Les images numériques prennent du volume pour copier fidèlement la réalité

LA VISION INFORMATIQUE DU RELIEF

RENAUD KERIVEN

RENAUD KERIVEN est Ingénieur des Ponts et Chaussées, au centre d'enseignement et de recherche en mathématiques, informatique et calcul scientifique de l'ENPC.

Nos deux yeux nous fournissent chacun une image plane, et pourtant le cerveau reconstruit en permanence un monde en trois dimensions. Comment transmettre cette faculté à l'ordinateur ? Des techniques intuitives ne donnent pas de résultats convaincants. La réponse est à chercher du côté d'un des domaines mathématiques les plus actifs aujourd'hui.

La photographie est certainement le moyen le plus répandu de partager un souvenir, de faire connaître un endroit. Et demain ? La vidéo, bien sûr ! Envoyer une cassette vidéo par la poste n'est pas aujourd'hui chose courante, encore moins l'équivalent numérique qui consisterait à le faire par Internet : dans les deux cas, c'est encore trop encombrant, peu pratique. Mais demain, ce sera évidemment quotidien. Un petit bonjour ? Un souvenir de vacances ? Un simple clic de souris, et le tout sera expédié par le WEB, son et animation compris…

Troisième dimension. Et pourtant, ce ne sera pas encore une représentation complète de la réalité que recevra le destinataire. Réfléchissez. A l'image fixe de la photographie, la vidéo apporte une dimension supplémentaire, le temps, mais il en manque encore une : la troisième dimension en espace. Tout cela reste désespérément plan. Entendons-nous bien : il existe aujourd'hui plusieurs moyens de donner une sensation de relief, que ce soit en observant la scène d'un point imposé fixé d'avance (cinéma en relief) ou en s'y déplaçant (procédés de réalité virtuelle). Nous ne disposons pas encore de véritables projections tridimensionnelles, sortes d'« hologrammes animés » dont rêvaient les auteurs de science-fiction, mais c'est déjà un premier pas. Non, le vrai problème, celui auquel nous allons nous intéresser ici, c'est l'acquisition de la forme tridimensionnelle du sujet considéré. Lorsque tel réalisateur filme ses comédiens avec deux caméras, il en enregistre deux séries d'images simultanées prises depuis deux emplacements légèrement distincts. Ces séries d'images, une fois reprojetées et observées avec un appareillage adapté (lunettes polarisantes ou autres), reproduiront la sensation de relief. C'est, en bien plus sophistiqué, un procédé de stéréoscopie vieux comme la photographie. Qu'il ne prenne pas l'envie au spectateur de voir ce qui se passe de l'autre côté de la scène qui lui est proposée : il ne voit et ne peut voir que le relief qu'il aurait observé, si ses deux yeux avaient été à la place des deux caméras.

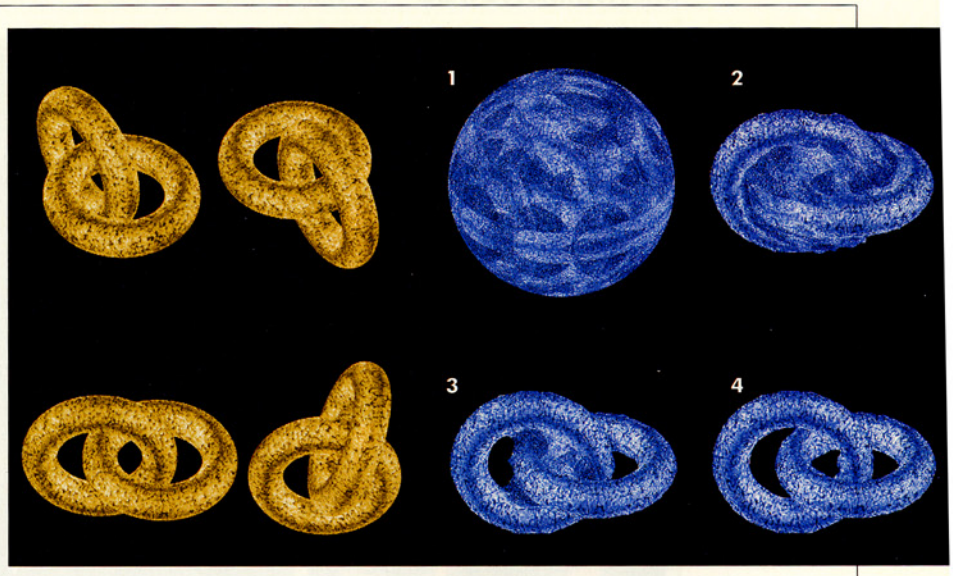
Un véritable enregistrement tridimensionnel, ce n'est pas saisir deux images d'un même objet pris sous des angles différents. C'est mémoriser l'information tridimensionnelle en elle-même —« il y a un objet à tel endroit de l'espace, il a telle forme, telle couleur, etc. » —, c'est dresser une carte tridimensionnelle complète de ce qui est observé.

Acquérir cette carte, voilà le difficile problème auquel devra se confronter le cinéma du futur. Pourtant, notre cerveau résout ce problème en permanence, justement à partir des images planes qui lui sont transmises : nos deux yeux transmettent deux images différentes du même …

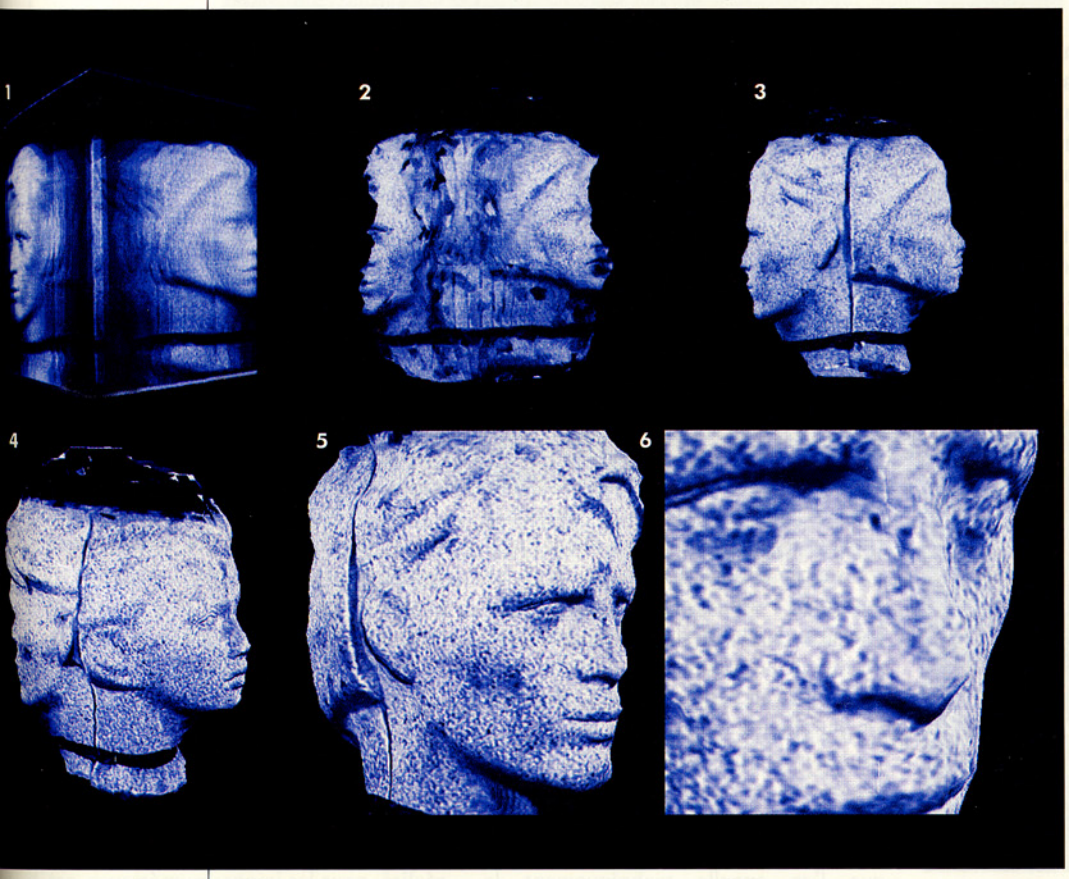
En jaune, plusieurs images planes d'un double buste en polystyrène. Ces photos nous suffisent pour percevoir immédiatement la forme de l'objet. Muni d'un algorithme développé par l'INRIA et l'ENPC, l'ordinateur y parvient aussi, *après quelques étapes intermédiaires (en bleu) : à partir d'une première* supposition assez vague, il déforme la surface, affine progressivement les traits et parvient à retrouver les contours des bustes. Bien sûr, on ne peut voir ici qu'une image plane de ce que la machine a obtenu, mais ce sont de véritables représentations en trois dimensions que la machine a reconstituées.

environnement, images que le cerveau exploite pour extraire de l'information sur la forme et la position de ce que nous voyons. Ce que le cerveau arrive à faire avec nos yeux, il est tentant d'essayer de le réaliser avec un système artificiel fondé sur des caméras reliées à un ordinateur. Imaginez cela : vous venez de filmer votre petit dernier effectuant ses premiers pas. Sitôt connecté au Caméscope, votre PC en extrait les images et commence à reconstituer le tout en trois dimensions. Quelques instants de patience, une petite connexion Internet, et vous pourrez faire admirer en 3D à ses grands-parents les exploits du jeune marcheur.

Retrouver la position d'un point dans l'espace lorsqu'on en possède deux images différentes… A la fois très simple dans le principe et très compliqué dans la pratique. Pourquoi ? Regardez d'un seul œil ce point M. Vous ne pouvez pas dire où se trouve M exactement, mais vous en savez déjà beaucoup : vous savez dans quelle direction il est. Mais à quelle distance ? Il vous faut ouvrir votre deuxième œil pour obtenir une deuxième direction.



Ces anneaux imbriqués sont un piège pour l'ordinateur : peut-il reconnaître deux objets physiquement séparés ? S'il se contentait de déformer une surface, il échouerait ici. Pourtant, on le voit partir d'une supposition inexacte, la sphère, qu'il parvient à couper pour obtenir les deux anneaux.
En fait, l'ordinateur se place pour son calcul non pas dans l'espace, mais dans un hyperespace de dimension quatre… non représenté ici, et pour cause ! Dans cet hyperespace, les deux anneaux appartiennent à une seule et même hypersurface, et les séparer ne pose pas plus de problèmes que de reconnaître un seul objet.

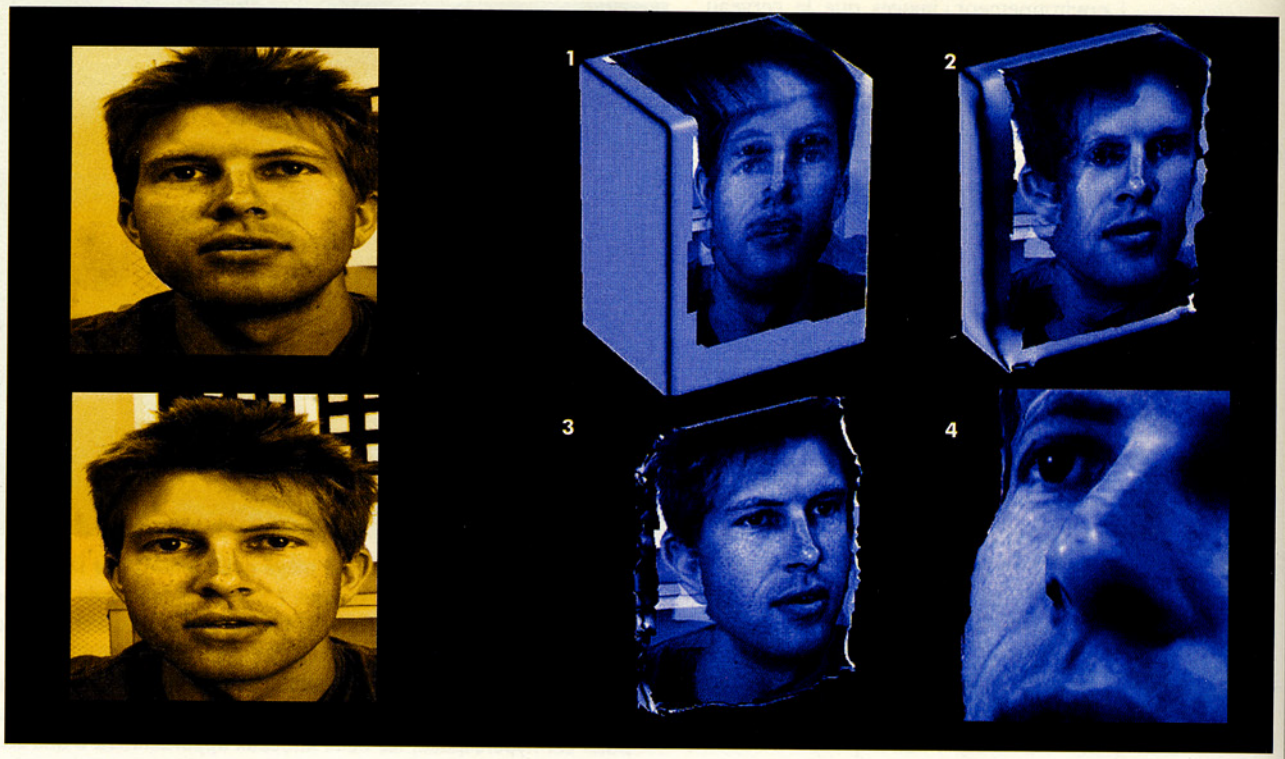


Vous saurez ainsi immédiatement que M se situe à l'intersection des deux droites (deux droites dans l'espace ne se coupent pas toujours, mais ici, oui). Procédez de la même manière pour tous les points d'un objet observé et vous aurez sa position et sa forme dans l'espace.

Voilà pour le principe. Dans la pratique, la réalisation informatique du procédé soulève un problème. Un seul, mais qui serait suffisant pour réduire tous les espoirs à néant : pour un point m1 donné dans la première image, la machine ne sait pas quel point m2 de la deuxième image lui correspond. Dès lors, impossible de retrouver le point de l'espace dont les points m1 et m2 seraient les images ! Des solutions existent malgré tout. La première consiste à éclairer les objets avec un laser. Seul le point éclairé par le laser sera visible dans les deux images. Pas de doute possible : le point vu dans la première image correspond à celui vu dans la deuxième, d'où la position du point éclairé dans l'espace.

Comment tenir compte des parties cachées par d'autres objets ou par l'objet lui-même sur les photographies ?

Pour reconstituer tout un objet, il faudra alors le balayer point par point avec le laser : c'est le principe du « plan laser », véritable scanner tridimensionnel. Efficace mais lent ! Imaginez un instant que la nature ait doté l'homme du même système ! Une deuxième solution consiste pour un point m1 donné à rechercher dans la deuxième image quels points m2 lui sont les plus semblables, c'est-à-dire lesquels sont le plus vraisemblablement les images d'un même point de l'espace. Si plusieurs points m2 possibles sont détectés, il faudra que l'ordinateur choisisse tant bien que mal tout seul lequel retenir ou qu'il demande à un opérateur de choisir pour lui. De telles méthodes automatiques ou semi-automatiques sont couramment utilisées, notamment en cartographie à partir d'images aériennes.



A partir de deux portraits (en jaune) seulement, l'ordinateur reconstruit la forme du visage en trois dimensions. C'est un premier pas vers un système d'identification : il suffirait de deux caméras au-dessus d'une porte et de quelques calculs pour reconnaître le visage d'une personne.

Nous sommes pourtant encore loin de l'objectif annoncé : acquérir une carte tridimensionnelle complète de ce qui est observé. Parmi les problèmes en suspens, notons-en deux principaux. Premièrement, ne traiter que deux images d'un objet ne permet de reconstituer que la partie de l'objet qui fait face à l'observateur. Il faudrait donc prévoir un procédé qui puisse reconstruire tout l'objet, à condition bien sûr que l'observateur en fasse tout le tour. Plus subtil est le deuxième problème, à la base de bien des dysfonctionnements de la deuxième des solutions décrites ci-dessus. Lorsqu'on observe un objet, certaines parties en sont cachées, par d'autres objets ou par l'objet lui-même. Ces parties cachées dépendent de la position de l'observateur. Pensez aux murs cachés par les toits dans le cas de la cartographie : sur deux images différentes, ce ne seront pas les mêmes murs qui seront visibles. En conséquence, rechercher absolument un point m2 correspondant à un point m1 donné est parfois source d'erreur. Dans le cas d'un observateur faisant tout le tour d'un ensemble d'objets, ce problème risque fort de devenir le principal obstacle, les objets se cachant les uns les autres et jamais de la même manière.

C'est néanmoins à ces deux problèmes que répond le procédé de reconstruction tridimensionnelle dont nous allons exposer succinctement le principe. Il a été développé conjointement par Olivier Faugeras de l'INRIA de Sophia Antipolis, professeur au Massachusetts Institute of Technology et par l'auteur[1]. Ce procédé n'a pu aboutir que grâce à de récentes avancées mathématiques sur la théorie des évolutions de surfaces d'une part, et d'autre part, sur leurs méthodes de simulation informatique, c'est-à-dire sur ce que l'on appelle techniquement le domaine des équations aux dérivées partielles (EDP), domaine qui a donné à la France l'une de ses récentes médailles Fields en la personne de Pierre-Louis Lions. L'ordinateur travaille en fait sur les surfaces délimitant les objets à reconnaître. Il émet d'abord une hypothèse sur la forme et la position de ce qui est observé, hypothèse consistant en une ou plusieurs surfaces. A partir de là, il vérifie la justesse de l'hypothèse en chacun des points de ces surfaces. Il faut pour cela déterminer, pour un point donné, les caméras qui doivent le voir et celles pour lesquelles il est caché. S'il y a bien un objet au point en question, alors les images qu'en ont les caméras qui le voient doivent être très semblables.

La bonne solution requiert d'ajouter une dimension, le calcul se faisant dans un espace de dimension quatre

A partir de ces tests en chacun des points de l'hypothèse, les surfaces sont déformées en de nouvelles surfaces dont la mesure de validité est plus grande, et ainsi de suite jusqu'à l'obtention de surfaces stables représentant fidèlement les objets observés.

Tests de validité, puis déformation : en pratique, comment réaliser ces opérations sur ordinateur ? Cela suppose d'abord de quantifier précisément les ressemblances entre deux points A et B. En réalité, pour les images numériques, ces points sont des pixels* d'une couleur donnée, composée de trois couleurs élémentaires : vert, rouge et bleu. On pourrait se limiter à l'examen des couleurs en chacun des points A et B, procédé intuitif, mais trop imprécis. Il faut en fait s'intéresser à de petites fenêtres rectangulaires centrées en A et B. Ce sont ces deux

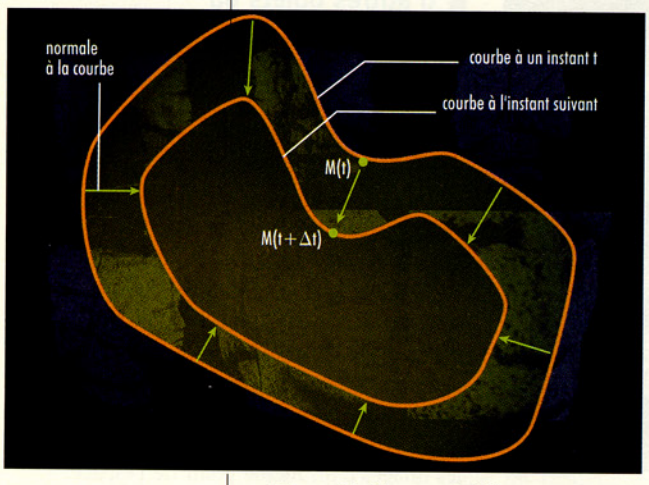


Figure 1. Comment déformer une courbe plane ? La solution intuitive consiste à déplacer chaque point selon la normale à la surface. Mais un système de vision informatique établi ainsi donnerait des résultats faux ! Les chercheurs ont dû trouver une transformation mathématique plus élaborée.

(1) Renaud Keriven. « Equations aux dérivées partielles et espaces d'echelle : applications à la vision par ordinateur. » Rapport de Thèse. 1997. Disponible sur demande à keriven@cermics.enpc.fr.

(2) Nicholas Ayache. *Vision stéréoscopique et perception multisensorielle : applications à la robotique mobile.* InterEditions, 1989.

fenêtres que l'on compare, et même plus précisément les pixels qui les composent. De plus, même en supposant, ce qui n'est qu'une approximation, qu'un objet émet la même quantité de lumière dans toutes les directions de l'espace, les deux images ne sont pas nécessairement prises dans les mêmes conditions d'éclairage, ni avec des caméras identiques. Il faut donc faire plus qu'une simple comparaison de couleurs des pixels de chaque fenêtre. La solution consiste à calculer ce qu'on appelle un facteur de corrélation normalisé entre les fenêtres, ce qui revient à comparer non plus les couleurs des pixels, mais la façon, dont ces couleurs varient au sein de chaque fenêtre. Ce facteur varie entre –1 pour des fenêtres totalement anti-corrélées et +1 pour des fenêtres totalement corrélées.

Déformer des surfaces. Une fois réglée cette question de corrélation, il reste encore à savoir comment déformer les surfaces. Mathématiquement, elles évoluent suivant une équation aux dérivées partielles, exactement comme nombre de phénomènes physiques variant au cours du temps. Concrètement, on transforme la surface par de petites variations imposées en chacun de ses points. Ces points évoluent selon la normale* à la surface. La longueur du déplacement, appelé vitesse normale, est bien sûr différente en chaque point M. Dans notre cas, elle dépend de la position de M, de la forme de la surface en M et des facteurs de corrélation des différentes caméras voyant M. On fait varier la vitesse normale de telle sorte que le procédé s'arrête aux points où l'objet est détecté (elle s'annule quand la corrélation est bonne). De plus, on l'adapte suivant des critères mathématiques rigoureux assurant une évolution « en douceur » de la totalité de la surface vers la solution.

Supposons, pour simplifier, que nous n'ayons plus à déformer une surface dans l'espace, mais une courbe dans un plan, et que cette courbe soit fermée. Une façon naïve de simuler son évolution sur ordinateur consiste à la mémoriser sous la forme d'un ensemble de points, que l'on déforme en faisant bouger les points le long de la normale (fig. 1). Cette manière de procéder est mauvaise car très instable : en certains endroits, les points se resserrent tandis qu'en d'autres ils s'écartent. L'approximation numérique devient vite incorrecte, et les résultats, faux. La bonne méthode, appelée « méthode des courbes de niveau », consiste à rajouter une dimension, et à inscrire cette courbe dans une surface évoluant, elle, dans l'espace.

On choisit une surface initiale dont la courbe est le niveau zéro, c'est-à-dire la trace de la surface sur le plan horizontal d'altitude zéro (fig. 2). C'est maintenant la surface qu'il s'agit de faire évoluer, mais pas n'importe comment : on veut que son niveau zéro finisse par représenter la solution ! C'est là qu'il faut faire appel à la théorie des EDP. Elle fournit en fait une équation qui permet de déformer la surface exactement de sorte que la courbe, elle, évolue vers la solution. Première qualité : en pratique, on est assuré de trouver la bonne solution. Cette technique possède aussi une propriété remarquable : rien n'empêche le niveau zéro de se séparer en plusieurs morceaux (fig. 2). La détection d'objets multiples se fait donc sans effort supplémentaire. Enfin et surtout, elle s'étend naturellement aux dimensions supérieures, soit, dans notre cas, à l'évolution d'une surface et non plus d'une courbe (il faut alors faire évoluer une hypersurface en dimension 4 dont le niveau zéro est la surface à déformer). Et cela marche ! Les étapes successives montrent de premières surfaces approximatives qui se déforment, se précisent et finissent par coller parfaitement aux objets photographiés. Les parties se cachant mutuellement ne trompent en rien l'algorithme.

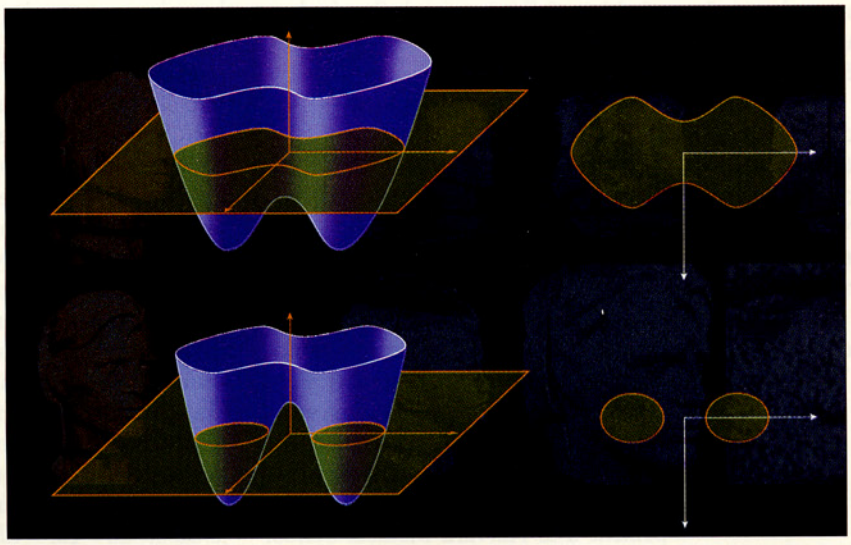


Figure 2. Une courbe plane (ici en orange) peut être vue comme la trace d'une surface de l'espace (en bleu). Les transformations mathématiques sur la surface se répercutent alors sur la courbe. En exploitant cette astuce en vision informatique, on est assuré d'obtenir le bon résultat. La courbe peut se séparer aussi en plusieurs morceaux, même si la surface, elle, reste d'un seul tenant.

La plus belle application de ces recherches, remplacer la vue humaine quand elle est défaillante, est encore loin

Jusqu'ici, les méthodes d'acquisition à base de caméras étaient sans justification mathématique rigoureuse, et surtout peu probantes. Elle ne reconstituaient les objets que partiellement, et se trompaient quand ils se cachent mutuellement. Ce nouveau procédé ouvre de nombreux horizons. Nous avons déjà évoqué le cinéma tridimensionnel ou l'acquisition automatique de mondes virtuels. Nous pourrions aussi citer la reconnaissance des formes. Les résultats dans ce domaine seraient grandement améliorés si l'on travaillait sur des données 3D plutôt que sur des images planes. La robotique est également demandeuse[2] : un robot mobile autonome doit aujourd'hui se contenter de sonars pour réagir vite et de plans lasers pour les tâches précises. Ces sonars sont d'ailleurs inutilisables dans le vide, donc en exploration spatiale, où pourtant éviter un obstacle est vital ! Les spécialistes du trucage et des effets spéciaux seraient ravis de pouvoir extraire toute l'information tridimensionnelle possible d'une séquence filmée. Ils pourraient alors très facilement modifier un acteur, le remplacer par un autre, ou même ajouter des objets ou d'autres acteurs à la séquence. Le résultat serait plus réaliste qu'avec les méthodes actuelles. Ce n'est là qu'une première avancée, car en réalité la véritable motivation de tels travaux est de pouvoir imiter un jour la vision humaine et la remplacer quand elle est déficiente. Le chemin est encore long **R.K.** ■

Pour en savoir plus

■ Olivier Faugeras. *Three-dimensional Computer Vision : a Geometric Viewpoint.* MIT Press, 1993.

■ Berthold K.P. Horn. *Robot Vision.* MIT Press, 1986.

■ Thierry Viéville. *A few Steps Towards 3D Active Vision.* Springer Series in Information sciences, 1995.

***PIXEL**
petit carré formant l'unité élémentaire d'une image numérique.

***NORMALE À UNE SURFACE**
direction perpendiculaire au plan tangent à la surface. Plus simplement, c'est la direction qui « fuit » par rapport à la surface.

WEB+
web.larecherche.fr