v_1 + Q_{22} + s_2\Pi_1} \quad (10)$$

$$v_2 = \frac{Q_{10}u_1 + Q_{11}v_1 + Q_{12} + s_1\Pi_1}{Q_{20}u_1 + Q_{21}v_1 + Q_{22} + s_2\Pi_1} \quad (11)$$

Et pour la troisième équation,
nous avons (avec $\Pi_2 = 1/Z_2$) :

$$\Pi_2 = Q_{20}u_1 + Q_{21}v_1 + Q_{22} + s_2\Pi_1 \quad (12)$$

Cette équation que nous n'utilisons pas ici, permet de fusionner différentes cartes de proximités au sein d'une séquence d'images, comme réalisé par ailleurs [35].

Pour obtenir la carte de proximités, nous allons nous servir des équations (10) et (11). A partir de ces équations et en notant la ligne i de la matrice Q par Q_i, nous avons deux relations nous permettant de remonter à la proximité :

$$\Pi_1 \underbrace{[s_0 - u_2 s_2]}_{d_0} = \underbrace{u_2(Q_2 m_1) - (Q_0 m_1)}_{n_0} \quad (13)$$

$$\Pi_1 \underbrace{[s_1 - u_2 s_2]}_{d_1} = \underbrace{u_2(Q_2 m_1) - (Q_1 m_1)}_{n_1} \quad (14)$$

Nous avons donc deux équations pour une seule inconnue. Or nous ne savons pas quelle est la meilleure de ces estimations.

Tout d'abord, on sait que la proximité est toujours positive.

On doit donc avoir $\frac{n_0}{d_0} > 0$ et $\frac{n_1}{d_1} > 0$.

Si l'une de ces deux relations n'est pas vérifiée, alors il faut choisir l'autre. Si aucune des deux n'est vérifiée, c'est que la mise en correspondance est mauvaise, on fixe alors la proximité à une valeur par défaut comme pour les petites régions.

Si par contre, ces deux relations sont vérifiées, on minimise au moindre carré la norme pondérée de la disparité. On veut donc minimiser le critère suivant :

$$L(\Pi_1) = [n_0 - \Pi_1 d_0]^2 + [n_1 - \Pi_1 d_1]^2$$

Ce qui revient, après dérivation par rapport à Π, à :

$$\Pi_1 = \frac{n_0 d_0 + n_1 d_1}{d_0^2 + d_1^2}$$

Nous avons à présent une carte de proximités que l'on peut affiner grâce à la régularisation.

6. affinage et régularisation de la carte de profondeurs

Nous avons préféré travailler sur la proximité plutôt que sur la profondeur pour des raisons de stabilité numérique. C'est donc sur ce champ que va se porter la régularisation. Une raison de ce choix concerne, par exemple, l'initialisation de la carte de proximités en cas d'utilisation de ce module sans passer par la segmentation région. Il paraît alors plus simple de considérer les objets loin dans la scène, voire à l'infini. Or, cette notion est difficile à quantifier en termes de profondeur. En revanche, elle correspond à une proximité nulle puisque la proximité est l'inverse de la profondeur $\mathbf{Z}$.

Pour représenter notre système, nous avons choisi une forme simple pour l'énergie. Le terme de vraisemblance ne porte que sur l'intensité en chaque point et nous avons pris une contrainte du premier ordre sur le modèle de la carte de disparités. Nous obtenons donc une énergie de la forme suivante à minimiser :

$$\begin{aligned} \Xi(\Pi_1) &= \int_u \int_v (\lambda \parallel I_1(m_1) - I_2(m_2) \parallel^2 \\ &+ \parallel \nabla\Pi_1 \parallel^2) du dv \end{aligned} \quad (15)$$

où λ pondère l'influence du terme de vraisemblance par rapport à la contrainte. Physiquement λ permet de contrôler le facteur de lissage : pour λ petit, la proximité est très lisse et pour λ grand, seule la mesure locale est prise en compte. Ce choix du critère "le plus simple possible" est lié à un soucis de traitement rapide et le fait que nous pouvons escompter de bons résultats initiaux grâce à notre initialisation.

Hormis pour le calcul du terme de vraisemblance, il nous faut les coordonnées de m_2 qui est le correspondant de m_1 dans la première image. Notons F la fonction de mise en correspondance précédemment décrite et explicitée dans (10) et (11). Cette fonction F dépend alors des coordonnées du point m_1 et de la proximité associée :

$$m_2 = F(m_1, \Pi_1)$$

L'énergie donnée par la relation (15) devient alors :

$$\begin{aligned} \Xi(\Pi_1) &= \int_u \int_v (\lambda \parallel I_1(m_1) - I_2(F(m_1,\Pi_1)) \parallel^2 \\ &+ \parallel \nabla\Pi_1 \parallel^2) du dv \end{aligned} \quad (16)$$

Comme nous travaillons sur des images, nous utilisons la forme discrétisée de cette énergie. Dans ce cas là, la minimisation sur toute l'image revient à minimiser l'énergie suivante :

$$\begin{aligned} \tilde{\Xi}(\Pi_1) &= \textstyle\sum_{m_1} \lambda [I_1(m_1) - I_2(F(m_1,\Pi_1))]^2 \\ &+ \parallel \nabla\Pi_1(m_1)^2 \parallel \end{aligned} \quad (17)$$

Ce que nous voulons trouver en minimisant cette énergie, c'est la proximité Π_1, il nous faut donc dériver cette énergie par rapport à Π_1, ce qui nous donne l'équation normale suivante en chacun des pixels :

$$0 = 2[\lambda(I_2(F(m_1,\Pi_1)) - I_1(m_1)) \frac{\partial I_2}{\partial m_2} \frac{\partial m_2}{\partial \Pi_1} + \Delta\Pi_1] \quad (18)$$

puisque la dérivée de la norme au carré du gradient ($\nabla\Pi_1$) est le laplacien ($\Delta\Pi_1$).

Or, dans le cas discret nous pouvons prendre l'estimée suivante pour le Laplacien [26] :

$$\Delta\Pi_1 = 4(\overline{\Pi_1} - \Pi_1)$$

où $\overline{\Pi_1}$ est la moyenne de la proximité sur les huit voisins les plus proches de m_1. Plus précisément, c'est une combinaison de l'opérateur de moyennage sur les quatre plus proches voisins avec ce même masque obtenu pour des axes diagonaux :

$$\frac{1}{4}\Delta\Pi_1 = \frac{4}{5}\begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & -1 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 \end{bmatrix} + \frac{1}{5}\begin{bmatrix} \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & -1 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} \end{bmatrix} = \frac{1}{20}\begin{bmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{bmatrix}$$

Cette équation est vérifiée si et seulement si :

$$\Pi_1 = \frac{\lambda}{4}[I_2(F(m_1,\Pi_1)) - I_1(m_1)] \frac{\partial I_2}{\partial m_2} \frac{\partial m_2}{\partial \Pi_1} + \overline{\Pi_1} \quad (19)$$

qui fournit une équation de récurrence sur Π [26], dont la convergence est assurée [6]. C'est à l'aide de cette relation que l'on pourra raffiner l'estimée de la carte de proximités de manière itérative en l'appliquant en chacun des pixels de l'image. Entre l'itération t et $t+1$ la régularisation s'exprimera par :

$$\Pi_1^{t+1} = \frac{\lambda}{4}[I_2(F(m_1,\Pi_1^t)) - I_1(m_1)] \frac{\partial I_2}{\partial m_2} \frac{\partial m_2}{\partial \Pi_1} + \overline{\Pi_1^t} \quad (20)$$

Nous avons alors deux méthodes pour implémenter ce processus itératif :

– la méthode de Jacobi où l'on calcule la régularisation à partir de la carte de proximités obtenue à l'étape précédente, la remise à jour étant effectuée une fois tous ces calculs faits. Cette technique se prête bien à la parallélisation.

– la méthode de Gauss Seidel qui remet à jour la carte de proximités dès que la régularisation est calculée. Cette méthode permet de ne pas avoir à garder en mémoire la correction en chacun des pixels avant de remettre à jour. Si cette méthode n'est pas facilement parallélisable, elle converge cependant plus rapidement que la méthode de Jacobi. C'est cette méthode que nous avons choisie dans notre implémentation.

Reprenons à présent la relation de régularisation (20). Nous voulons aussi bien pouvoir utiliser cette régularisation seule pour estimer la carte de proximités, que pouvoir raffiner celle obtenue à partir de la segmentation en régions. Cela revient à accorder plus d'importance à l'estimation des proximités de la carte, que l'on a dès le début de la régularisation. Nous avons alors rajouté un terme, afin de tenir compte de cette connaissance en complémentarité avec le moyennage. La relation (20) va alors être de la forme :

$$\begin{aligned} \Pi_1^{t+1} = \tfrac{\lambda}{4}[I_2(F(m_1,\Pi_1^t)) - I_1(m_1)] \frac{\partial I_2}{\partial m_2} \frac{\partial m_2}{\partial \Pi_1} \\ +(1-\alpha)\overline{\Pi_1^t} + \alpha\Pi_1^t \end{aligned} \quad (21)$$

où α est un terme de pondération, qui correspond à une méthode classique [6] de "sur-relaxation".

λ permet de contrôler le lissage et α permet de contrôler la convergence de l'algorithme. Pour α proche de 1, le système se base sur la valeur précédente, ce qui freine la convergence, mais améliore la stabilité du système (convexification). Pour α proche de 0, on privilégie le voisinage précédent. On se rapproche alors de la méthode sans pondération de l'équation (20).

7. résultats

Voici tout d'abord les résultats obtenus en appliquant uniquement la régularisation à une paire stéréo d'images synthétiques d'une pyramide.

On donne alors comme initialisation de la carte de proximités, un plan fronto-parallèle [26].

Malgré la simplicité du critère et grâce à un mode à plusieurs résolutions, comme on peut trouver dans [26], la convergence

est assez rapide : on applique pour cinq ou six taux de lissage différents de deux à cinq étapes de régularisation. Les temps d'exécution sont de l'ordre de 8 secondes CPU pour une itération de la régularisation de la carte de proximités.

On voit que l'on arrive à récupérer un profil sensé de la pyramide, où les discontinuités de profondeur ont été prises en compte lors du processus de régularisation au lieu d'être atténués.

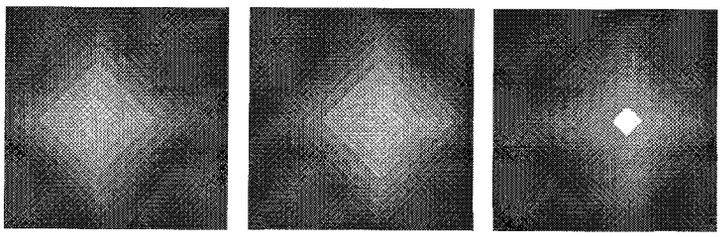

Figure 4. – A gauche : une paire stéréoscopique d'images synthétiques 512×512 d'une pyramide. A droite : une des deux vues obtenues après le plus fort lissage nécessaire pour le mode multi-échelle de la régularisation.

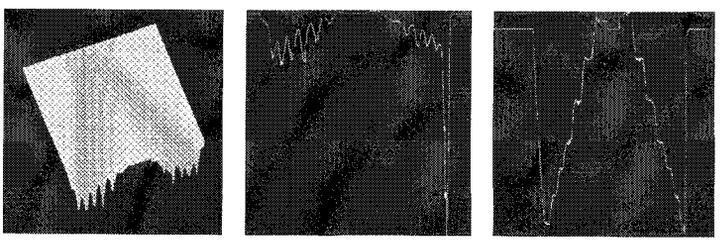

Figure 5. – De gauche à droite on peut voir : la représentation 3D de la première étape de convergence, puis une coupe de cette même carte de proximités et une autre coupe après quelques itérations de la régularisation au même niveau de lissage.

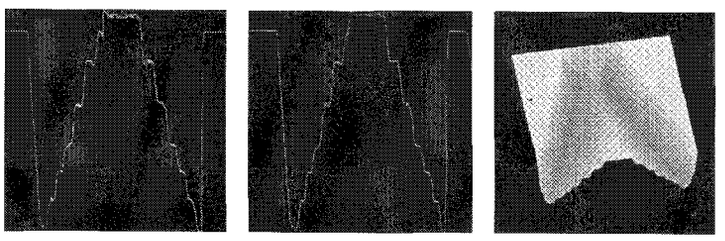

Figure 6. – Dans les deux images de gauche, on voit tout d'abord l'évolution d'une coupe de la carte de proximités, au cours de la régularisation couplée avec un mode à différentes résolution. L'image de droite représente la carte de proximités 3D de la pyramide obtenue en fin de traitement.

Nous avons voulu par cette deuxième expérience, montrer la qualité de l'initialisation de la carte de proximités dans le cas d'une paire d'images très texturées. Nous remercions tout particulièrement Michael Otte de l'université de Karlsruhe pour nous avoir fourni cette séquence.

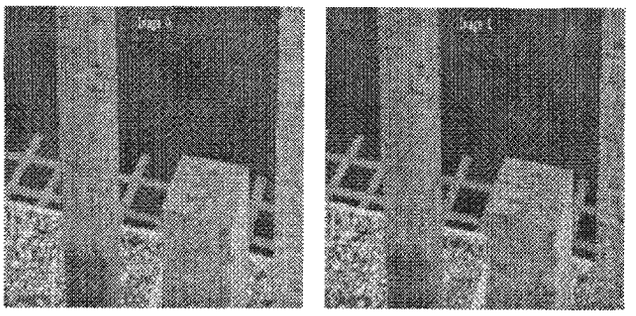

Figure 7. – Paire d'images d'une scène d'intérieur comportant différents piliers. On peut remarquer la très faible disparité entre ces deux vues.

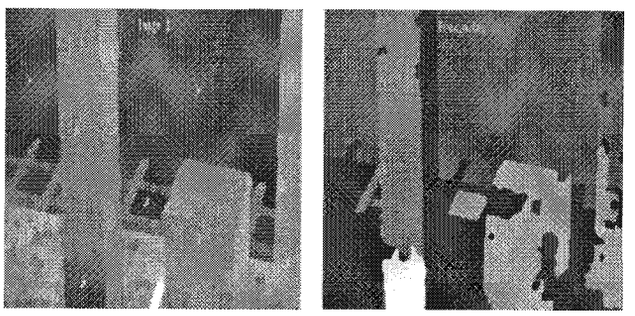

Figure 8. – L'image de gauche représente les déplacements trouvés pour les régions mises en correspondance. L'image de droite est l'initialisation correspondante de la carte de proximités. Plus les régions sont claires, plus elles sont proches.

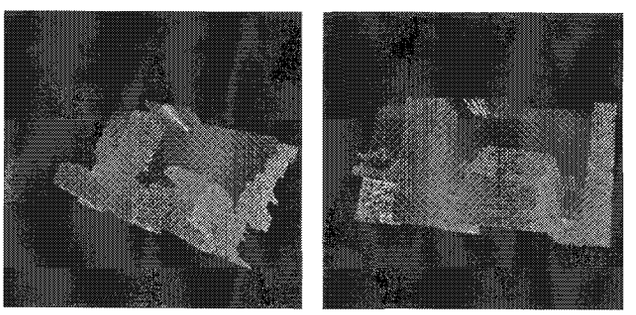

Figure 9. – Deux images de reconstruction 3D initiales de la scène pour deux valeurs de seuils sur le score de corrélations différentes.

Ici, nous avons travaillé sur deux images acquises à l'aide de la caméra principale de la tête artificielle. Ces images sont rectangulaires de taille 440×217. On peut remarquer sur les images de départ que le bruit à l'acquisition est important. De plus, on a un cas de réflectance important au niveau de l'écran du moniteur. On pourra suivre son comportement en sachant que l'hypothèse de base est que nous considérons des surfaces lambertiennes.

Ces images ont été segmentées après l'ajustement automatique du paramètre de lissage. Ensuite, par la mise en correspondance des régions détectées, on initialise la carte de disparités nécessaire au processus de régularisation. Les temps de calcul sont de l'ordre de 3.5 secondes CPU par image pour la segmentation en régions,

de 5.3 secondes CPU par image pour récupérer les informations correspondant à chacune des régions et la modélisation par plan. La mise en correspondance des régions est de l'ordre de 3 secondes CPU et la régularisation se fait à peu près en 2.9 secondes CPU par étape.

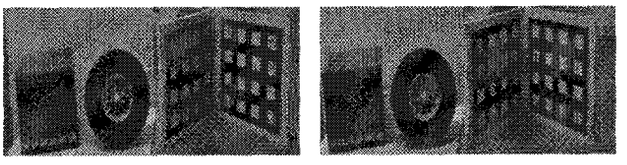

Figure 10. – Images de départ acquises sur la tête.

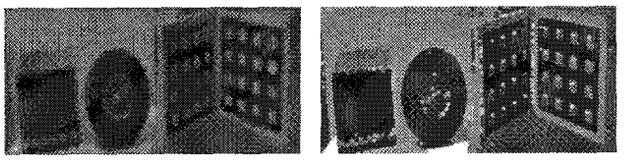

Figure 11. – De gauche à droite : image après ajustement du taux de lissage, image segmentée en régions par l'algorithme"toboggan". Chaque carré marque la présence d'une région.

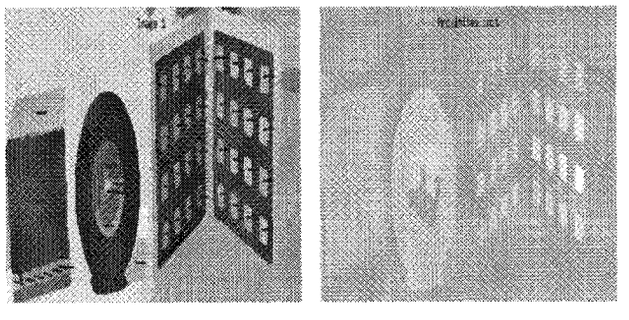

Figure 12. – L'image de gauche représente les déplacements trouvés pour chacune des régions corrélées. On peut remarquer qu'un bon ajustement du seuil de tolérance sur le score de corrélation et sur la taille des régions à mettre en correspondance, permet d'obtenir un champ cohérent de déplacement entre les deux vues. L'image de droite est l'initialisation de la carte de proximités. Plus les régions sont claires, plus elles sont proches.

8. conclusion

Au niveau d'implémentation actuel, notre problème majeur vient des caractéristiques de la carte de proximités que nous donnons en entrée de la régularisation. Par la segmentation en régions, on remarque que la carte obtenue est constituée de morceaux de plans fronto-parallèles. Cela veut dire qu'aux bords de ces bouts de plans, nous avons des discontinuités importantes. Or, le modèle de carte que nous avons pour le moment dans le processus de régularisation, est un modèle continu au lieu de continu par

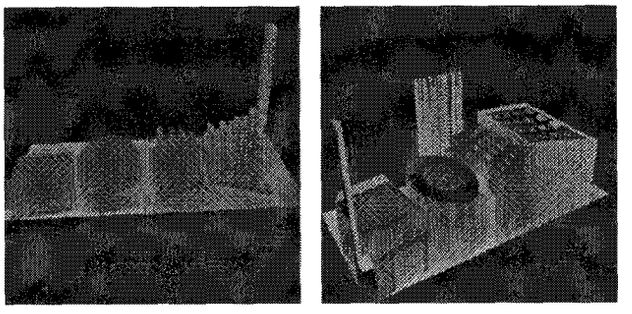

Figure 13. – Ces deux images sont des représentations 3D des cartes de proximités initiales, obtenues pour deux valeurs de seuil sur le score de corrélation. Dans l'image de droite, on voit que ce seuil est plus élevé. En effet, de nouvelles régions de la grille de calibration ont été mises en correspondance, au risque de valider des corrélations erronées de régions de petite taille. On peut remarquer que l'on a retrouvé la disposition relative des objets dans la scène, comme par exemple la roue qui est effectivement sensiblement plus proche que le moniteur.

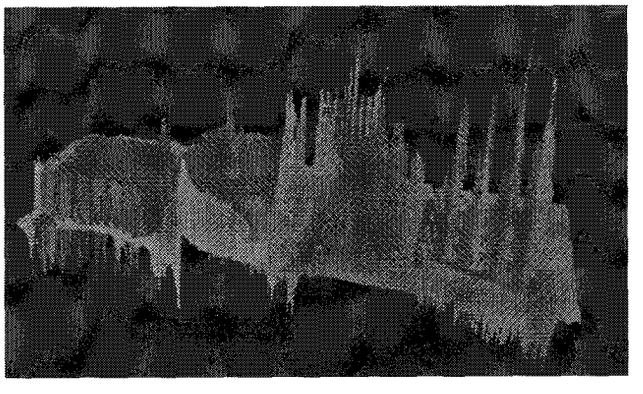

Figure 14. – Vue 3D de la carte de disparité après quelques étapes de régularisation. Les problèmes que l'on voit apparaître aux bords des régions lors de la régularisation, proviennent du fait que l'étape de segmentation impose des discontinuités franches. Actuellement nous ne gérons pas suffisamment des discontinuités. De nombreuses améliorations possibles sont exposées dans la section suivante.

morceaux. Ce phénomène limite pour l'instant la qualité des résultats obtenus.

Pour pallier ce problème, comme chaque pixel possède des informations sur la région à laquelle il appartient, nous pouvons expliciter ces discontinuités et en tenir compte, comme suggéré par [26]. Par exemple, le terme de moyennage peut être limité aux voisins du point considéré qui appartiennent eux aussi à la même région, ce qui va permettre de réaliser un mécanisme tenant compte des discontinuités et des occultations. Cela permettra aussi de conserver les discontinuités franches détectées en sortie de la segmentation en régions.

D'autre part, si nous avons choisi de partir sur une représentation en régions de l'image, c'est parce qu'elle nous paraît beaucoup plus appropriée à la détection et au traitement des occultations et discontinuités de proximité ou de profondeur, qui sont, comme nous l'avons déjà fait remarquer, les deux principaux problèmes à gérer lors d'une reconstruction dense 3D.

En effet, les frontières des régions sont les parties de l'image où peuvent se localiser les discontinuités. On pourra alors, par les différentes méthodes présentées, les localiser et ainsi adapter la régularisation de façon très simple, à l'aide de fonctions de visibilité comme on peut le trouver dans [34].

La représentation en régions est aussi adaptée au problème des occultations. En effet, parmi les frontières des régions détectées, il y a bien évidemment les frontières solides des objets dans la scène. Les occultations provenant du fait qu'un objet recouvre partiellement un autre, elles seront elles aussi localisées aux bords des régions. Là aussi la régularisation pourra alors être adaptée.

Par contre, s'il faut accorder beaucoup plus d'importance à l'estimation de la proximité obtenue en sortie de la segmentation en régions, elle doit alors être la plus robuste et fiable possible. De fait, plus nous y accorderons de l'importance, plus elle devra être précise et détaillée. Il faudra surtout plus de renseignements sur leurs connexités, des modèles cinématiques pour prédire des mouvements importants, des modèles quadratiques sur l'intensité pour remonter à la normale de la surface, toute chose aisément implémentable dans le code actuel. Par exemple, pour le moment, les régions considérées comme trop petites ne sont pas corrélées. Ceci entraîne évidemment des "trous" dans la carte de proximités estimée. Nous pourrions alors envisager grâce à ces nouvelles informations, une étape de rattachement de ces petites régions de l'image aux plus importantes. Ainsi, la carte serait plus complète et plus juste au sens de la régularisation.

D'autre part, il faudrait aussi intégrer le fait que si certaines régions ne sont pas corrélées, la proximité minimale ou moyenne assignée est peu significative. Dans ce cas là il faudrait pondérer les termes de la régularisation afin de moins prendre en compte cette information initiale pour ces régions spécifiques. Elles seraient ainsi corrigées plus fortement que les régions mises en correspondance.

Nous avons donc pu réaliser par la coopération de deux modules de vision, un mécanisme de reconstruction dense d'une scène, qui est à la fois rapide et efficace, et dont les paramètres les plus sensibles s'adaptent automatiquement à la scène observée, ce qui est tout à fait nécessaire en vision active.

La maquette logicielle utilisée a été volontairement réduite puisqu'il fallait montrer que l'algorithme "marche simplement". Mais de multiples améliorations peu coûteuses peuvent être envisagées. La nécessité de ces améliorations dépendant de l'application considérée, leur description sort du cadre de cette étude.

De plus, même à ce stade relativement restreint de sophistication, le présent algorithme a pu être utilisé au sein d'un système de détection et d'observation de cibles 3D en mouvement comme rapporté en [30].

Remerciements

Nous remercions Olivier Faugeras. Nous remercions aussi Gérard Giraudon et Luc Robert pour leur aide et leurs conseils lors de l'élaboration de ce travail.

BIBLIOGRAPHIE

[1] T. Aach and A. Kaup, M.A.P. estimation of dense disparity-fields for stereoscopic images. In *international Conference on Image Procesing, Singapore*, pages 1113-117, September 1992.

[2] A. L. Abbott and N. Ahuja, Surface reconstruction by dynamic integration of focus, camera vergence, and stereo. In *International Conference on Computer Vision*, pages 532-543, Tampa, FL, Decembre 1988.

[3] A. L. Abbott and N. Ahuja, Active Surface reconstruction by integrating focus, vergence stereo and camera calibration. In *Proceeding of the 3 rd ICCV, Osqka*, pages 489-492, 1990.

[4] N. Ayache, *Artificial Vision for Mobile Robots.* MIT Press, Cambridge, Massachusetts, 1989.

[5] L. Cohen, L. Vinet, P. Sander, and A. Gagalowicz, Hierarchical region based stereo matching. In *CVPR' 89 San-Diego*, pages 416-421, 1987.

[6] P. Duchateau et D. W. Zachmann, *Theory and problem of Partial Differential Equations.* Schaum's oytline series, Mc Graw-Hill Book Compagny, 1986.

[7] J. Fairfield, Toboggan contrast enhancement. in *Application of Artiricial Intellingence, Machine Vision and Robotics*, volume 1708, pages 221-229. Proceedings of S.P.I.E., 1980.

[8] O.D. Faugeras, *Three-dimensional Computer Vision : a geometric viewpoint.* MIT Press, Boston, 1993.

[9] O.D. Faugeras, B. Hotz, H. Mathieu, T. Viéville, Z. Zhang, P. Fua, E. Théron, L. Moll, G. Berry, J. Vuillemin, P. Bertin, and C. Proy, Real time correlation-based stereo : algorithm, implementations and applications. Technical report 2013, INRIA, 1993.

[10] W. Foerstner and A. Pertle, Photogrammetric standard methods and digital images matching techniques for high precision measurements. *Patterm recognition in pratice*, pages 55-72, 1986.

[11] P. Fua, A parallel stereo algorithm that produces dense depth maps and preserves images features. Technical Report 1369, Institut National de recharche en Informatique et en Automatique, January 1991.

[12] D. Geman and G. reynolds, Constrained restoration and the recovery of discontinuities. *Transactions on Pattern Analysis and Machine Intelligence*, 14(3), March 1992.

[13] W. Hoff and N. Ahuja, Extracting surfaces from stereo images : An integrated approach. in *1st ICCV, London*, pages 284-294, 1987.

[14] B K P Horn, *Robot Vision.* MIT Press, Cambrodge, Massachusetts, 1986.

[15] B K P Horn, Height and gradient from shading. *International journal of Computer Vision*, 5 : 1 : 37- 75, 1990.

[16] J.J. Hwang and E.L. Hall. Matching of featured objects using relational tables form stereo images. *Computer Graphics and Image Processing, 20 : 22-42, 1982.*

[17] S. Mitter, J. Marroquin and T. Poggio, probabilistic solution of ill-posed problems in computation vision. *Journal of the American Statistical Association*, 82(397) : 76-89, March 1987.

[18] T. Kanade and M. Okutomi, A stereo matching algorithm with an adaptative window : Theory and experiment. Artificial intelligence CMU-CS-90-120, Carnegie Mellon, April 1990.

[19] R. Deriche , L. Robert and O. D. Faugeras, Dense Depth Map Reconstruction Using Multiscale Regularization. In *2 nd Singapore International Conference on Image Processing*, pages 13-127, Singapore , September 1992.

[20] H. Maître and W. Luo, Using models to improve stereo reconstruction. *IEEE trans. on Pattern Analysis and Machine Intelligence, 14 : 269-277, 1992.*

[21] R. March, A regularization model for stereo vision with controlled continuity. *Pattern Recognution Letters*, pages 259-263, october 1989.

[22] J.L. Marroquin, Surface reconstruction preserving discontinuities. Artificial Intelligence A. I. MEMO 792, Massachusetts Institute of Technology, August 1984.

[23] K. Pahlavan, J.O. Ekhlund, and T. Uhlin, Dynamic fixation. In 4th ICCV, Berlin, pages 404-411. IEEE Society, 1993.

[24] G. Randall, *parallélisation d'un algorithme de stéréoscopie trinoculaire.* PhD thesis, University of Orsay, Dept of Comp. Science, 1991. PhD thesis.

[25] Y. Remion, H. Maître, and J.L. Krahe, Recalage pare zones de 2 vue stéréoscopiques d'un univers composé d'objets plans. In *Cognitiva'87, La Villette, Paris*, pages 175-180, 1987.

[26] L. Robert, *Stéréovision : de la mise en correspondence de courbes à l'analyse photogrammatrique de la scène*. PhD thesis, Ecole polytechnique, Dept of Comp. Science, 1993. PhD thesis.

[27] P.S. Toh and A.K. Forrest, Occlusion detection in early vision. in *Third International Conference on Computer Vision, Osaka,* pages 126-132, December 1990.

[28] R. Vaillant and O.D. Faugeras, Using extremal bounderies for 3d object modeling. In *IEEE transaction on >pattern Analysis and Machine Intelligence,* volume 14, pages 157-173, February 1992.

[29] A. Verri and T. Poggio, Motion field and optical flow : differences and qualitative properties. Technical report AI Memo 917, MIT Press, Cambridge, 1986

[30] T. Viéville, E. Clergue, R. Enciso, and H. Mathieu, Experimentating with 3d vision on a robotic head. *Robotics and Autonomous Systems*, 14(1), 1995.

[31] T. Viéville and Q.T. Luong, Computing motion and structure in image sequences without calibration. In *The 12th Int. Cof. on Pattern Recognition, Jerusalem*, pages 420-426, 1994.

[32] A. Witkin, D. Terzopoulos, and M. Kass, Signal matching trough scale space. *International journal of Computer Vision*, pages 133-144, 1987.

[33] B. Wrobel-Dautcourt, Surfaces tridimentionnelles obtenues par appariement stéréoscopique de régions. in *Sixième AFCET en reconnaissance des formes et intelligence artificielle, Paris*, pages 299-307, 1987.

[34] N. Yokoya, Stereo surface reconstruction by multiscale-multistage regularization. Technical report 90-45, Electrotechnical Laboratory, November 1990.

[35] Z. Zhang, R. Deriche, Q-T. Luong, and O. Faugeras, A robust approach to image matching : Recovrey of the epipolar geometry. In *Proc. International Symposium of Young Investigators on Information/Computer/Control,* pages 7-28, Beijing, China, February 1994.

Manuscrit reçu le 24 Mars 1994.

LES AUTEURS

Emmanuelle CLERGUE

Emmanuelle Clergue est titulaire du D.E.S.S. de l'Ecole Supérieure des Sciences de l'Ingénieur, de Sophia-Antipolis et a obtenu le D.E.A. de Vision Robotique de l'Université de Nice. Elle achève une Thèse d'Université à l'Institut Eurecom, au sein du groupe Multimedia pour des travaux en Vision Artificielle. Elle a été stagiaire de l'INRIA en tant qu'élève ingénieur et stagiaire de DEA où elle a réalisé deux études scientifiques publiées dans des revues internationales.

Thierry VIÉVILLE

Thierry Viéville est un chercheur de l'Institut National de Recherche en Informatique et Automatique (INRIA) où il travaille au sein du projet de Vision Artificielle du Pr. Olivier Faugeras. C'est un Ingénieur Bio-Médical de l'Ecole Nationale Supérieure des Télécommunications, Maître en Mathématiques Fondamentales, Docteur en Neuro-Science, et Habilité à diriger des Recherches en Sciences de l'Ingénieur. Ces travaux concernent les applications des mathématiques à la vision artificielle, l'analyse du mouvement, la vision active ainsi que les liens entre vision artficielle et vision biologique.