Evaluation automatique de l'imitation dans l'interaction

Stéphane Michelet Catherine Achard Mohamed Chetouani

Sorbonne Universités, UPMC Univ Paris 06, UMR 7222, ISIR, F-75005, Paris, France

4 Place Jussieu 75005 Paris
{ stephane.michelet, catherine.achard, mohamed.chetouani}@isir.upmc.fr

Résumé

L'imitation entre deux partenaires humains est un des concepts clés du comportement social et de l'interaction. Son évaluation automatique fait partie des verrous liés au développement de robots socialement-adaptés. Des recherches en psychologie ont établi que l'imitation est définie par trois paramètres principaux : l'orientation dans la relation (qui mène l'interaction), le délai entre les partenaires, et le degré d'imitation. Dans ce papier, nous proposons une méthode non supervisée permettant de mesurer l'imitation entre deux partenaires en termes de délai et de degré et ce, en étudiant uniquement des données gestuelles. Dans un premier temps, des points d'intérêts spatio-temporels sont détectés afin de sélectionner les régions les plus importantes des vidéos. Ils sont ensuite décrits à l'aide d'histogrammes pour permettre la construction de modèles bag-of-words dans lesquels l'information spatiale est réintroduite. Le degré d'imitation et le délai entre les partenaires sont alors estimés grâce à une cross-corrélation entre les deux modèles bag-of-words. Les résultats présentés mettent en avant l'efficacité de cette méthode non supervisée pour fournir une mesure continue du degré d'imitation. Des études ont également été menées pour gérer au mieux le compromis entre la durée d'observation et la continuité de la mesure ou pour fixer un délai maximal entre un stimulus et sa réponse.

Mots Clef

Interaction, Imitation, Bag-of-words, Apprentissage non supervisé.

Abstract

Imitation between human partners is a key concept of human social behaviour and interactions. Its automatic evaluation is one challenge of the development of socially adapted robots. Researches in psychology have established that imitation is defined by three main parameters : orientation in the relation (who is leading the interaction), the delay between the partners, and the degree of imitation. In this paper, we propose an unsupervised method permitting the measurement of imitation between two partners in terms of delay and degree just by studying motor imitation. First, spatio-temporal interest points are detected in order to select the most important regions in videos. They are then described using histograms in order to construct bag-of-words models in which spatial information has been reintroduced. Degree of imitation and delay between partners are then estimated thanks to cross-correlation between two bag-of-words models. Results presented highlight the efficiency of this unsupervised method in providing a continuous measurement of the degree of imitation. Studies have also been made in order to address the compromise between duration of observation and continuity of the measure, or to fix a maximum delay between a stimulus and a response.

Keywords

Interaction, Imitation, Mimicry, Bag-of-words, unsupervised learning.

1 Introduction

Les interactions sont des échanges multimodaux dans lesquels des processus dynamiques prennent place [6]. Alors que les interactions ont été étudiées en psychologie pendant des années [14, 3], leur étude automatique et leur modélisation computationnelle font partie des challenges majeurs de ces dernières années [19, 13].

L'étude de la dynamique des communications humaines a suscité un intérêt croissant ces dernières années dans des domaines très variés, comme le traitement du signal social [13, 2], la robotique sociale [16, 1] ou les neurosciences [9]. Par exemple, dans cette dernière catégorie, Dumas et al. [9] ont modélisé et mesuré l'activité cérébrale de deux partenaires en interaction. Ils ont montré que lorsque les partenaires sont engagés dans une interaction, leurs activités cérébrales peuvent être modélisées par des oscillateurs couplés, démontrant ainsi une influence réciproque entre les partenaires. Dans ce cadre de communication humaine, l'imitation joue un rôle important. En effet, Chartrand et al. ont établi un lien entre le degré de mimétisme, la perception de la fluidité de l'interaction et le degré d'appréciation de l'interaction entre les partenaires [5].

Les interactions et l'imitation en particulier font intervenir plusieurs modalités, de manière plus ou moins forte. Citons à titre d'exemple la posture du corps, les hochements de tête, le regard, la prosodie, les tours de paroles, ...
Dans ce papier, nous nous concentrons sur l'imitation gestuelle qui permet de mieux comprendre la coordination des actions entre les partenaires, à la fois temporellement et en terme de forme [3], tâche qui constitue un verrou pour le développement de robots socialement adaptés.
Quelques travaux récents de la littérature détaillent des méthodes permettant d'accéder à l'imitation. Ainsi, Sun et al. [18] calculent la corrélation entre des séries de caractéristiques visuelles et en déduisent un score d'imitation global. Les caractéristiques encodées représentent des quantités de mouvement, ce qui mène plus à une mesure de synchronie qu'à une mesure d'imitation. Feese et al. [10] quant à eux s'intéressent au leadership et calculent des co-occurrences basées sur des évènements discrets. Dans [4], Bilakhia et al. apprennent des modèles de régression à partir des caractéristiques du premier sujet afin de prédire les caractéristiques du second, et utilisent l'erreur de reconstruction pour mesurer l'imitation. Les caractéristiques visuelles utilisées proviennent des expressions faciales. Dans [20], Xiao et al. utilisent la divergence de Kullback-Leibler entre deux mixtures de gaussienne modélisant les fréquences des mouvements de tête et comparent les scores obtenus à une annotation externe du comportement.
Des recherches en psychologie [11, 12] ont établi que l'imitation est définie par trois paramètres principaux : l'orientation dans la relation (qui mène l'interaction), le délai entre les partenaires et le degré d'imitation. La méthode proposée dans ce papier diffère des approches traditionnelles dans le sens où elle tire parti des études menées en psychologie et permet de mesurer l'imitation entre deux partenaires en termes de délai et de degré et ce, en étudiant uniquement des données gestuelles. Elle diffère également de beaucoup de travaux dans le sens où elle conduit à une mesure continue de l'interaction et ce, dans le but de modéliser la communication sociale de manière fine. La méthode s'appuie sur des modèles Bag-Of-Words (BOW) pour caractériser chaque vidéo puis sur une mesure de similarité pour estimer les paramètres de l'imitation.
Dans la suite de cet article, la section 2 présente la modélisation spatio-temporelle utilisée ainsi que la mesure de similarité choisie. Le protocole expérimental est ensuite décrit dans la section 3, avant les résultats expérimentaux (section 4) et la conclusion (section 5).

2 Mesure de l'imitation avec une modélisation par bag-of-words

2.1 Modélisation spatio-temporelle par bag-of-words

Le modèle Bag-Of-Words (BOW) a été largement utilisé par la communauté afin de réduire les coûts computationnels dans l'analyse d'images ou de vidéos. Il

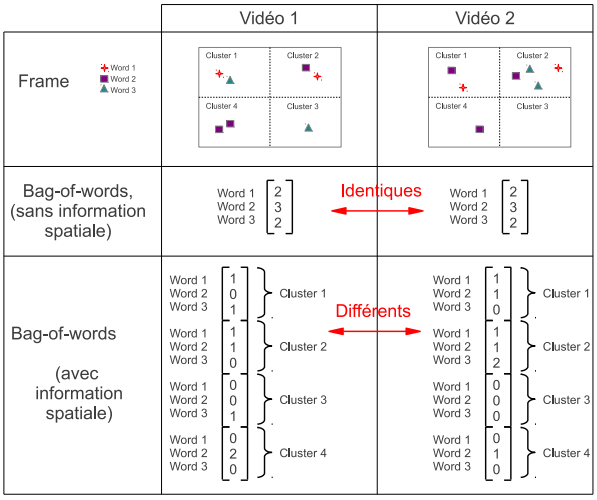

FIGURE 1 – Intérêt du spatial dans les BOW.

consiste à décrire une scène à partir de l'ensemble des mots visuels qui la constitue. Afin de remédier à son principal défaut nous avons cependant réintroduit l'information spatiale dans la modélisation des séquences. Les Points d'Intérêts Spatio-Temporels (STIP) sont extraits dans un premier temps en utilisant l'approche proposée par Dollár et al. [8]. Ils sont ensuite caractérisés avec des Histogrammes de Flot Optique (HOF) [15] estimés sur un volume spatio-temporel centré autour de chaque point (3x3x2), menant à un descripteur de dimension 90. Comme l'information spatiale est indispensable à l'étude de l'imitation gestuelle (cf figure 1), elle est réintroduite en construisant deux dictionnaires séparés. Le premier dictionnaire porte seulement sur la position des points. Ses k représentants obtenus à partir de l'algorithme des k-means sont des centres spatiaux ($Cluster1$ à $Cluster4$ sur la figure 1) qui correspondent aux zones spatiales contenant le plus d'information. Afin de détecter l'imitation quelle que soit la position du participant dans l'image, toutes les coordonnées sont exprimées dans un espace centré sur le visage de l'utilisateur. Les visages sont détectés dans chaque image en utilisant un Détecteur de Visage appelé IntraFace [21]. Un filtre de Kalman est ensuite appliqué afin de stabiliser les résultats et palier aux rares observations manquantes. Le second dictionnaire est appris en considérant seulement les descripteurs visuels et amène à des mots visuels ($Word1$ à $Word3$ dans la figure 1). Le vecteur final composant le BOW est la combinaison des histogrammes des mots visuels estimés sur les différents clusters spatiaux, comme illustré sur la figure 1. Chaque point va contribuer à toutes les composantes du BOW en fonction (inversement proportionnel) de sa distance aux centres spatiaux et visuels.

Une fois les vidéos décrites à l'aide d'un BOW spatio-temporel, une mesure de similarité est utilisée pour les comparer.

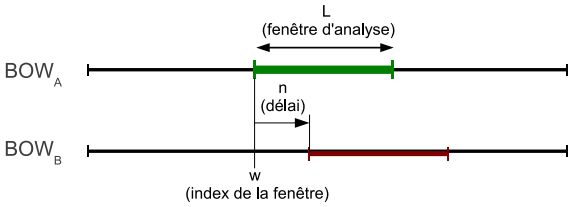

FIGURE 2 – Illustration des variables utilisées.

2.2 Mesure de similarité

La cross-corrélation est une des méthodes les plus simples et les plus utilisées pour comparer deux séries temporelles. Pour estimer localement, au temps k, la corrélation entre deux signaux, deux paramètres doivent être considérés : la durée L de la fenêtre temporelle pendant laquelle les signaux sont observés et le décalage temporel (délai) n entre ces signaux, comme illustré sur la figure 2. Si $W_A(k) et W_B(k)$ représentent les BOW caractérisant les séquences A et B, la corrélation estimée avec un délai n et une fenêtre d'observation de longueur L est donnée par :

$$[W_A \star W_B]_n(k) = \sum_{m=k}^{k+L} \frac{W_A(m)^T . W_B(m+n)}{||W_A(m)||_2 * ||W_B(m+n)||_2} \quad (1)$$

où $||x||_2$ est la norme L2 de x.
Comme dans le cadre d'une interaction naturelle, le délai n est inconnu et comme de plus il varie dans le temps, les corrélations sont estimées pour plusieurs délais n et le degré de l'imitation est estimé en chaque temps k avec :

$$degré(k) = \max_{n \in [-n_{max}, \ldots, +n_{max}]} ([W_A \star W_B]_n(k)) \quad (2)$$

D'autres métriques auraient pu être utilisées pour comparer les séquences et obtenir une mesure d'imitation, comme par exemple le Dynamic Time Warping (DTW). Cependant, le temps de calcul de la cross-corrélation est très inférieur à celui du DTW. De plus, tandis que la cross-corrélation permet de remonter au délai entre les partenaires (donnée pertinente dans le cadre d'une interaction), celui-ci est beaucoup plus difficilement accessible avec le DTW (ce délai est variable au sein même de la fenêtre d'interaction avec le DTW). Ces raisons justifient le choix de l'utilisation de la cross-correlation.

3 Base de données

A part les travaux récents de Sun et al. [17], nous n'avons pas trouvé dans la littérature de base de données publiques et annotées permettant d'étudier le mimétisme. Or cette base, plus dédiée à l'analyse des expressions faciales, ne présente que très peu de gestes et de mouvement. Ceci nous a amené à créer une nouvelle base qui pourra aider les chercheurs dans leur travail pour développer de nouveaux

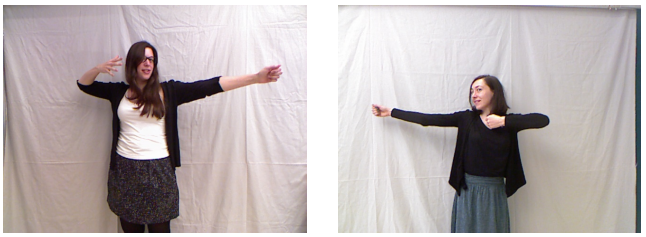

FIGURE 3 – Participant 1 (à gauche) faisant une action et participant 2 (à droite) l'imitant en miroir.

algorithmes, en sautant les phases de collecte de données et d'annotation. Cette base comprend des séquences d'imitation dyadique, sans gestes prédéfini. Ainsi, nous avons demandé à deux participants, se tenant face à face, à une distance d'environ 2 mètres, de jouer à un jeu d'imitation : le meneur fait des gestes de manière libre tandis que son partenaire doit l'imiter en miroir (figure 3). Les gestes effectués n'ont pas été appris, ne contiennent aucune sémantique et ne sont pas segmentés dans le sens où ils s'enchainent de manière libre, sans passer par une étape de repos. Le caractère non contraint de cette base nous a naturellement conduit à mettre en place une approche non-supervisée.
Chacune des deux personnes est filmée à une vitesse de 25 frames par seconde et les deux vidéos sont synchronisées. La base est composée de 9 dyades et chaque dyade effectue deux sessions, pour lesquelles chaque participant a été meneur puis imitateur. Ainsi un total de 18 interactions durant environ 2 minutes ont été enregistrées. Pour finir, seules les 2700 premières images (108 secondes) de chaque vidéo ont été gardées afin de ne pas privilégier certaines vidéos plus longues, ce qui aurait faussé les résultats.
Les 18 participants étaient volontaires, et comptaient 8 femmes et 10 hommes en équipes mixtes. La plupart d'entre eux (16) ont accepté la publication de leur image. L'âge moyen est de 22,9 ans.

4 Résultats

Pour tous les résultats, les dictionnaires de mots visuels et de centres spatiaux ont été appris sur une base externe, présentée dans [7]. Cette base est composée de 10 minutes de vidéos de vue frontale de participants faisant des gestes, comme illustré sur la figure 4. Bien que le point de vue soit comparable à la base de données présentée dans la section 3, la luminosité et le protocole expérimental sont cependant différents.
Une fois les deux dictionnaires appris, le processus complet de mesure de l'imitation a été appliqué sur la base présentée section 3. Les vidéos d'imitation correspondent aux vidéos couplées de la base. Pour les vidéos de "non-imitation", nous avons repris l'idée présentée par Bernieri dans [3], où des vidéos de pseudo interactions sont créées en mélangeant les dyades des vidéos d'interaction. Ainsi

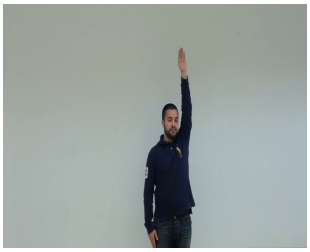

FIGURE 4 – Base de données externe, proposée dans [7].

nous disposons d'un total de 18 vidéos d'imitation et de 153 vidéos de non-imitation.
Pour chaque couple de vidéos (imitation ou non), le processus est appliqué et donne à chaque instant un degré d'imitation qui est seuillé pour décider de l'imitation ou la non-imitation. Une matrice de confusion est alors calculée, en supposant que l'imitation est présente à chaque instant dans les vidéos d'imitation, et est toujours absente dans les vidéos de non-imitation. Cette hypothèse est cependant abusive puisqu'il existe de l'imitation fortuite dans les vidéos de non-imitation, ou que l'imitation n'est pas forcément totale dans les vidéos d'imitation.
Grâce à la matrice de confusion il est possible de calculer les taux de vrai positif et faux positif et ainsi de tracer la courbe ROC (Receiver Operating Characteristic). L'aire sous la courbe (AUC) est un bon indice d'efficacité du processus qui nous permettra de valider l'approche proposée en introduisant dans un premier temps l'influence du nombre de clusters spatiaux et visuels sur la mesure. Dans un second temps, nous étudierons et interpréterons les paramètres caractérisant l'imitation comme la durée de la fenêtre d'analyse et le délai maximal autorisé.

4.1 Influence du nombre de mots visuels et de clusters spatiaux

Dans un premier temps, nous étudions la contribution des mots spatiaux et visuels sur la mesure d'imitation. Pour cela la durée de la fenêtre d'analyse et le délai maximum ont été choisis comme un compromis entre le temps de calcul et la validité de la mesure d'interaction. Ainsi ils ont été fixés respectivement à 5 secondes et 10 secondes. La taille des deux dictionnaires va de 1 (ce qui revient à ne pas utiliser le dictionnaire) à 128, comme illustré figure 5 où l'aire sous la courbe ROC est présentée en fonction du nombre de mots.
Nous pouvons remarquer que si les informations spatiales ne sont pas utilisées (première ligne du tableau), les résultats restent relativement faibles à moins que le nombre de mots n'augmente de manière drastique (128). De plus, un nombre faible de catégories spatiales (4) suffit à améliorer les résultats de manière notable.
D'un autre côté, on pourrait penser que la seule localisation spatiale des points pourrait suffire à mesurer l'imitation entre les partenaires. Cependant, la première colonne de la figure 5 montre que le spatial seul a des résultats faibles

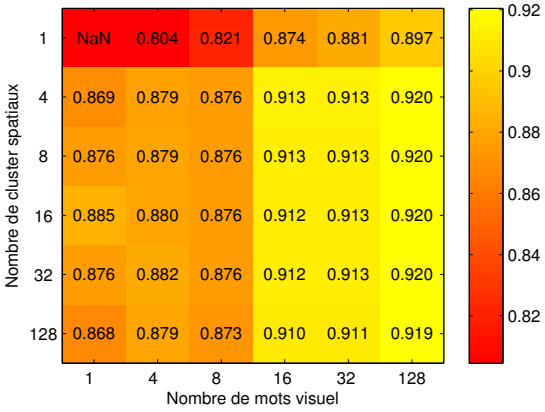

FIGURE 5 – AUC en fonction des nombres de mots visuels et spatiaux.

et donc, que les deux types d'information sont nécessaire.
Pour conclure, la combinaison de 4 clusters spatiaux avec 16 mots visuels marque un tournant dans les résultats, l'AUC passant de 0,88 à 0,91. L'augmentation des dictionnaires ne permet ensuite que des améliorations mineures (< 0.01) par rapport à la complexité engendrée. Ainsi, ces deux dictionnaires seront conservés dans la suite.

4.2 Délai maximum et durée de la fenêtre d'analyse

Deux paramètres utilisés pour calculer la corrélation et présentés dans la section 2.2 influencent les résultats et ont également une signification physique dans l'interaction. Il s'agit de la durée de la fenêtre d'étude et du décalage temporel entre les partenaires. Concernant la fenêtre d'étude, la question posée est la suivante : combien de temps faut-il observer une interaction pour décider s'il y a ou non imitation entre les partenaires ? Il parait évident que plus l'observation sera longue, plus fiable sera la mesure. Cependant le but est d'obtenir une mesure locale et instantannée de l'imitation, ce qui suppose une fenêtre temporelle la plus courte possible. Ainsi, un compromis doit être réalisé. Le second paramètre est le délai temporel maximal entre les partenaires. Là aussi, si un délai trop court est rédhibitoire car les personnes mettent un certain temps à réagir dans des interactions naturelles, un délai trop long n'est pas non plus envisageable : si deux personnes font le même geste à 20 secondes d'écart, peut-on considérer que c'est de l'imitation et qu'il s'agit d'une réponse à un stimulus ? Là aussi, un compromis doit être trouvé. Bien souvent, les psychologues [11] utilisent un délai ad hoc de 3 secondes, mais est-ce optimal ? Afin d'optimiser ces paramètres, mais aussi d'étudier leur importance, nous faisons varier le délai maximum autorisé entre 0 et 40 secondes et la durée de la fenêtre d'analyse entre 5 et 100 secondes. Les résultats (AUC) sont présentés figure 6.
Le premier résultat, auquel on pouvait s'attendre, est que

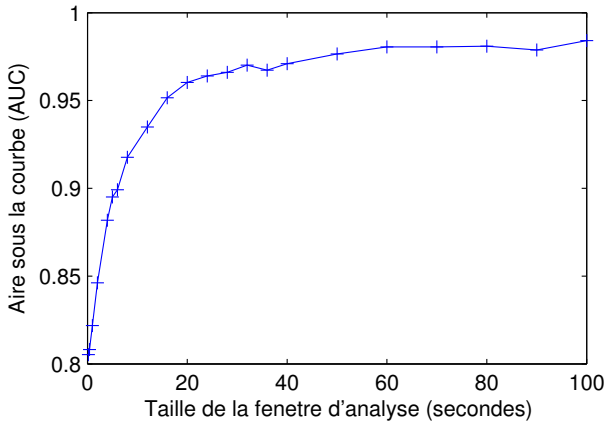

FIGURE 7 – AUC en fonction de la taille de la fenêtre d'analyse, pour un délai fixé à 15 frames.

plus la fenêtre d'analyse est longue, meilleur le score est. Le second résultat est l'existence d'un délai optimal, indépendamment de la durée de la fenêtre d'étude. Dans notre cas ce délai de 15 frames, correspond à 0,6 seconde de décalage temporel entre les deux partenaires. Ce temps est court relativement aux 3 secondes habituellement utilisées par les psychologues dans l'étude des interactions naturelles [11]. Cependant ce résultat est cohérent avec la construction de notre base : il était demandé explicitement aux participants de s'imiter, ce qui se fait dans des temps très courts. Conformément à nos attentes, l'augmentation du délai maximum autorisé diminue les résultats. Cet effet prévisible est dû à la flexibilité introduite qui permet, pour les données artificielles, de trouver de l'imitation là où il n'y en a pas, alors que dans le cas des vidéos d'imitation, le délai optimal situé autour de 15 frames est trouvé.
La figure 7 montre l'évolution de l'AUC en fonction de la taille de la fenêtre d'analyse pour un délai maximum égal au délai optimal de 15 frames.
Comme nous l'avions déjà constaté, l'AUC augmente avec la taille de la fenêtre. Cependant, au-delà de 750 frames (30 secondes), cette augmentation est faible. Ainsi, il n'est pas nécessaire d'observer l'interaction sur une très longue durée pour détecter l'imitation de manière fiable. Le choix de la longueur de la fenêtre d'étude, qui est un compromis entre la qualité et la continuité de la mesure d'imitation, peut maintenant être fait en fonction de l'application, en connaissance de cause. Par exemple, une fenêtre de 10 secondes permet déjà d'avoir une bonne mesure de l'imitation ($AUC > 0.92$) tandis qu'une une fenêtre de 20 secondes amène à de meilleurs résultats ($AUC > 0.96$), proches du maximum ($AUC = 0.98$), mais moins de réactivité.

5 Conclusion

Dans ce papier, quatre contributions majeures ont été apportées. Tout d'abord, face au manque de base de données d'imitation motrice dans la littérature, nous avons introduit une nouvelle base d'imitation dyadique gestuelle. Celle-ci est composée de 9 dyades réalisant une imitation gestuelle libre à tour de rôle (meneur/imitateur), ce qui conduit à 18 séquences d'imitation. Des pseudo-séquences de non imitation sont ensuite automatiquement générées en mixant les différentes dyades.

La seconde contribution est l'introduction d'une méthode de mesure d'imitation entre deux partenaires, issue des techniques de reconnaissance d'action non supervisée. Cette méthode, qui permet d'estimer le degré d'imitation à chaque instant de la séquence, utilise des modèles bag-of-words dans lesquels l'information spatiale a été réintroduite. Une mesure de cross-corrélation des différents bag-of-words conduit à l'estimation de l'imitation. Contrairement à la plupart des approches proposées dans la littérature, aussi bien la temporalité que la forme des gestes sont pris en compte. De plus, nous accédons à une mesure continue de l'imitation fort utile pour une modélisation fine des interactions.

Dans un troisième temps, nous nous sommes intéressés au délai temporel maximum admissible entre les partenaires : si la réponse à un stimulus ne peut pas être instantanée, nous avons montré que laisser trop de latitude dans le délai détériore les résultats en facilitant la détection d'imitation fortuite. Ainsi, il est possible, pour un scénario, ou un type d'interaction donné, de définir un délai maximum optimal.

La dernière contribution répond à la question de temporalité : pendant combien de temps est-il nécessaire d'observer une interaction pour décider de la présence ou non d'imitation ? Nous avons ainsi confirmé notre première intuition : plus la fenêtre d'observation est longue, plus la mesure est fiable. Nous avons également montré que passé 20 secondes, les améliorations sont minimes et que dès 10 secondes, des mesures relativement fiables sont obtenues. Le choix final devra être réalisé en fonction de l'application et du compromis réactivité du système versus qualité de la mesure.

Les futurs challenges qui se présentent à nous sont d'étudier plus en profondeur un autre paramètre important de l'imitation qui est l'orientation dans la relation. De plus, la méthode proposée devra être réévaluée sur une base de données plus naturelle, dans laquelle l'imitation est plus diffuse et moins présente. Enfin, ces travaux sur l'imitation vont être étendus à des interactions naturelles et ainsi évoluer vers des mesures de synchronie.

Remerciement

Ce travail a été effectué dans le cadre du Labex SMART et a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir portant la référence ANR-11-IDEX-0004-02 ainsi que dans le cadre de l'ANR SYNED-PSY (ANR-12-SAMA-06) du programme Santé Mentale et Addictions.

délai (secondes)	fenêtre d'analyse (secondes)																						
	0.2	0.4	1	2	4	5	6	8	12	16	20	24	28	32	36	40	50	60	70	80	90	100
0	0.776	0.780	0.795	0.823	0.859	0.875	0.882	0.897	0.920	0.935	0.946	0.951	0.951	0.960	0.965	0.959	0.971	0.971	0.972	0.977	0.978	0.981
0.2	0.788	0.792	0.807	0.834	0.870	0.888	0.891	0.908	0.928	0.945	0.955	0.958	0.959	0.966	0.963	0.967	0.975	0.977	0.978	0.979	0.979	0.983
0.4	0.799	0.802	0.816	0.843	0.879	0.895	0.898	0.916	0.934	0.950	0.959	0.963	0.965	0.969	0.966	0.970	0.977	0.979	0.979	0.980	0.978	0.985
0.6	0.805	0.808	0.822	0.846	0.882	0.895	0.899	0.918	0.935	0.952	0.960	0.964	0.966	0.970	0.967	0.971	0.977	0.981	0.981	0.981	0.979	0.984
0.8	0.807	0.809	0.822	0.845	0.881	0.893	0.898	0.917	0.934	0.951	0.960	0.963	0.966	0.969	0.966	0.970	0.976	0.980	0.981	0.981	0.979	0.983
1	0.805	0.807	0.820	0.842	0.879	0.891	0.896	0.915	0.932	0.950	0.959	0.963	0.964	0.968	0.966	0.970	0.976	0.980	0.981	0.981	0.979	0.983
1.2	0.802	0.803	0.816	0.839	0.876	0.888	0.894	0.913	0.930	0.948	0.958	0.962	0.963	0.967	0.965	0.969	0.976	0.979	0.980	0.980	0.978	0.983
2	0.785	0.787	0.801	0.827	0.863	0.875	0.887	0.906	0.924	0.941	0.953	0.959	0.960	0.964	0.963	0.967	0.973	0.977	0.977	0.977	0.975	0.981
4	0.752	0.754	0.770	0.797	0.839	0.852	0.865	0.884	0.912	0.929	0.942	0.946	0.950	0.957	0.956	0.963	0.970	0.973	0.974	0.977	0.973	NaN
8	0.712	0.715	0.731	0.761	0.806	0.821	0.833	0.854	0.886	0.908	0.921	0.931	0.937	0.942	0.946	0.950	0.956	0.962	0.967	0.969	0.973	NaN
20	0.668	0.671	0.687	0.717	0.762	0.781	0.796	0.822	0.857	0.880	0.891	0.901	0.909	0.914	0.920	0.927	0.944	0.957	NaN	NaN	NaN	NaN
40	0.622	0.624	0.635	0.666	0.709	0.731	0.751	0.777	0.805	0.833	0.851	0.858	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

FIGURE 6 – AUC en fonction du délai maximum autorisé entre les partenaires et de la longueur de la fenêtre d'analyse.

Références

[1] S. Al Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, A. Mahdhaoui, J. C. Martin, S. Ondas, C. Pelachaud, J. Urbain, and M. Yilmaz. Generating Robot/Agent backchannels during a storytelling experiment. *ICRA'09*, pages 3749–3754, 2009.

[2] Tobias Baur, Ionut Damian, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. NovA : automated analysis of nonverbal signals in social interactions. In *Human Behavior Understanding*, pages 160–171. 2013.

[3] F. J. Bernieri, J. S. Reznick, and R. Rosenthal. Synchrony, pseudo synchrony, and dissynchrony : Measuring the entrainment process in mother-infant interactions. *Journal of Personality and Social Psychology*, Vol. 54(2) :243–253, 1988.

[4] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual detection of behavioural mimicry. In *ACII 2013*, pages 123–128, 2013.

[5] Tanya L Chartrand and John A Bargh. The chameleon effect : The perception–behavior link and social interaction. *Journal of personality and social psychology*, Vol. 76(6) :893, 1999.

[6] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal synchrony : A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, Vol. 3(3) :349–365, 2012.

[7] Emilie Delaherche, Sofiane Boucenna, Koby Karp, Stéphane Michelet, Catherine Achard, and Mohamed Chetouani. Social coordination assessment : Distinguishing between shape and timing. In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, pages 9–18. 2013.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005*, pages 65–72, 2005.

[9] Guillaume Dumas, Jacqueline Nadel, Robert Soussignan, Jacques Martinerie, and Line Garnero. Inter-brain synchronization during social interaction. *PLoS ONE*, Vol. 5(8) :e12166, 2010.

[10] Sebastian Feese, Bert Arnrich, Gerhard Tröster, Bertolt Meyer, and Klaus Jonas. Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion. In *PASSAT 2012 and SocialCom 2012*, pages 520–525, 2012.

[11] Ruth Feldman. Infant–mother and infant–father synchrony : The coregulation of positive arousal. *Infant Mental Health Journal*, Vol. 24 :1–23, 2003.

[12] Ruth Feldman. Parent–infant synchrony and the construction of shared timing ; physiological precursors, developmental outcomes, and risk conditions. *Journal of Child Psychology and Psychiatry*, Vol. 48(3-4) :329–354, 2007.

[13] Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups : A review. *Image and Vision Computing*, Vol. 27(12) :1775–1787, 2009.

[14] Marianne LaFrance. Nonverbal synchrony and rapport : Analysis by the cross-lag panel technique. *Social Psychology Quarterly*, 42 :66–70, 1979.

[15] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *ICCV*, pages 432–439, 2003.

[16] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. Predicting listener backchannels : A probabilistic multimodal approach. In *Intelligent Virtual Agents*, pages 176–190. 2008.

[17] Xiaofan Sun, Jeroen Lichtenauer, Michel Valstar, Anton Nijholt, and Maja Pantic. A multimodal database for mimicry analysis. In *Affective Computing and Intelligent Interaction*, pages 367–376, 2011.

[18] Xiaofan Sun, Anton Nijholt, Khiet P. Truong, and Maja Pantic. Automatic visual mimicry expression analysis in interpersonal interaction. In *CVPRW 2011*, pages 40–46, 2011.

[19] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. Bridging the gap between social animal and unsocial machine : A survey of social signal processing. *IEEE Transactions on Affective Computing*, Vol. 3(1) :69–87, 2012.

[20] Bo Xiao, Panayiotis G. Georgiou, Chi-Chun Lee, Brian Baucom, and Shrikanth S. Narayanan. Head motion synchrony and its correlation to affectivity in dyadic interactions. In *ICME 2013*, pages 1–6, 2013.

[21] Xuehan Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR 2013*, pages 532–539, 2013.