Texte de la 253e conférence de l'Université de tous les savoirs donnée le 9 septembre 2000.

TRAITEMENT NUMERIQUE DES IMAGES ET VISION PAR ORDINATEUR

par Olivier FAUGERAS

Le mot d'image recouvre une réalité très diverse : on parle d'images naturelles lorsqu'on entend des images du monde dans lequel nous vivons, on parle d'images de synthèse lorsqu'on entend des images qui ont été calculées par un programme d'ordinateur sans que la scène représentée existe nécessairement, on parle enfin, et sans épuiser le sujet, d'images médicales quand on pense à une représentation picturale du corps humain obtenue à l'aide des modalités d'exploration telles que l'Imagerie par Résonance Magnétique ou IRM.
Les images d'aujourd'hui sont dans leur immense majorité des images numériques c'est-à-dire que les différents paramètres dont elles dépendent tels que le temps, les coordonnées d'espace, l'intensité lumineuse, la couleur, ont été contraints à ne prendre leurs valeurs que dans des ensembles de nombres entiers. Ces processus dits de quantification et d'échantillonnage permettent de représenter sans perte d'information les images comme des tableaux de chiffres sur lesquels on peut calculer. Les tableaux étant très grands, il est en général nécessaire pour effectuer les calculs de faire appel à un ordinateur, l'image numérique et son traitement sont nés. Ces traitements peuvent être classés grossièrement en deux catégories : si le résultat du calcul est une nouvelle image, meilleure que l'original selon certains critères de qualité, on parle d'amélioration ou de restauration d'images, si le calcul a pour but d'extraire de l'image une certaine information pour effectuer une tâche ou prendre une décision, on parle d'analyse d'images.

Un peu d'histoire

Du point de vue historique on peut dire que le traitement et l'analyse des images numériques sont nés aux États Unis dans les années soixante, sous l'impulsion des militaires nord américains. Le point de vue des chercheurs de l'époque était soit pragmatique (« il faut que le traitement marche et résolve le problème qui m'est posé »), soit naïf comme dans le cas de Marvin Minski, professeur au Massachusetts Institute of Technology, qui confia dit-on à quelques étudiants désœuvrés la mission de résoudre le problème de la vision (comment calculer automatiquement à partir de quelques images d'une scène la structure tridimensionnelle de celle-ci). Il est notoire que les étudiants échouèrent puisque ce problème est aujourd'hui encore largement ouvert.
Cette situation perdura jusqu'au début des années 1980 où elle changea soudainement avec l'arrivée en scène d'une personnalité tout à fait remarquable, David Marr, neurophysiologiste et mathématicien ; anglais d'origine, il fut recruté au MIT par le même Marvin Minski et mit sur pied un programme de recherche en vision par ordinateur qui fut interrompu par sa mort prématurée. L'une des idées force de Marr est que la vision biologique, donc humaine, a pour but de construire une représentation tridimensionnelle des surfaces des objets qui nous entourent. Cette idée dont on peut penser qu'elle est un peu réductrice s'est avérée très féconde en permettant de définir un programme de recherche précis : comprendre en détail comment une telle

représentation peut être construite. Sa fécondité n'aurait pas été aussi grande si elle n'avait pas été complétée par deux autres idées qui ont permis d'échafauder une méthodologie pour atteindre ce but.
La première est que la vision ainsi définie est un problème de traitement de l'information qui peut s'étudier en quelque sorte indépendamment de l'organisme qui le résout : l'animal ou l'ordinateur ont à effectuer la même tâche. La seconde idée est une idée réductionniste : les différentes fonctions visuelles telles que la perception de la couleur, des textures, du mouvement ou de la distance sont dans une très large mesures indépendantes les unes des autres et peuvent donc être étudiées séparément. On sait aujourd'hui qu'il ne s'agit là au mieux que d'une approximation mais qui a permis de faire avancer de manière très significative notre connaissance de la perception visuelle.
À partir de ces prémisses, David Marr propose de mener l'étude en distinguant soigneusement trois niveaux : Le niveau de la théorie computationnelle où l'on construit la théorie du processus visuel étudié, où l'on analyse les équations qui le décrivent et les contraintes qui le caractérisent ; le niveau de l'étude algorithmique, où l'on cherche à construire les représentations, les schémas numériques et les algorithmes, qui permettront de résoudre effectivement les équations qui ont été proposées au premier niveau ; enfin le niveau de la réalisation physique de ces algorithmes où l'on précise comment ces représentations peuvent être construites, ces calculs effectués soit dans un ordinateur, dans une rétine ou un cerveau.
Outre David Marr, il est évident aujourd'hui que les progrès remarquables de notre connaissance des mécanismes biologiques et computationnels de la perception visuelle ont été accomplis grâce aux apports de deux sciences que sont les mathématiques et l'informatique. Les mathématiques ont apporté des outils d'analyse précis comme la géométrie algébrique, la géométrie différentielle, le calcul des variations, la théorie des équations aux dérivées partielles qui, comme en physique, ont permis de construire des théories computationnelles de nombreuses fonctions visuelles. L'informatique a apporté ses théories des représentations des structures de données et des algorithmes, de l'analyse de leur complexité, ses langages et ses outils de programmation, ses architectures à base de microprocesseurs permettant d'expérimenter avec des structures de calcul non conventionnelles.
Je vais prendre deux exemples qui illustrent assez bien, je crois, ces apports. Le premier est tiré de la géométrie, le second du calcul des variations.

Géométrie et vision par ordinateur

Felix Klein, mathématicien allemand de la fin du XIX^{e} siècle, s'intéressa beaucoup aux propriétés des figures géométriques qui ne changent pas, dites invariantes, lorsque l'on applique à ces figures des transformations. Celles-ci peuvent être des déplacements (rotations, translations) ou bien des transformations plus compliquées. L'une des notions importantes dans ce contexte est celle de groupe de transformations qui permet de donner un sens à la concaténation de plusieurs transformations. Les déplacements forment un groupe mais des groupes plus grands comme le groupe affine ou le groupe projectif jouent un rôle important en vision par ordinateur. Le jeune Felix Klein dès 1872, alors qu'il était professeur à l'Université d'Erlangen, développe l'idée selon laquelle les propriétés importantes des figures géométriques sont celles qui sont invariantes sous l'action d'un groupe particulier. Réciproquement, toute quantité invariante d'une figure exprime une propriété géométrique intéressante de cette figure. Le programme d'Erlangen comme on l'appelle aujourd'hui a donc consisté à explorer de manière systématique cette relation entre

groupes de transformations et invariants des figures géométriques. Ce programme a été couronné de succès en mathématiques bien sûr mais aussi en physique où il a donné naissance à des interprétations nouvelles des lois de conservation des grandeurs physiques et où il a fini par envahir pratiquement toutes les branches.

La raison de son succès en vision par ordinateur est qu'une caméra est une machine projective au sens où elle effectue une opération de projection de l'espace tridimensionnel sur la rétine qui elle, est bidimensionnelle. On trouve une illustration de cette remarque dans des gravures d'Albrecht Dürer comme celles reproduites en figure **{fig:1}** où l'on voit un peintre s'efforcer par des moyens mécaniques, un modèle, de respecter les lois de la perspective récemment découvertes à l'époque. La version moderne de ce modèle utilise la géométrie projective ce qui permet de distinguer deux types de paramètres, ceux qui définissent la position et l'orientation de la caméra dans l'espace, dits paramètres extrinsèques, et ceux qui définissent la façon dont l'image fournie par la caméra est distordue, dits paramètres intrinsèques. Ces paramètres sont en général inconnus et l'un des problèmes que l'on rencontre dans de nombreuses applications est celui de les estimer.

La méthode traditionnelle, photogrammétrique, pour réaliser cette estimation est d'utiliser une grille d'étalonnage, c'est-à-dire un objet de géométrie connue : l'observation de la distortion de la géométrie de l'image de cet objet permet de remonter aux paramètres intrinsèques de la caméra. Cette méthode est néanmoins difficile, voire impossible, à utiliser dans certaines applications comme la réalité augmentée. Un apport des chercheurs en vision par ordinateur est d'avoir remarqué que les paramètres intrinsèques de la caméra pouvaient être calculés directement à partir de l'image, sans faire intervenir de grille d'étalonnage, grâce à un objet purement mathématique : il s'agit d'un cercle de rayon imaginaire situé dans le plan à l'infini. Ce cercle s'appelle l'ombilic. L'image de l'ombilic par la caméra est une conique, également imaginaire, dont les paramètres sont liés de manière très simple à ceux de la caméra. Comme l'ombilic est toujours présent dans l'environnement on comprend qu'on puisse, grâce à lui, se passer d'une grille d'étalonnage.

Prenons un exemple. La figure **{fig2:photos-inria}** montre quelques images du site de l'INRIA Sophia-Antipolis prises avec un appareil de photo. Aucun des paramètres de l'appareil n'était connu lors de la prise de vue. Néanmoins, ces paramètres ont pu être calculés avec précision en utilisant l'ombilic comme grille d'étalonnage. Une fois ceux-ci connus, on peut évaluer la structure tridimensionnelle de la scène par des méthodes de type stéréoscopique : on obtient le modèle géométrique de la figure. Ce modèle peut alors être enrichi à l'aide des images qui ont servi à le calculer et on obtient l'image de la figure **{fig4:inria-images }** dans laquelle sont combinées géométrie, intensité lumineuse et textures.

Une autre application de la géométrie projective sont les mosaïques d'images. Si l'on prend plusieurs images d'une scène d'un même point de vue, la géométrie de ces images peut être décrites très simplement à l'aide du groupe projectif du plan. Concrètement cela veut dire que, étant données deux images quelconques de l'ensemble, on peut passer de l'une à l'autre à l'aide d'une transformation de ce groupe qu'on appelle une homographie. Pour illustrer mon propos, regardons la figure **{fig:5carlton}** où sont présentées quelques images de la ville de Cannes prises d'un même point de vue. La figure **{fig6:carlton-mosaique}** montre comment on peut assembler ces images dans une même structure qu'on appelle une mosaïque : dans celle-ci, l'une des images a été prise comme référence (celle qui a conservé sa forme rectangulaire, au centre de l'image), et on appliqué à chacune des autres une transformation homographique ce qui a pour effet de les rendre compatibles avec l'image de référence au prix d'une distorsion parfois significative. La mosaïque ainsi obtenue a de multiples applications, notamment dans les domaines de la réalité virtuelle ou de la réalité augmentée.

Calcul des variations et vision par ordinateur

Dans la section précédente, nous nous sommes principalement attachés à retrouver la structure tridimensionnelle de la scène à partir de quelques images sans nous soucier d'y identifier ou d'y reconnaître des objets ou des formes. Le but de cette section est de montrer comment une autre grande classe de techniques issues des mathématiques modernes, le calcul des variations, permet dans de nombreux cas de répondre à de telles questions.
La solution utilise deux grandes idées. La première est que l'objet recherché est représenté par un modèle mathématique, disons un ensemble de courbes et de surfaces qui dépendent d'un certain nombre de paramètres. La seconde idée est que l'on peut exprimer l'adéquation du modèle aux mesures, c'est-à-dire aux images, à l'aide d'un critère qui mesure une sorte d'énergie : plus l'énergie est faible, meilleure est l'adéquation et inversement, plus l'énergie est élevée et moins bonne est l'adéquation. Le problème est donc de rechercher quels sont les paramètres du modèle qui minimisent l'énergie. Une fois ces paramètres trouvés, si l'énergie correspondante est suffisamment basse, on en conclut que le modèle est effectivement présent dans la scène.
La question est maintenant de savoir comment minimiser notre énergie. La réponse se trouve précisément dans la théorie du calcul des variations qui permet, partant d'une hypothèse de départ (un jeu de paramètres plausible) correspondant à un modèle initial, de faire évoluer ces paramètres, par exemple en résolvant une équation aux dérivées partielles (EDP), jusqu'à ce que l'on ne puisse plus faire décroître l'énergie. Cette EDP va donc déformer le modèle initial et le transformer en un meilleur modèle, meilleur au sens où le modèle transformé explique mieux les mesures, c'est-à-dire comme nous le disions plus haut, les images.
Montrons ces idées à l'œuvre sur un exemple. En figure **{fig7:coureur}** nous voyons trois images simultanées d'un homme en train de courir prises par trois caméras de télévision. Le but est de caractériser sa course en étudiant les vitesses de paramètres tels que son centre de gravité ou les angles de certaines de ses articulations. Pour cela on utilise un modèle tridimensionnel tel que celui qui apparaît en figure **{fig8:modele}.** Ce modèle est défini par un grand nombre de paramètres et on va chercher à les estimer à partir des trois séquences vidéo. Comment faire ? On part d'une hypothèse et on place le modèle dans une certaine position dans la scène. À partir de cette hypothèse, on va calculer les images de ce modèle telles qu'elles seraient formées par les trois caméras et on va les comparer aux trois images réellement obtenues. Elles seront en général différentes et on va alors modifier (à l'aide de techniques issues du calcul des variations) les paramètres du modèle afin de rapprocher le plus possible les images hypothétiques et les images réelles. La figure **{fig9:coureur-modele}** montre l'un des résultats obtenus : on a superposé les images hypothétiques aux images réelles de la figure et on voit que les paramètres du modèle ont été correctement estimés : la superposition est bonne.

L'avenir

Je voudrais conclure par une évaluation des résultats obtenus par les techniques de vision par ordinateur. Je ne souhaite pas laisser le lecteur sur l'impression que le problème de la vision tel que je l'ai défini au début de ce texte est aujourd'hui résolu car ce n'est pas le cas. Certes des applications de plus en plus complexes et spectaculaires faisant intervenir de la perception visuelle peuvent être résolues automatiquement par informatique mais quand on y regarde de plus

près, l'homme est souvent encore dans la boucle c'est-à-dire qu'il est nécessaire de faire appel à ce merveilleux système de perception qu'est la perception visuelle humaine pour pallier les défauts des systèmes de perception visuelle artificiels. Quelles sont les raisons de cet échec partiel ? Elles tiennent en quelques mots. Les systèmes de vision par ordinateur sont dans une large mesure incapables de s'adapter à des variations de l'environnement qui n'ont pas été explicitement prévues dans leurs programmes et de s'enrichir à partir de leurs erreurs, leurs capacités d'apprentissage sont voisines de zéro. Or ceci est exactement à l'opposé de ce que l'on observe chez les singes dits supérieurs et chez l'homme. Il semblerait donc que la vision par ordinateur soit passée pour le moment à côté de quelque chose d'essentiel qui n'a pas encore été pris en compte dans les théories computationnelles.

Pour tenter d'aller de l'avant il me paraîtrait naturel de retrouver un peu de l'esprit de David Marr qui, je le rappelle, prônait l'étude simultanée ou en tout cas en liaison très étroite de la vision biologique et de la vision machine. Or si l'on fait l'état des lieux en cette année 2000, on s'aperçoit que très peu de laboratoires travaillant sur la perception visuelle ont à la fois une activité dans le domaine de la vision biologique et dans celui de la vision computationnelle. Comment faire pour que des liens se tissent à nouveau entre les deux communautés ?

Une piste s'ouvre peut-être du côté de ce qu'on appelle aujourd'hui les méthodes non invasives d'exploration fonctionnelle du cerveau. L'Imagerie par Résonance Magnétique (IRM), la Magnétoencéphalographie (MEG) complétée par l'Electroencéphalographie (EEG), la caméra à positron (PET) permettent de mesurer certains aspects de l'activité cérébrale avec des précisions spatiales et temporelles qui s'améliorent chaque année. Appliquées à la perception visuelle elles peuvent fournir des renseignements précieux sur son fonctionnement qui pourraient potentiellement enrichir les théories et les algorithmes développés en vision par ordinateur. Inversement certaines des idées développées en traitement d'images numériques sont susceptibles de trouver des applications en vue d'améliorer la qualité et la précision des mesures fournies par l'IRM, la MEG et l'EEG, la PET.

On voit donc bien là s'esquisser une synergie entre des mondes trop éloignés les uns des autres aujourd'hui (neuro-sciences de la perception, informatique, mathématiques et vision par ordinateur) mais dont le rapprochement annoncé laisserait espérer le développement d'une théorie commune de la perception visuelle biologique et artificielle qui déboucherait sur une meilleure compréhension fondamentale de la perception visuelle en général et sur une efficacité accrue des méthodes de vision par ordinateur.

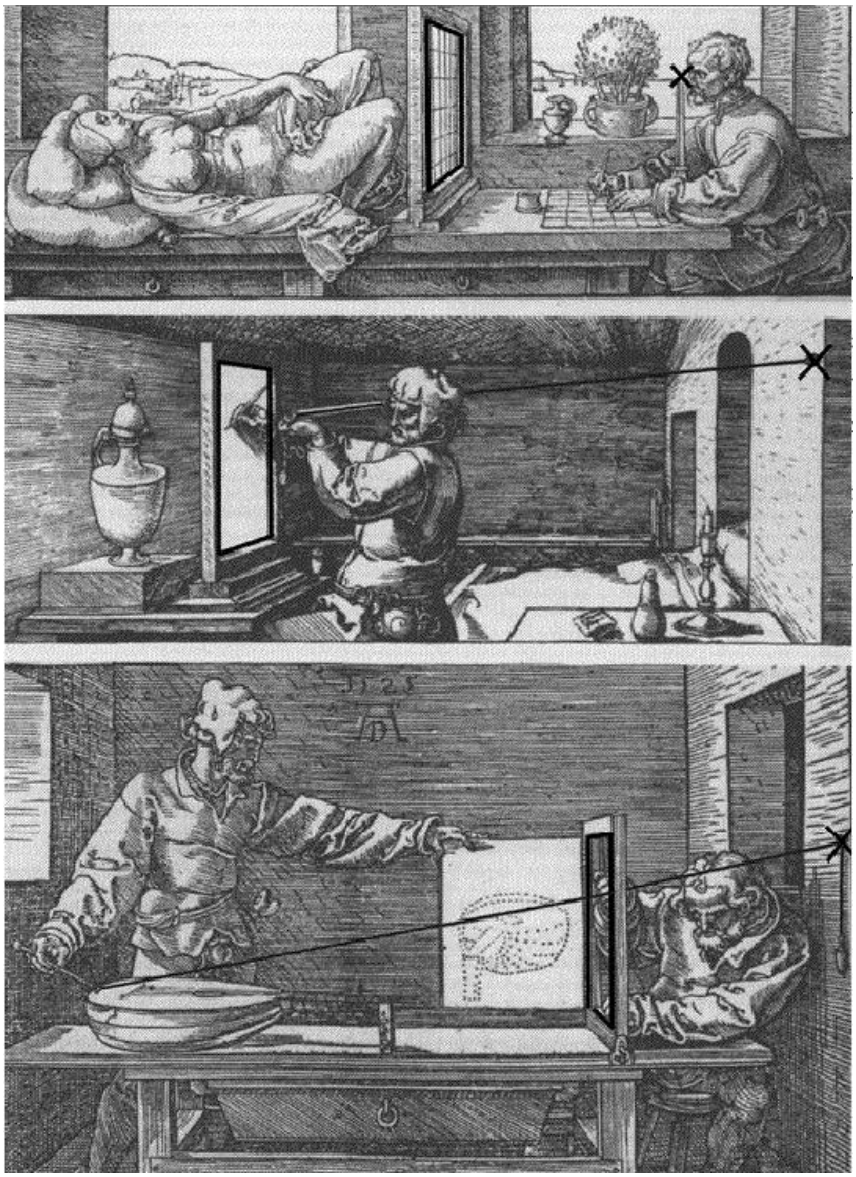

fig1

fig2

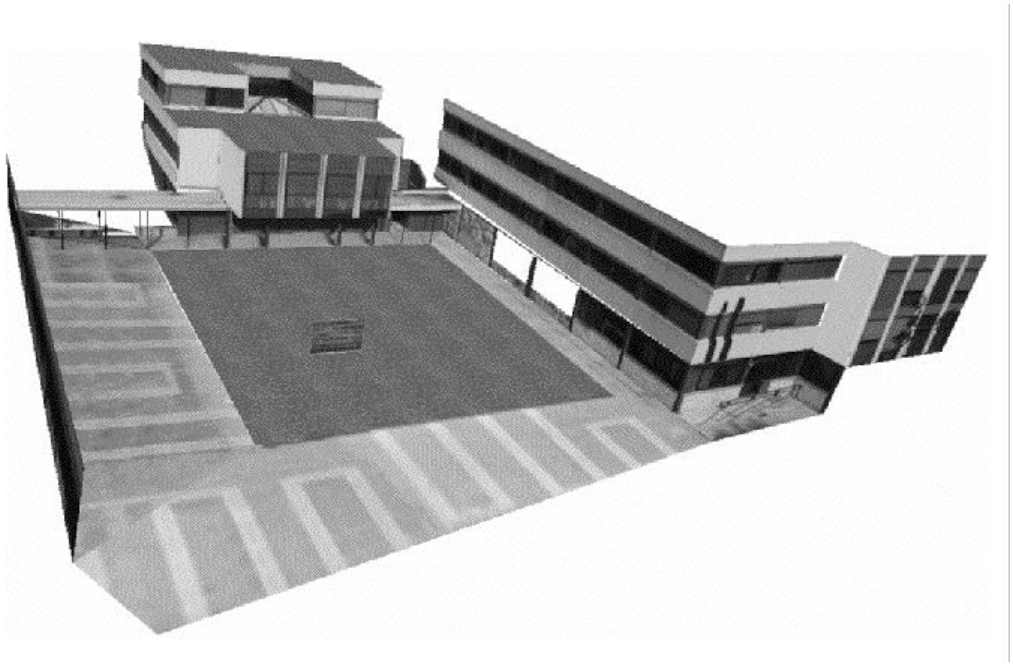

fig4

fig 5

fig6

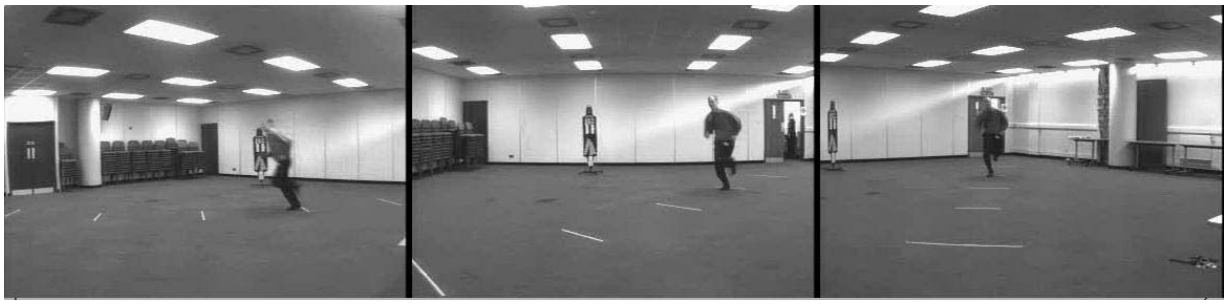

fig7

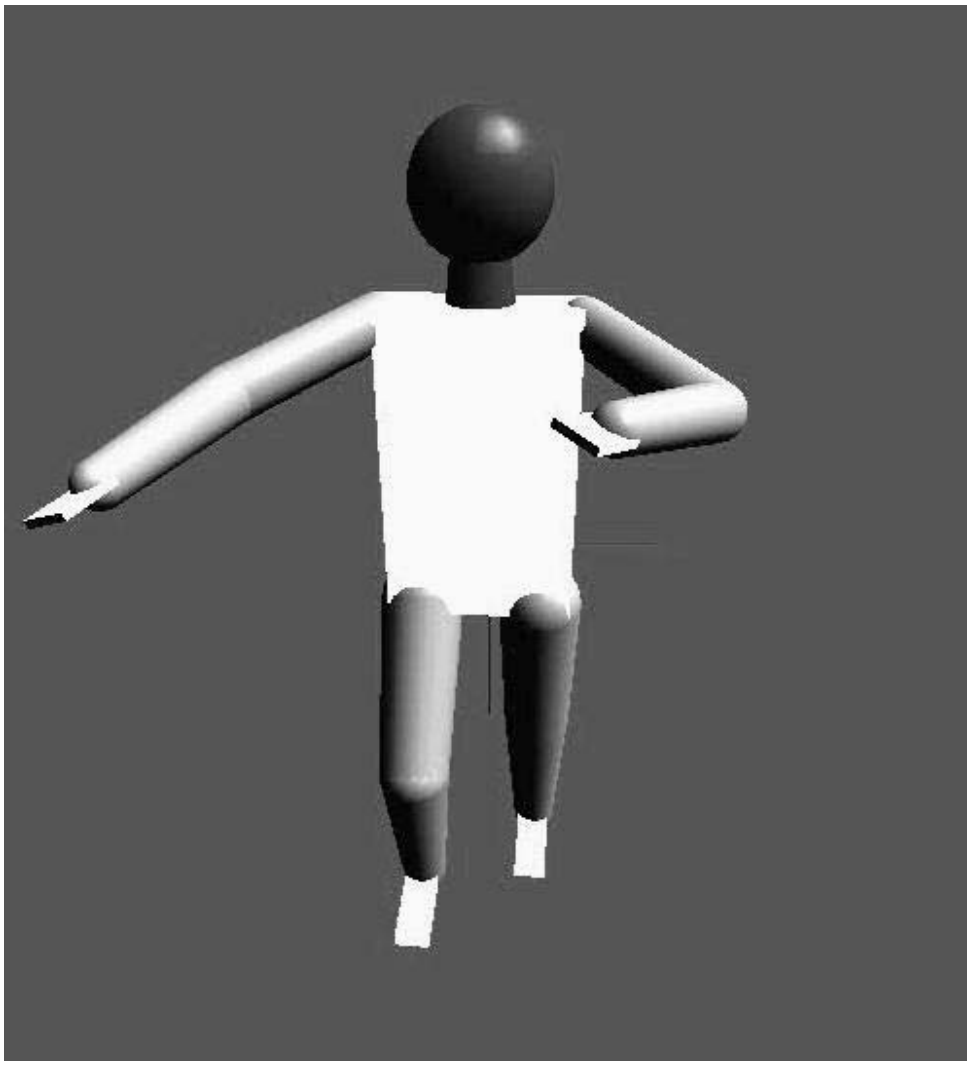

fig 8

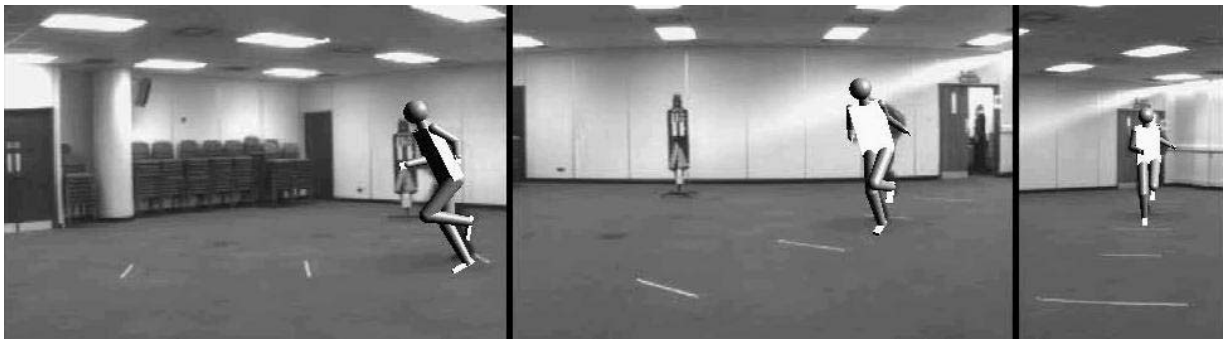

fig9