Détection Automatique des groupes d'interactions

Maisonnasse Jérôme
PRIMA, GRAVIR-IMAG
INRIA Rhône-Alpes, France
38349 St. Ismier.
jerome.maisonnasse@inrialpes.fr

Brdiczka Oliver
PRIMA, GRAVIR-IMAG
INRIA Rhône-Alpes, France.
38349 St. Ismier.
oliver.brdiczka@inrialpes.fr

ABSTRACT
This work brings a new perception to intelligent environments for human activity recognition. To understand human activities, we need to differentiate users concerned and their interpersonal links. We propose a new group interaction detection tool based on an assumption that conversational turn taking is synchronized inside groups. We assess HMM-based and algorithmic approaches to build a detector which is able to segment interaction groups from speech detection. Both approaches show good results thus confirm the conversational hypothesis.

Categories and Subject Descriptors
I.5.3 [PATTERN RECOGNITION]: Clustering – *Algorithms, Similarity measures.*

General Terms
Algorithms, Performance, Human Factors,

Keywords
Groups of interaction clustering, Markov Model, speech detection, conversationnal analysis, intelligent environment, ubiquitous computing.

Résumé
Ce travail apporte une nouvelle perception aux environnements intelligents pour la reconnaissance de l'activité humaine. La prise de conscience d'une activité implique au préalable de différencier les protagonistes et leurs interdépendances. Nous proposons un outil de détection des groupes d'interactions basé sur une hypothèse conversationnelle, c'est-à-dire sur la synchronisation des prises de parole à l'intérieur des groupes. Deux approches sont évaluées afin de fournir un détecteur capable de discrétiser les groupes d'interactions. Ces dernières donnent de bons résultats, ce qui confirme la qualité de l'hypothèse conversationnelle.

1. INTRODUCTION
L'informatique ubiquitaire ouvre la voie vers des applications de plus en plus proches des activités humaines. Les environnements intelligents disposent de plus en plus de moyens de perception et d'action afin d'interagir et d'offrir des services au moment le plus opportun, en phase avec l'activité des utilisateurs. Il est alors important de détecter qui sont les utilisateurs potentiels du système et quels sont leurs liens par rapport à l'activité. Dans un même espace physique, plusieurs individus peuvent se regrouper et se fédérer autour d'une même activité, ou bien avoir des activités parallèles et complètement indépendantes les unes des autres.

Cette agrégation des utilisateurs en petits groupes indépendants permet une approche multi-échelle de la détection de l'activité. Le plus bas niveau de détection apporte une vision spécifique de l'activité principale en considérant chaque groupe comme une unité indépendante où les individus interagissent suivant un but donné. A un même moment, il est possible d'observer des activités similaires jouées en parallèle ou des activités radicalement différentes à l'intérieur des groupes. Le but de notre approche n'est pas de reconnaître l'activité en tant que telle, mais plutôt de délimiter les individus engagés dans cette activité où l'interaction verbale joue un rôle majeur et fédérateur.

Avec la reconnaissance de l'activité de haut niveau, la dynamique de fusion et de séparation des groupes au cours du temps nous permet de prendre conscience de l'émergence d'une activité globale. Lorsque plusieurs groupes indépendants fusionnent, on peut imaginer que la façon dont ces groupes fusionnent soit une information importante pour la détection de changement dans les activités en cours. Par exemple, dans une réunion de travail, les individus se regroupent par appartenance à une équipe et interagissent spontanément au sein de leur groupe. Dès que le meeting démarre, les individus se fédèrent autour d'une même tâche. Ce scénario illustre la manière dont la fusion de plusieurs groupes en un seul peut servir à la détection du commencement d'un meeting dans un environnement intelligent. Le même raisonnement est applicable pour la détection de la fin d'un meeting.

Dans la suite du document, nous nous intéresserons à travers la section 2 à la reconnaissance automatique des interactions. Puis nous présenterons successivement nos deux approches dans la section 3. Les résultats, ainsi qu'une discussion en section 4, précéderont la conclusion sur cette recherche en section 5.

2. ETAT DE L'ART
La reconnaissance de l'activité humaine à partir d'événements de parole est souvent employée dans le contexte de l'analyse de groupes. Dans cette perspective, le groupe et ses membres sont définis à l'avance. L'objectif est alors de reconnaître certaines actions clés de la part des membres [11] ou d'analyser de façon globale le type de meeting [3] à partir de la fréquence des tours de

UBIMOB 05, May 31. - June 3, 2005, Grenoble, France.

parole et leurs durées. Mais ce type d'approche spécifie implicitement l'étude d'un seul et unique groupe prédéfini dans un contexte particulier. Aucune détection n'est envisagée sur les dépendances entre les individus et leur appartenance à un ou plusieurs groupes. L'analyse des interactions à partir d'un grand ensemble de données enregistrées permet de reconstruire les réseaux sociaux d'un ensemble d'individus. Pour cela les conversations sont détectées et analysées [4]. Cependant, la détection automatique des conversations à partir d'événements de parole, à savoir qui parle et quand, avec l'information mutuelle [1] nécessite une durée d'enregistrement assez importante.

Dans ce travail, nous cherchons à développer un détecteur de groupes d'interactions temps réel, robuste et le plus généralisable possible, ayant pour but de définir les limites qui regroupent les individus autour d'une même activité intentionnelle. Cette dernière est dans un grand nombre de cas l'activité principale [8], dans la mesure où l'activité verbale nécessite un certain niveau d'attention. A noter que, l'activité d'un ensemble d'individus peut aussi, à certains moments, devenir l'activité d'autres qui porteront attention, donnant lieu à des focalisations collectives éphémères [10], ce que Goffman appelle les « cadres de participation »[9]. Notre objectif est de rendre visible cette dynamique afin que des environnements intelligents puissent en tirer parti dans la reconnaissance d'activités.

L'analyse conversationnelle [13] et la psychologie sociale ont mis en évidence un certain nombre de points remarquables sur les interactions interindividuelles. Partant du postulat selon lequel les interactions verbales au sein d'un groupe sont régulées [7], intentionnelles et composées de conversations successives qui visent à agir sur un pot commun d'informations [5][6], il est alors possible d'extraire une hypothèse conversationnelle pour la détection des groupes d'interactions. A partir de ces constatations, nous avons formulé une hypothèse conversationnelle comme critère significatif pour la détection de groupes d'interactions.

3. NOS APPROCHES

Nous présentons ici deux approches. La première relève d'une méthode algorithmique qui repose sur une matrice de distances dans l'espace des probabilités d'appartenir au même groupe. La deuxième méthode emploie une modélisation de Markov cachée. Toutes les données en entrée sont produites avec une détection automatique de la parole [14] à partir d'enregistrements audio. Nous savons à quel moment une personne commence et s'arrête de parler.

Les deux approches s'appuient sur l'hypothèse conversationnelle de base : lorsque deux personnes parlent en même temps, il est peu probable qu'elles appartiennent au même groupe. Si les deux individus ne parlent pas, on ne peut rien dire sur leur appartenance. Enfin, si une personne parle sur les deux alors il est probable qu'elles appartiennent au même groupe. Nous tirons profit de la régulation des prises de parole, considérant que les interactions verbales sont ordonnées à l'intérieur de chaque groupe, et désordonnées entre des groupes distincts. En ce qui concerne la notion de groupe, nous avons pris le parti de définir une configuration minimum de deux individus.

3.1 Approche algorithmique

Les interactions sont représentées par une matrice de distances symétrique M de dimension *n* x *n* à l'instant *t* dans l'espace de probabilité d'appartenir au même groupe, où *n* est le nombre de personnes enregistrées. Pour des raisons de simplicité et de sémantique, une interaction implique une relation symétrique entre deux individus. Si *i* parle à *j* et *j* écoute *i*, alors *i* et *j* interagissent également. Les m_{ij}^t varient entre 0 et 1. La distance 0 signifie que i et j sont très proches et appartiennent au même groupe, et la distance 1 signifie que *i* et *j* n'appartiennent pas au même groupe. A chaque instant *t*, nous évaluons la matrice de distance des interactions courante A qui est de même dimension que M où $a_{ij}^t = d^t(i,j)$. Soit *d* la distance entre deux personnes *i* et *j* définie par l'hypothèse conversationnelle.

$$d^t(i,j) = \begin{cases} 0 & si\ (parle\ _i^t \vee\ parle\ _j^t) \wedge \neg (parle\ _i^t \wedge\ parle\ _j^t) \\ 1 & si\ parle\ _i^t \wedge\ parle\ _j^t \\ 0.5 & si \neg parle\ _i^t \wedge \neg parle\ _j^t \end{cases}$$

Le facteur temps est représenté par un paramètre α, qui est un paramètre de mélange des deux matrices. Plus α est fort, moins la matrice d'interaction M est modifiée par la matrice d'interaction courante A $M^{t+1} = \alpha.M^t + (1-\alpha).A^t$

Avant de séparer les groupes, nous calculons au préalable de façon algorithmique l'ensemble des configurations de groupes possibles. Afin de fusionner et séparer les différents groupes, nous vérifions à partir de M que les plus grandes distances entre les individus à l'intérieur des groupes soient inférieures à la plus petite des distances entre les groupes. Lorsque ce critère est vérifié, il donne la configuration des groupes à l'instant *t.* Si ce critère n'est pas vérifié, nous considérons que tous les individus sont regroupés en un seul groupe.

3.2 Approche modèle de Markov caché

Cette approche repose sur l'estimation de la probabilité de prise de parole des personnes. On se base sur l'hypothèse que cette probabilité change selon la configuration des groupes. Pour déterminer la configuration actuelle, on utilise un modèle de Markov caché (*Hidden Markov Model*). Un modèle de Markov caché est une chaîne de Markov de premier ordre [12] dont chaque état génère une observation. Seules les observations sont visibles. Le but est de dériver la séquence des états « cachés » ayant généré ces observations. Dans notre cas, un état du modèle de Markov caché correspond à une configuration possible de groupe. Les observations du modèle correspondent à une discrétisation des événements de prise de parole comme indiqué sur les figures 1 et 2. Une observation est un vecteur contenant une valeur binaire (parole, non parole) pour chaque personne enregistrée. À chaque état du modèle de Markov caché est associée une distribution de probabilité d'observations. Cette distribution est basée sur l'hypothèse conversationnelle. On suppose que la probabilité que deux personnes appartenant au même groupe parlent en même temps est faible alors que cette probabilité est élevée dès que ces personnes ne sont plus dans le même groupe. Dans le cas d'un groupe comportant toutes les personnes, la probabilité de prise de parole d'une seule personne est élevée alors que la probabilité de prise de parole en parallèle est faible. La probabilité de transition entre les états est fixée à un niveau très faible par rapport à la probabilité de rester dans le même état. Ceci permet de stabiliser la détection car la fréquence d'observations est très haute par rapport à la fréquence de

changement de groupe (16ms par rapport à environ 30 secondes) et suppose que des changements de groupes surviennent dans des délais raisonnables. Afin de détecter les groupes en temps réel, on applique l'algorithme de Viterbi [12] sur le flux d'observations arrivantes. Étant donné un modèle de Markov caché et une séquence d'observations, l'algorithme de Viterbi calcule la séquence la plus probable d'états qui génèrent les observations. Dans notre cas, la séquence d'états correspond à la suite de changements de configurations de groupe. Nous calculons la séquence d'états pour la fenêtre des 5000 dernières observations, ce qui correspond à une fenêtre d'observation de 80sec. L'état calculé pour la dernière observation de la fenêtre représente la configuration actuelle de groupe. L'algorithme de Viterbi est recalculé toutes les 10 observations, ce qui correspond à un déplacement de la fenêtre de 10 observations et une estimation de la configuration de groupe toutes les 160 ms.

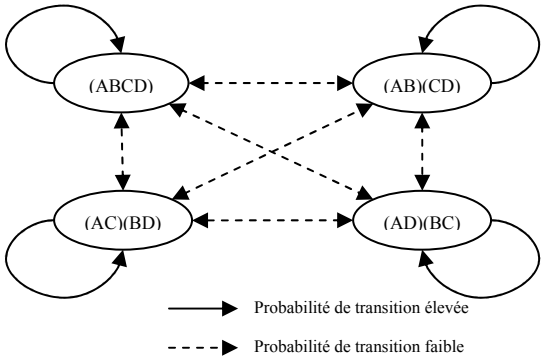

Figure 1. Représentation du HMM pour 4 configurations de groupes

Numéro d'Observation	Prise de Parole			
	A	B	C	D
0	0	0	0	0
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
4	0	1	0	0
5	0	1	0	1
6	0	1	1	0
7	0	1	1	1
8	1	0	0	0
9	1	0	0	1
10	1	0	1	0
11	1	0	1	1
12	1	1	0	0
13	1	1	0	1
14	1	1	1	0
15	1	1	1	1

Figure 2. Les états et les observations d'un modèle de Markov caché pour 4 personnes

4. Expérience et Résultats

Dans cette section, nous décrivons la façon dont nous avons évalué les deux approches. Puis nous commentons les résultats obtenus sur les scénarios de test.

4.1 Scénario d'évaluation

Nous avons réalisé trois expériences afin d'enregistrer les interactions entre quatre individus, interagissant suivant un script prédéfini de configurations de groupes. Chaque individu est équipé d'un micro-cravate. Pour les besoins de l'évaluation, nous contraignons l'utilisation du micro-cravate afin de minimiser le plus possible les fausses détections. Les sujets sont libres de leurs mouvements et de leurs discussions. La figure 3 ci-dessous, montre la configuration d'interactions de deux groupes indépendants.

Figure 3. Deux groupes indépendants interagissent. Le 1er groupe est au premier plan et le 2ème , au deuxième plan.

4.2 Résultats et discussion

Dans les deux approches, un résultat est obtenu toutes les 160 ms. Pour l'approche algorithmique, α est fixé à 0.0075 pour l'ensemble des expériences. En ce qui concerne l'approche Markovienne, les probabilités des états et de transitions sont prédéfinies à la main suivant l'hypothèse conversationnelle. Les matrices de confusion indiquent pour chaque configuration de groupes les bonnes et les mauvaises détections, avec en colonne le résultat des détections et en ligne la réponse attendue.

Table 1. Matrice de confusion Algorithme

Groupes	(ABCD)	(AB)(CD)	(AC)(BD)	(AD)(CB)
(ABCD)	**0.909**	0.055	0.01244	0.02329
(AB)(CD)	0.31775	**0.66095**	0.02018	0.00110
(AC)(BD)	0.26772	0.0	**0.72950**	0.00276
(AD)(BC)	0.20733	0.0	0.01920	**0.77346**

Table 2. Matrice de confusion HMM

Groupes	(ABCD)	(AB)(CD)	(AC)(BD)	(AD)(CB)
(ABCD)	**0.87535**	0.02737	0.06245	0.03460
(AB)(CD)	0.08535	**0.86707**	0.04549	0.00207
(AC)(BD)	0.21531	0.01057	**0.77411**	0.0
(AD)(BC)	0.03926	0.03629	0.07951	**0.84492**

Avec l'approche HMM, nous obtenons 84.8 % de détections correctes de configuration de groupe et 77.8% de détections correctes avec l'approche algorithmique. Pour 4 personnes, nous avons 4 configurations possibles. Les figures 4, 5 et 6 donnent un aperçu de la détection des groupes au cours du temps. Une ligne sur deux correspond à une configuration de groupes : ligne 0=> (ABCD), ligne 2 => (AB)(CD), ligne 4 => (AC)(BD) et ligne 6 => (AD)(BC). La ligne rouge indique les configurations de groupes réelles et attendues en sortie du détecteur.

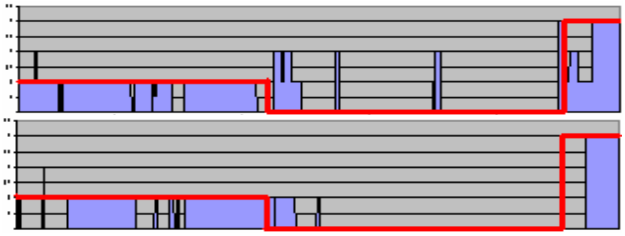

Figure 4. 1er Expérience : en haut résultat HMM, en bas résultat Algorithmique (durée : 9'22'')

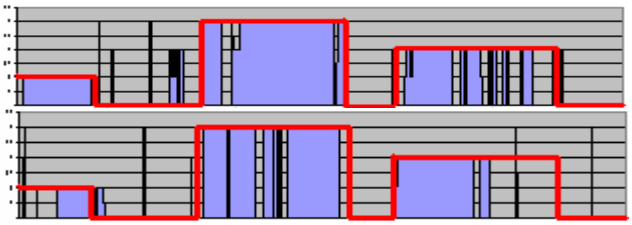

Figure 5. 2ème Expérience : en haut résultat HMM, en bas résultat Algorithmique (durée : 15'16'')

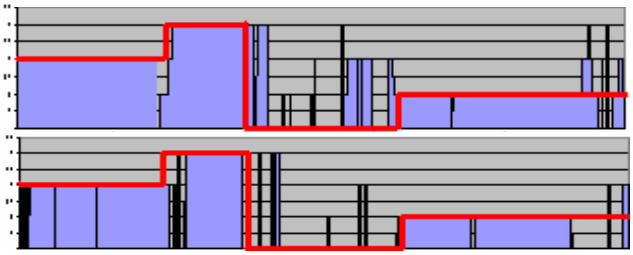

Figure 6. 3ème Expérience : en haut résultat HMM, en bas résultat Algorithmique (durée : 16'19'')

Les performances obtenues avec les deux modélisations suffisent à valider la qualité de l'hypothèse conversationnelle pour distinguer des groupes d'interactions. L'approche algorithmique de par sa simplicité suffit à la détection des groupes d'interactions et définit une performance plancher qui nous servira de référence. L'approche Markovienne, avec l'algorithme de Viterbi et la fenêtre glissante d'observations, se montre plus robuste bien qu'un peu plus sensible aux fausses détections (voir graphiques ci-dessus). Cette instabilité pourrait être corrigée à partir d'un meilleur ajustement des probabilités de transitions entre les états qui représentent les différentes configurations de groupes. L'approche algorithmique semble moins sensible aux fausses détections mais cela dépend du paramètre α qui tend à homogénéiser les réponses lorsqu'il est fort, tout en diminuant le nombre de détections.

5. Conclusions

Nous présentons un détecteur de groupes d'interactions robuste et fonctionnant en temps réel. Ce dernier se base sur la seule information des événements de parole et sur l'hypothèse de synchronisation des tours de parole à l'intérieur des groupes. Le résultat obtenu avec les deux approches est honorable dans la mesure où cette seule information ne peut suffire à désambiguïser toutes les situations. De plus, le modèle doit être étendu pour détecter les individus isolés et en nombre supérieur. Cette limitation pourra être réduite en utilisant d'autres indices comme l'orientation de la tête et les distances interindividuelles qui semblent être de bons indicateurs [2]. Une approche multimodale doit être alors envisagée. En ce qui concerne les algorithmes, pour le modèle Markovien, un apprentissage sur un ensemble représentatif de données pourrait valider les probabilités définies à la main et les ajuster. Une autre façon de voir l'apprentissage consiste à spécialiser la détection pour un contexte donné, ce qui diminuera les performances en généralisation mais augmentera les performances pour le cas particulier. Concernant l'approche algorithmique, on révisera la méthode de séparation des groupes qui ne convient que pour un petit nombre d'utilisateurs. D'autres approches de clustering peuvent être envisagées à partir de la matrice de distances, ainsi qu'une optimisation du paramètre de α.

6. REFERENCES

[1] Basu S. *Conversational Scene Analysis.* Ph.D. Thesis. MIT Department of EECS. September, 2002.

[2] Beattie, G. *The regulation of speaker turns in face-to-face conversation;some implications for conversation in sound-only communication channels*, Semiotica, 34, pp.55-70, 1981

[3] Burger, S., MacLaren, V., and Yu, H., *The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style,* Proc. of ICSLP 2002 (International Conference on Spoken Language Processing), Denver, CO, USA, 2002.

[4] Choudhury, T. and Pentland, A., *Characterizing Social Interactions Using the Sociometer*, Appears in: Proceedings of NAACOS 2004, June 2004

[5] Clark H. *Using Language*. Cambridge, England : Cambridge University Press, 1996.

[6] Clark H., Schaefer E., F. *Contributing to Discourse*, Cognitive Science 13, 259-294, 1989.

[7] Duncan, S., *Some signals and rules for taking speaking turns in conversation*, Journal of Personality and social psychology, 23(2), pp.283-293, 1972.

[8] Goffman, E. (1974). *Frame analysis*. New York: Harper Row.

[9] Goffman, E. (1987). *Façons de parler*. Paris: Minuit. (Edition originale, 1981, A. Kihm, Trad.).

[10] Joseph, I., *Attention distribuée, attention focalisée. Les protocoles de la coopération au PCC de la ligne A du RER*. Sociologie du Travail, 4/94, 563-587, 1994.

[11] McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D., *Automatic Analysis of Multimodal Group Actions in Meetings,* in "IEEE Transactions on Pattern Analysis and Machine Intelligence (to appear)", 2004.

[12] Rabiner, L. A tutorial on Hidden Markov Models and selected applications in speech recognition, Proc. IEEE 77(2):257-286, 1989.

[13] Sacks, H., Schegloff, E., et Jefferson, G. (1974) *A simplest systematics for the organization of turn-taking for conversation*. Language, 50, 696-735.

[14] Vaufreydaz, D., IST-2000-28323 FAME: Facilitating Agent for Multi-Cultural Exchange, (WP4) European Commission project IST-2000-28323 October 2001.