Interfaces multimodales pour un assistant au voyage

Alain Goyé, Eric Lecolinet, Shiuan-Sung Lin, Gérard Chollet

GET-ENST
46, rue Barrault
75634 Paris Cedex 13
goye | elc | lin | chollet@enst.fr

Catherine Pelachaud

IUT de Montreuil - University of Paris 8
140, rue de la Nouvelle France
93100 Montreuil, France
c.pelachaud@iut.univ-paris8.fr

Xiaoqing Ding

Dept. of Electronic Engineering
Tsinghua University
Beijing, 100084, China
dingxq@tsinghua.edu.cn

Yang Ni

Institut National des Télécommunications
Département Electronique et Physique
9,Rue Charles Fourier
91011 Evry Cedex-France
yang.ni@int-evry.fr

RESUME

Dans le cadre du développement d'un assistant personnel pour voyageurs, nous étudions trois types d'interfaces multimodales pour PDA : 1) la combinaison de Control menus et d'entrées vocales pour contrôler des interfaces zoomables vers des bases de données de type graphique ou textuelles, 2) l'affinement d'images captées par la caméra intégrée, basé sur la corrélation d'une série d'images, pour améliorer la reconnaissance de caractères, et 3) des agent conversationnels animés enrichis de comportements et de gestes culturels. Nous présentons ici ces trois modalités et leur intégration dans l'application globale.

MOTS CLES : interfaces zoomables, Control menus, affinement d'images, agent conversationnels culturels.

ABSTRACT

As a part of a project to develop a personal assistant for travellers, we are studying three types of multimodal interfaces for a PDA: 1) a combination of Control menus and vocal inputs to control zoomable user interfaces to graphical or textual databases, 2) refinement of pictures captured by the integrated camera, based on correlating a series of pictures, in order to enhance character recognition, and 3) embodied conversational agents able to communicate via synchronized speech and culture-dependent nonverbal behaviors (face, gaze and gesture). We describe here these three modalities and their integration in the main application.

KEYWORDS : zoomable user interfaces, Control menus, image refinement, embodied conversational agents.

INTRODUCTION

Ces travaux se placent dans le contexte du projet Lingtour, dont l'objectif est de développer un assistant personnel (PDA) permettant à un voyageur à l'étranger d'accéder facilement à des informations pratiques et culturelles, de communiquer avec la population et de s'orienter dans une ville. En collaboration avec l'université Tsinghua de Pékin, nous développons une démonstration à l'usage des touristes occidentaux en Chine.

OBJECTIFS

Parallèlement à un travail sur la gestion multilingue des informations, l'objectif du projet est d'exploiter au mieux les possibilités du PDA pour la multimodalité : utiliser conjointement, en l'absence de clavier, les entrées de l'écran tactile, du microphone et de la caméra, et exploiter alternativement ou simultanément les possibilités graphiques et sonores, selon le contexte, pour échanger l'information.

APPROCHES TECHNIQUES

L'application est construite autour des accès à 3 bases de données liées entre elles, contenant respectivement des informations linguistiques, géographiques et touristiques. Son architecture générale est représentée Figure 1. Du point de vue des IHM, les aspects innovants du projet résident essentiellement dans trois aspects : 1) l'interrogation des bases de données par une combinaison d'interfaces utilisateur zoomables utilisant des Control menus 2D, et d'entrées vocales utilisant un moteur de reconnaissance multilingue, 2) la capture et le traitement d'images pour l'extraction et l'interprétation du texte qui s'y trouve, et 3) la délivrance des informations selon diverses modalités, incluant des agents conversationnels

dont le comportement est adapté au contexte social (tenant compte de l'utilisateur) et technique (PDA).

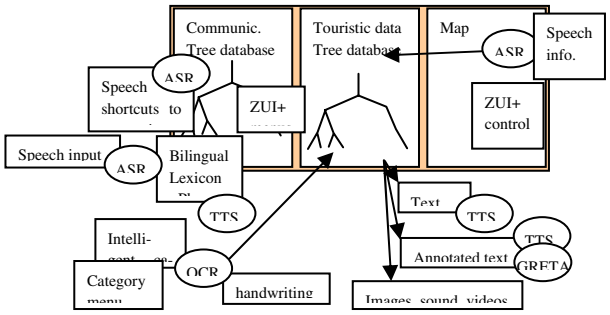

Figure 1: Architecture de l'application.

L'interface « geste et voix »

L'emploi d'interfaces utilisateurs zoomables (ZUIs), contrôlées par des Control menus, semble particulièrement adapté au PDA qui dispose d'un écran de taille réduite. Pourtant peu d'applications sur PDA intègrent ce type d'interface. Les ZUIs sont basées sur le concept de zoom sémantique, révélant progressivement différents niveaux de détails de l'information. Les Control menus [1] (Figure 2) permettent de sélectionner et de contrôler des actions (par exemple un déplacement, un zoom, etc.) en un seul et même geste. Cette propriété permet d'éviter les changements de contexte, l'utilisateur n'étant plus obligé de manipuler plusieurs interacteurs situés à des endroits différents de l'écran pour effectuer une même action. Ces menus offrent par ailleurs une meilleure adéquation entre l'action logique voulue par l'utilisateur et l'interaction physique requise par l'interface. Les interfaces classiques requièrent généralement une décomposition des actions en sous-actions nécessitant des interactions différentes (par exemple lorsque l'utilisateur veut "zoomer" l'interface il lui faut généralement d'abord effectuer la sous-action "se mettre en mode zoom" en cliquant sur un bouton, puis la sous-action "zoomer" en manipulant la souris, un potentiomètre, etc.). Cette décomposition semble avoir un impact non négligeable sur la fluidité de l'interaction comme en attestent les travaux de Buxton [4] et d'autres études (en particulier sur les Flow menus, une technique d'interaction proche des Control menus et légèrement postérieure [7]). Enfin, de même que les Pie et Marking menus dont ils s'inspirent [5, 6], grâce à leur disposition spatiale les Control menus permettent aux utilisateurs de se remémorer les gestes servant à sélectionner les actions. Lorsque ces gestes sont suffisamment rapides, ces menus n'apparaissent plus à l'écran. Le passage du mode novice (apparition du menu) au mode expert (plus d'apparition) se fait donc implicitement en fonction du niveau d'habitude de l'utilisateur. Pour les raisons précédentes, nous avons choisi d'utiliser les Control menus afin de faciliter la navigation et localiser plus rapidement une information dans un contexte qui peut être très vaste. Ce choix semble etre validé par les observations informelles effectuées lors d'une précédente étude (concernant la navigation dans des bases de données bio-génétiques [2] et une bibliothèque électronique [3]). Dans notre application l'utilisateur contrôle ainsi au stylet la navigation dans deux types de données : a) dans un plan de ville, b) dans un lexique comportant les mots et phrases utiles au touriste, hiérarchisés en catégories telles que : hébergement/hôtel/réservation….

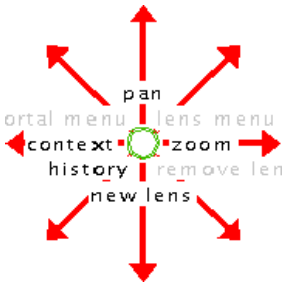

Figure 2: Control menu circulaire.

Nous ajoutons à cet usage des Control menus, la modalité « voix ». Notre application intègre un moteur de reconnaissance vocale qui fonctionne sur un vocabulaire limité, mais qui est indépendant du locuteur et ne requiert aucun apprentissage. La reconnaissance dans différentes langues partage des modèles acoustiques communs, ce qui facilite les extensions futures à de nouvelles langues. Les modèles peuvent également être adaptés à l'utilisateur et aux conditions d'usage. L'information vocale est employée différemment selon le contexte : a) au cours de la navigation dans le plan, l'utilisateur peut pointer un objet du plan et accéder par un menu vocal à diverses informations sur cet objet : description/horaires/tarifs-d-entrée/comment-y-aller… b) au cours de la navigation dans le lexique, comme raccourci d'accès aux catégories, puis pour l'accès à une entrée, mot ou phrase. La traduction de cette entrée sera affichée, et éventuellement synthétisée dans la langue cible. A terme, la convivialité des entrées vocales pourra être améliorée par une fonction de capture des mots-clés ("word spotting"). Bien entendu ces usages de la voix en entrée sont doublés par une alternative graphique, pour permettre l'usage du dispositif dans un environnement bruyant.

La caméra « intelligente »

Nous employons la caméra qui est de plus en plus souvent intégrée aux assistants personnels, pour la capture et la reconnaissance d'informations textuelles en langue étrangère [8] (Figure 3). La reconnaissance de caractères – chinois en particulier – atteint aujourd'hui de bonnes performances. Cependant la qualité et notamment la résolution d'une image saisie d'un peu loin dans la rue, avec une caméra bon marché, n'est souvent pas suffisante pour cette application. Nous travaillons donc à l'affinement des images, par corrélation et recalage d'une

séquence d'images successives. L'exploitation des légères différences dues au mouvement naturel de la main qui tient l'appareil, doit permettre de reconstituer l'image avec une résolution supérieure à celle du capteur.

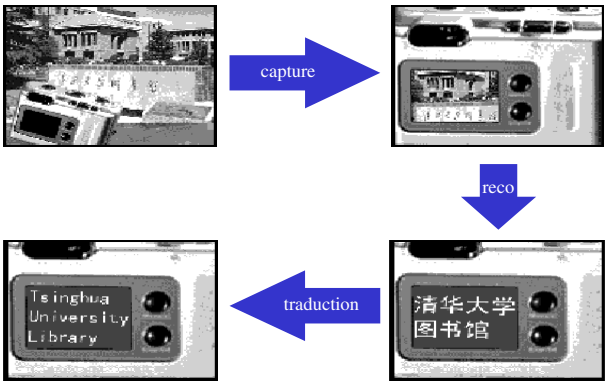

Figure 3: usage de la caméra.

Pour limiter la charge de calcul cette opération peut être effectuée sur une sous-partie de l'image. Cette sous-partie peut être sélectionnée semi-automatiquement lors de l'étape de délimitation et segmentation préalable à la reconnaissance. Le texte une fois reconnu peut être traduit, localement ou par accès à un serveur distant via un service de radiocommunication. Pour faciliter la traduction faite localement, un menu vocal permet de choisir le contexte, en indiquant s'il s'agit d'un panneau de bus / de rue / d'un monument, etc. Le texte traduit dans la langue de l'utilisateur, peut également être restitué par synthèse vocale.

Des agents conversationnels culturels

Les agent conversationnels (ACs) [9] permettent de transmettre une information de manière plus attractive et conviviale qu'une simple synthèse vocale (Figure 4). Les expressions nonverbales permettent de disambiguer un discours, de renforcer certains mots ou parties du discours… Elles fournissent des informations aussi bien au niveau syntactique que sémantique et émotionnel. Dans un contexte multiculturel, une démonstration visuelle peut aussi être le meilleur vecteur d'enseignement de certains usages.

Figure 4: expressivité d'un agent conversationnel.

Dans un premier temps, nous travaillons au portage sur PDA d'agents animés. La complexité et le niveau de détail de l'animation doivent être adaptés à la puissance et à la taille d'écran de l'appareil. Cependant, malgré de grands progrès récents en matière de réalisme, les agents actuels ne connaissent qu'un type de comportement, qui reflète le plus souvent la culture occidentale. Nous travaillons donc à concevoir un agent qui adapte son comportement au contexte culturel et social de l'utilisateur. La même information doit en effet être délivrée différemment, par exemple, à un Français et à un Chinois, mais aussi à un journaliste et à un particulier. Le comportement et l'animation de l'agent sont dirigés par une représentation sémantique indépendante de la langue, basée sur le standard XML-XSD. Ce langage permet une description de la fonction communicative des gestes ainsi que des signaux composant les gestes. Nous ajoutons à ces attributs classiques une sur-couche intégrant des attributs spécifiques à la culture. Par exemple, la culture influera sur le choix d'un geste (sourire ou hochement de tête), sur la durée d'un regard… Ces influences peuvent porter sur la définition d'un signal, mais aussi sur son intensité, le masquage d'un signal par un autre, etc.

RESULTATS ATTENDUS

A l'issue de ce projet, qui couvre l'année 2003, nous souhaitons être en mesure de démontrer : 1) la possibilité d'intégrer sur un PDA les parties existantes des diverses interfaces présentées ici : a) Control menus 2D, capture et reconnaissance de texte, b) agents conversationnels ; et 2) les bénéfices des améliorations que nous proposons pour chacune de ces fonctionnalités : intégration de commandes vocales dans les menus, affinement des images par corrélation, et enrichissement des agents par des attributs culturels.

BIBLIOGRAPHIE

1. Pook, S., Lecolinet, E., Vaysseix, G. et Barillot, E., *Control Menus: Execution and Control in a Single Interactor.* Proc. ACM conf. on Human Factors in Computing Systems (CHI) 2000, 263-264. ACM Press.

2. Pook, S., Lecolinet, E., Vaysseix, G. et Barillot, E., *Context and interaction in zoomable user interfaces*. Proc. conf. on Advanced Visual Interfaces (AVI) 2000, 227-231 & 317. ACM Press.

3. Plénacoste, P., Lecolinet, E., Pook, S., Dumas, C. et Fekete, J.D., *Bibliothèques : comparaisons entre le réel et le virtuel en 3D, 2D zoomable et 2D arborescent.* Actes conf. franco-britannique IHM-HCI (Interaction Homme-Machine / Human Computer Interaction) 2001.

4. Buxton, W., *Chunking and phrasing and the design of human-computer dialogue*. Information Processing, 1986, 475-480.

5. Hopkins, D., *The design of implementation of Pie menus*. Dr Dobb's journal of software tools, 1991, 16 (12), 16-26.

6. Kurtenbach, G., Fitzmaurice, G.W., Owen, R.N. et Baudel, T., *The Hotbox: efficient access to a large number of menu-items.* Proc. ACM conf. on Human Factors in Computing Systems (CHI) 1993, 231-327. ACM Press.

7. Guimbretière, F., Stone, M. et Winograd, T., *Fluid Interaction with High-resolution Wall-Size Displays.* ACM conf. on User Interface Software Technology (UIST) 2001. ACM Press.

8. Mao, Y., Dong, Q., Qi Y. et Chollet, G. *Realization of an Intelligent Camera capable of Character Recognition and Translation*. in Proc. of Sino-French Symp. on Speech and Language Processing, Beijing, October 2000. Disponible à l'adresse: http://www.tsi.enst.fr/~chollet/Projets/Chine/Lingtour/IntelCamera.doc.

9. Pelachaud, C., Carofiglio, V., De Carolis, B. et de Rosis, F., *Embodied Contextual Agent in Information Delivering Application*, First International Joint Conference on Autonomous Agents & Multi-Agent Systems, Bologna, July 2002.