



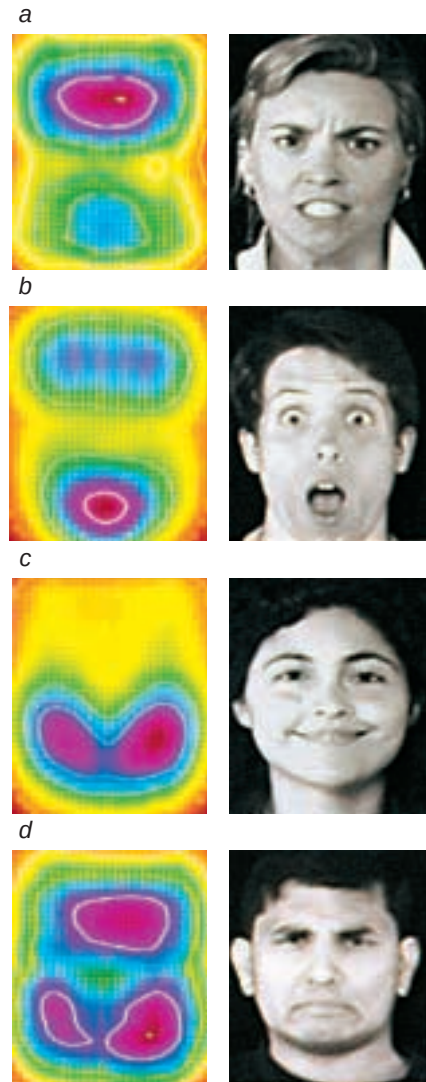
À l'Institut de technologie du Massachusetts, nous venons de mettre au point un ensemble de systèmes informatiques qui reconnaissent les visages, les expressions et les gestes. À l'aide de ces systèmes, nous avons réalisé des environnements qui s'approchent de la maison, du bureau ou de la voiture décrits précédemment. Ces «pièces intelligentes» sont équipées de caméras et de microphones dont les enregistrements sont transmis à un réseau d'ordinateurs. Ces derniers déterminent ce que font et disent les occupants de la pièce intelligente, qui peuvent ainsi utiliser leurs gestes, leurs voix et leurs expressions faciales, au lieu d'un clavier ou d'une souris, pour interagir avec des programmes informatiques, pour parcourir des bases de données multimédias ou pour voyager dans le royaume de la réalité virtuelle.

La première pièce intelligente fut construite dans notre laboratoire en 1991, et il existe aujourd'hui cinq pièces de ce genre à travers le monde, toutes reliées par le réseau téléphonique : trois à Boston, une au Japon et une au Royaume-Uni. Des installations similaires sont en projet à Paris, à New York et à Dallas.

Chaque pièce contient plusieurs micro-ordinateurs, chargés chacun d'une tâche spécifique. Quand une pièce intelligente doit analyser les images, les sons et les gestes, nous l'équiperons de trois ordinateurs, un par type d'interprétation. Si elle doit avoir des capacités supérieures, nous ajoutons des machines au réseau. Les ordinateurs utilisent tous la même méthode statistique, nommée analyse du maximum de vraisemblance : ils comparent les informations qu'ils reçoivent avec des modèles stockés en mémoire ; ils calculent la probabilité qu'un modèle corresponde aux données reçues, et choisissent le modèle le plus probable. Ces comparaisons permettent aux ordinateurs de répondre à une série de questions sur leurs utilisateurs, y compris déterminer leur identité et, parfois, ce qu'ils veulent.

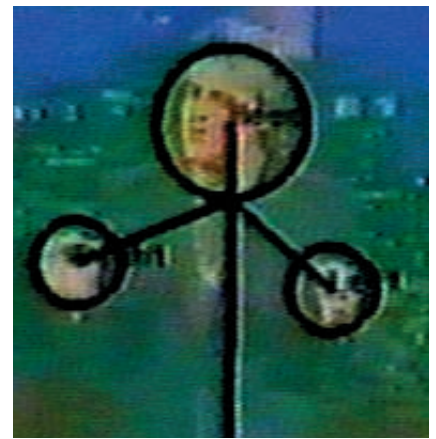
Où?

Avant de chercher à savoir ce que font ses occupants, la pièce doit les localiser. Aussi avons-nous mis au point un détecteur de personnes qui suit les mouvements d'un individu à travers la pièce. Nous l'avons vu, ce détecteur utilise l'analyse du maximum de



3. DANS LES CARTES DE MOUVEMENTS, des couleurs chaudes montrent les parties du visage qui ont le plus bougé. Certaines expressions produisent des motifs de couleur typiques. Lorsque quelqu'un est fâché, par exemple, c'est surtout la périphérie des sourcils qui se modifie (a). La surprise se reconnaît aux sourcils et à la bouche (b), la joie s'exprime surtout par la bouche (c), et le dégoût par le visage tout entier (d).

vraisemblance. Tout d'abord, il crée un modèle de la personne enregistrée par la caméra sous forme de sept taches de couleur reliées (deux pour les mains, deux pour les pieds, et trois taches distinctes pour la tête, le torse et les jambes). Il décrit chacune des taches de deux manières : sous la forme d'une série de valeurs pour la couleur et l'emplacement de la tache ; et sous la forme d'une «carte de support», une liste qui indique les pixels de l'image qui appartiennent à la tache (les pixels sont les éléments d'une image, semblables aux points qui composent l'image de votre téléviseur). Ensuite, pour modéliser



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