



4. UN DÉTECTEUR DE BLAGUES détecte les mots d'esprit et l'humour du comique Jay Leno. Le système informatique repère les changements dans le rythme, l'intonation et les gestes de J. Leno.

l'arrière-plan, le détecteur de personnes crée des surfaces dotées d'une certaine texture. À chaque point de ces surfaces correspond une valeur moyenne de couleur et une répartition autour de cette moyenne.

Chaque fois que le logiciel détecte une nouvelle image dans le flot d'images vidéo, le détecteur de personnes compare l'image avec les modèles qu'il a élaborés, ainsi qu'avec d'autres références. Tout d'abord, le système essaie de déterminer la forme que devrait prendre le modèle de taches dans la nouvelle image. Si, par exemple, le torse de la personne se déplaçait vers la droite à une vitesse de un mètre par seconde un dixième de seconde plus tôt, le système considère que le centre de la tache représentant le torse s'est déplacé de un dixième de mètre vers la droite. Ces estimations sont également fondées sur des modèles de mouvements calculés à partir de tests effectués par des milliers de personnes. Par exemple, nous avons observé que les taches qui correspondent au torse se déplacent plus lentement que celles qui correspondent aux pieds et aux mains.

Lorsqu'il a terminé ses prévisions, le détecteur de personnes calcule la probabilité que chaque pixel de la nouvelle image appartienne aux différentes taches : il calcule la différence entre les valeurs de couleur et de luminosité du pixel et les valeurs moyennes de couleur et de luminosité de toutes les taches ; puis il compare les résultats obtenus avec la répartition de chaque tache afin de déterminer la probabilité que la différence soit le fruit du hasard. Si, par exemple, la différence de luminosité entre un pixel et une tache est de dix pour cent et que les statistiques sur cette tache indiquent qu'une telle

différence ne se produit que dans un pour cent des cas, la probabilité que le pixel appartienne à la tache n'est que de 0,01. Les ombres ne présentent qu'un problème mineur dans la mesure où elles provoquent des différences de luminosité très marquées. Dès lors, le système recherche les ombres, des zones qui sont plus sombres que le niveau de luminosité moyen, et nivelle leur nuance de couleur et leur saturation à l'aide de la luminosité générale de la zone observée.

Le système doit aussi corriger les légères modifications de l'éclairage et de la disposition des objets dans la pièce, qui pourraient faire que l'ordinateur place certains pixels dans de mauvais modèles. Pour contourner cette difficulté, le système met continuellement à jour les pixels qui sont visibles derrière l'utilisateur, calculant une moyenne à partir des anciennes et des nouvelles informations. De la sorte, le système assure le suivi des modifications qui surviennent. Par exemple, lorsque l'utilisateur déplace un livre, l'ordinateur modifie son modèle de la pièce en deux endroits : à l'endroit où se trouvait le livre et à l'endroit où il a été déplacé. Au terme de ces calculs et de ces compensations, le système affecte chaque pixel de la nouvelle image au modèle qui lui correspond le mieux. Enfin, il met à jour les statistiques qui décrivent le modèle de taches et la scène d'arrière-plan, ainsi que celles qui anticipent la direction que prendront les mouvements des taches.

Qui et comment ?

Une fois que les pièces intelligentes savent où se trouvent leurs occupants, elles doivent déterminer qui sont ces

personnes et ce qu'elles disent. De nombreux algorithmes permettent aux ordinateurs de comprendre le langage, mais la plupart des systèmes de reconnaissance de la parole ne fonctionnent correctement que si l'utilisateur porte un microphone ou s'en trouve proche. Aussi avons-nous cherché un système qui permettrait à l'ordinateur de décoder le langage d'un visiteur lorsque celui-ci se déplace dans une pièce, même lorsque le bruit de fond est important.

Nous nous sommes aidés du système de détection qui suit en permanence la position de l'utilisateur. À l'aide de ses informations, le système de reconnaissance vocale commande électroniquement une série de microphones fixes, de sorte qu'ils ne renforcent que les sons qui proviennent de la direction de la bouche de l'utilisateur. Puisque le son voyage à une vitesse fixe, il parvient en des lieux différents à des moments distincts. Ainsi, si le système utilise les sons enregistrés par une série de microphones et leur ajoute les décalages temporels pour tenir compte de la position, il renforce le son qui provient de cette position. Ensuite, il compare le son global capté avec les sons de mots connus jusqu'à trouver une correspondance.

Une pièce intelligente doit également savoir qui y parle ou qui lui parle : cette identification est souvent aussi importante que de reconnaître les paroles prononcées. Comme la manière la plus rapide d'identifier un individu est de reconnaître son visage, nous avons mis au point un système de reconnaissance faciale. Pour utiliser la méthode du maximum de vraisemblance, ce système devait tout d'abord construire des modèles à partir de visages qu'il

Les «ordifringues»

Une pièce intelligente fonctionne un peu comme un majordome. Sans gêner ses occupants, elle cherche constamment à les aider. Ne pourrait-elle nous accompagner en permanence en écoutant et en voyant tout ce que nous faisons et disons et en anticipant nos besoins ? Des collègues de l'auteur cherchent à intégrer des caméras, des microphones et des ordinateurs dans les vêtements. Ces «vêtements intelligents», surnommés ordifringues, peuvent fournir des informations sur l'environnement de leur utilisateur, tels le nom de personnes qu'il rencontre ou le lieu de sa prochaine réunion. En outre, ils peuvent remplacer la plupart des ordinateurs et des assistants électroniques utilisés habituellement.

Le projet de vêtements intelligents de l'Institut de technologie du Massachusetts a commencé en 1992. L'intégration d'ordinateurs, de caméras, de microphones et autres capteurs aux vêtements portés par une personne donne au programme le point de vue de celui qui porte les vêtements.

Par exemple, en intégrant une caméra dans la monture d'une paire de lunettes, le logiciel de reconnaissance des visages peut identifier les interlocuteurs et murmurer leur nom à l'oreille. Si des microphones sont placés dans une veste, un logiciel de reconnaissance vocale

peut rappeler des données importantes : à l'évocation d'un contrat, par exemple, le logiciel pourrait afficher les données financières associées à ce contrat sur l'écran intégré aux lunettes. Si un récepteur du système GPS de localisation par satellite était intégré à vos chaussures, un logiciel de navigation pourrait vous indiquer votre chemin.

Nos prototypes actuels de vêtements intelligents utilisent des casques commerciaux qui assurent confort et confidentialité. Leurs unités centrales sont peu encombrantes. Nous avons mis au point des unités de saisie de données qui autorisent l'utilisation de ces machines dans presque tous les contextes. De surcroît, elles utilisent des connexions sans fil au réseau *Internet*.

Notre but est de proposer des objets suffisamment petits et légers pour pouvoir être utilisés constamment, comme les lunettes et les montres aujourd'hui, qui mettraient à votre disposition la puissance d'un ordinateur. Les vêtements intelligents ne sont pas encore très discrets, en particulier les objets portés sur la tête, donnent une allure de «cybernaute» (voir la photographie). Toutefois, l'ère de l'«informatique permanente» est proche. Deux de nos collègues se sont ainsi transformés en «cyborgs» : ils portent leurs ordifringues continuellement.



«connaissait». Nous avons utilisé une technique mathématique nommée analyse par vecteur propre, afin d'obtenir des «visages propres», à partir desquels n'importe quel autre visage est obtenu par simple combinaison linéaire.

Lorsque la caméra détecte un occupant dans la pièce, le système de détection trouve la position de son visage, le système d'identification extrait l'image du visage de l'arrière-plan et il normalise son contraste. Le système calcule alors la combinaison linéaire qui correspond au visage observé et il compare le modèle avec celui des personnes connues. Si le calcul des similarités fournit un résultat proche, le système considère qu'il a identifié l'utilisateur. À l'aide de cette méthode, nos pièces intelligentes ont reconnu des visages, parmi plusieurs centaines de personnes, avec 99 pour cent de succès.

L'expression faciale est au moins aussi importante que l'identité. Un programme d'enseignement, par exemple, devrait savoir si ses étudiants ont l'air de s'ennuyer. Aussi, lorsque notre pièce intelligente a trouvé et identifié le visage de quelqu'un, elle analyse son expression (*voir la figure 2*). Un autre ordinateur compare alors les mouvements faciaux enregistrés par la caméra avec des cartes qui décrivent les mouvements faciaux qui accompagnent plusieurs expressions du visage. À chaque expression correspond une combinaison caractéristique de mouvements musculaires. Lorsque nous sourions, par exemple, nous relevons les coins de la bouche et nous soulevons certaines parties du front ; cependant, lorsque nous faisons semblant de

5. DES VOITURES INTELLIGENTES sont étudiées par les informaticiens de l'Institut de technologie du Massachusetts et par leurs collègues du Centre de recherche de la Société Nissan. Andy Liu et l'auteur ont réalisé un système informatique qui détermine la nature des actions d'un conducteur quelques instants après qu'elles ont commencé. Ils ont obtenu ce résultat en couplant leur système de reconnaissance avec un simulateur de conduite automobile où des capteurs enregistraient les mouvements des mains, des jambes et des yeux de personnes qui conduisaient sur un trajet virtuel. À l'aide de ce système, ils ont identifié des activités types, auxquelles on peut maintenant comparer les mouvements des conducteurs en situation de conduite réelle : pour savoir ce qu'un conducteur s'apprête à faire, l'ordinateur compare ses mouvements aux modèles identifiés. Une voiture qui saurait quand vous avez choisi de changer de voie pourrait vous avertir si un camion se trouve dans votre angle mort.





sourire, nous ne bougeons que la bouche. Au cours de nos expériences, le système a correctement reconnu les expressions d'un petit nombre de sujets dans 98 pour cent des cas.

Quoi?

La reconnaissance du visage, de l'expression et du langage d'une personne n'est que la première étape. Les pièces intelligentes doivent ensuite placer ces perceptions dans leur contexte : les mêmes mouvements peuvent être interprétés de différentes manières selon les intentions de la personne qui les effectue. Lorsque nous conduisons une voiture, par exemple, nous cessons d'accélérer soit parce que nous souhaitons ralentir, soit parce que nous abordons un virage ; toutefois, dans ce dernier cas, nous tournons simultanément le volant. Pour savoir ce que nous faisons à un instant précis, un ordinateur embarqué dans une automobile doit prendre en compte la modification de l'ensemble des mouvements.

Pour concevoir un tel système, nous avons emprunté des techniques utilisées dans le domaine de la reconnaissance de la parole. Dans ces techniques, on modélise les mots sous forme de séquences de sons, nommées états internes. Chaque mot comprend une succession particulière d'états internes, qui sont soit des phonèmes (la plus petite unité d'une langue), soit des parties de phonèmes. Un système informatique essaie d'identifier des mots en comparant la séquence de sons qu'ils contiennent avec des modèles de mots, puis en sélectionnant la correspondance la plus probable.

Nous avons généralisé cette approche dans l'espoir de déterminer les intentions des personnes à partir de leurs mouvements. Nous avons élaboré un système informatique qui détermine, par exemple, si une personne qui lève le bras indique une direction ou si elle s'étire seulement. Le système reconnaît l'indication d'une direction en se référant à un modèle qui comprend trois états internes : lever le bras, le maintenir dans cette position et le ramener vers soi rapidement. L'étirement est plutôt un mouvement lent et régulier. Aussi, en observant ces états internes, caractérisés par l'accélération du mouvement de la main et par la direction prise par ce mouvement, notre système «comprend» ce que la personne est en train de faire.

À ce jour, nous avons réalisé plusieurs systèmes qui interprètent les actions humaines. Le système le plus simple permet aux occupants des pièces intelligentes d'utiliser leur corps pour commander des environnements virtuels : on utilise notamment la description de la forme de l'utilisateur afin de placer un modèle vidéo de celui-ci dans une scène de réalité virtuelle où se trouvent des représentations animées de créatures virtuelles. Ces dernières analysent des informations sur les gestes, les positions et les sons de l'utilisateur pour définir la manière dont elles interagiront avec lui. Par exemple, le chien virtuel Silas va chercher les objets qu'on lui lance : lorsqu'un utilisateur de la pièce intelligente imite les mouvements qui consistent à ramasser et à lancer la balle virtuelle de Silas, le chien «voit» l'image vidéo effectuer le même geste et va chercher son jouet. Silas s'assied et se met également sur le dos, sur commande.

L'occupant de la pièce peut aussi utiliser sa position et sa voix pour interagir avec un programme informatique. Les jeux informatiques, par exemple, sont beaucoup plus amusants quand on remplace les manettes de jeu par des environnements qui permettent de naviguer dans des espaces virtuels à trois dimensions : dans un jeu de combat, lorsqu'un adversaire apparaît à gauche, le joueur le tuera simplement en se tournant vers la gauche et en criant «Pan».

Pourquoi?

Les jeux en réalité virtuelle ne sont pas les seules applications des pièces intelligentes, loin de là. Ces dernières pourront, par exemple, aider les malentendants. Dans le langage des signes, une série de mouvements faits avec les mains assure la communication. Nous avons donc cherché à reconnaître ce langage : à partir de l'observation de nombreux exemples de mouvements des mains, nous avons construit un modèle pour chaque signe. Puis, en comparant les gestes effectués par une personne avec les modèles enregistrés, nous avons traduit un sous-ensemble de 40 mots du langage des signes en temps réel, avec 99,2 pour cent de réussite. Lorsque nous aurons augmenté la taille du vocabulaire compris par notre ordinateur, nous pourrons créer des interfaces pour les malentendants aussi fiables que les sys-



6. LE LANGAGE DES SIGNES est fondé sur un ensemble de gestes des mains. L'auteur et ses collègues ont mis au point un programme qui reconnaît nombre de ces gestes. Sur cette photographie, la caméra enregistre le mouvement des mains qui, en langage des signes, signifie «bicyclette». Le système informatique modélise son mouvement et le compare avec des modèles de signes connus. Lorsque la correspondance est trouvée, la machine prononce le mot «bicyclette».

tèmes de reconnaissance vocale d'aujourd'hui.

Les conducteurs d'automobile pourraient également bénéficier des pièces intelligentes. Chaque année, 10 000 Français sont tués dans des accidents de la route dus en majorité à des erreurs de conduite. En collaboration avec le centre de recherche de la Société Nissan, nous avons construit un véhicule qui enregistre les actions des conducteurs et qui fournit des indications utiles, des instructions et même des avertissements.

Afin de répertorier un ensemble de modèles de conduite caractérisés par les gestes que font les conducteurs lors-

qu'ils dépassent, suivent un véhicule, tournent, s'arrêtent, accélèrent ou changent de file, nous avons enregistré les mouvements des mains et des jambes d'un grand nombre de conducteurs installés dans un simulateur (voir la figure 5). Puis nous avons utilisé les modèles résultants afin de classer les actions de chaque conducteur testé aussi vite que possible. De manière assez surprenante, le système parvient à déterminer ce que font les conducteurs presque au moment où l'action commence. Il identifie les actions avec 86 pour cent de succès, une demi seconde après le début de l'action. Quand on lui laisse deux

secondes pour identifier l'action, il la détermine correctement dans 97 pour cent des cas.

Ainsi, dans des situations simples, des systèmes informatiques modestes peuvent repérer les mouvements des personnes, les identifier et reconnaître leurs expressions en temps réel. En combinant plusieurs petits systèmes, on obtient des pièces intelligentes où, sans l'aide de claviers ni de souris, les occupants commandent l'affichage à l'écran, jouent avec des créatures virtuelles et communiquent en langage des signes. À l'aide de nos systèmes, nous étudions des applications variées. Par exemple, nous construisons des prototypes de lunettes qui identifient le visage d'une personne et murmurent son nom à l'oreille. Nous réalisons aussi des écrans de télévision qui savent si un téléspectateur les regardent et nous envisageons même des cartes de crédit qui reconnaîtront leur propriétaire et sauront si elles ont été volées.

D'autres équipes de notre laboratoire tentent de doter nos pièces intelligentes de la capacité de percevoir l'attention et les émotions, pour essayer de mieux comprendre les actions et les motivations humaines. Rosalind Picard espère mettre au point un système capable de déterminer si les conducteurs ou les étudiants sont distraits. Aaron Bobick écrit un programme qui interpréterait les mouvements des sportifs : il envisage de créer une caméra de télévision qui, au rugby par exemple, distinguerait une feinte d'une attaque décisive et suivrait automatiquement l'action. Progressivement, grâce au perfectionnement des pièces intelligentes, les ordinateurs deviendront des assistants attentifs. Sommes-nous sur la voie d'un monde où la distinction entre les objets animés et inanimés s'estompera?

Alex PENTLAND dirige le Groupe Ordinateurs et perception au Laboratoire des médias de l'Institut de technologie du Massachusetts.

P. MAES, T. DARRELL, B. BLUMBURG et A. PENTLAND, *The ALIVE System : Full-Body Interaction with Autonomous Agents*, in *Proceedings of Computer Animation '95*, 1995.

I.A. ESSA et A.P. PENTLAND, *Facial Expression Recognition Using a Dynamic Model and Motion Energy*, in *Proceedings of the Fifth International Conference on Computer Vision*, IEEE Computer Society Press, 1995.

T. STARNER et A. PENTLAND, *Real-Time American Sign Language from Video Using Hidden Markov Models*, in *International*

Symposium on Computer Vision, 1995, IEEE Computer Society Press, 1995.

A. PENTLAND et A. LIU, *Toward Augmented Control Systems*, in *Proceedings of the Intelligent Vehicles '95 Symposium*, IEEE Industrial Electronics Society, septembre 1995.

C. Wren, A. AZARBAYEJANI, T. DARRELL et A. PENTLAND, *Pfinder : Real-Time Tracking of the Human Body*, in *Integration Issues in Large Commercial Media Delivery System*, sous la direction de A.G. Tescher et V.M. Bove, SPIE, vol. 2616, 1996.

Des articles et des rapports techniques sont disponibles sur le site Web : <http://www-white.media.mit.edu/vismod>