A New Computational Approach to Identify Human Social intention in Action

Mohamed Daoudi, Yann Coello, Paul Audain Desrosiers, Laurent Laurent Ott Ott

▶ **To cite this version:**

Mohamed Daoudi, Yann Coello, Paul Audain Desrosiers, Laurent Laurent Ott Ott. A New Computational Approach to Identify Human Social intention in Action. IEEE International Conference on AUTOMATIC FACE AND GESTURE RECOGNITION (FG 2018) , May 2018, Xi'an, China. hal-01692111v1

HAL Id: hal-01692111

https://hal.science/hal-01692111v1

Submitted on 24 Jan 2018 (v1), last revised 23 Mar 2018 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A New Computational Approach to Identify Human Social intention in Action

Mohamed Daoudi[1], Yann Coello[2], Paul Audain Desrosiers[2] and Laurent Ott[2]
[1] IMT Lille Douai, Univ. Lille, CNRS, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
[2] Univ. Lille, CNRS, UMR 9193 SCALab

***Abstract*— In this paper, we propose to analyze the trajectories of the human arm to predict social intention (personal or social intention). The trajectories of different 3D markers acquired by Mocap system, are defined in shape spaces of open curves. The results obtained in the experiments on a new dataset show an average recognition of about 68% for the proposed method, which is comparable with the average score produced by human evaluation. The experimental results show also that the classification rate could be used to improve social communication between human and virtual agents. To the best of our knowledge, this is the first paper which uses computer vision techniques to analyze the effect of social intention on motor action for improving the social communication between human and avatar.**

I. Introduction

Understanding what a conspecific is doing (for instance recognising other's action) is crucial for the control of our everyday social interactions. Understanding the reasons that drive the observed behaviour (for instance identifying other's intention) is however much more complicated. Previous literature has shown than when we observe a confederate grasping a bottle of water, we can anticipate whether the bottle is grasped to drink from it or to throw it away, based on variation of arm movement kinematics (e.g. [7]). But, what happens if the spatial constraints of the task are similar (same object, same location, same movement), but only the social intention changes (for example, moving the bottle on a table for a subsequent movement performed by either the actor or someone else). Recent researches in cognitive psychology have suggested that even in that case movement kinematics is affected. In particular, these studies showed that when we perform an action with a social instead of a personal intention, we amplify the spatial and temporal parameters of the motor action [11]. Furthermore, an observer is able to perceive these kinematic deviants and anticipate social intention in motor actions performed by others, in order to act in a complementary way ([9]; see [10] for a review). A challenge for the future is to determine whether these kinematic deviants can be registered and processed by a classifier in order to allow virtual and artificial systems (avatars, robots ...) to distinguish between different social goals in interactive context, including either humans or other artificial systems. The challenge is how human social intention can be predicted from the actions.

Action and activity recognition is one of most active areas in computer vision. However, very few attempts have addressed the issue of intention from motion. Zunino et al. [13] propose an approach of intention prediction from motion based on covariance-based representations. The use of covariance matrices is extended to the case of temporal sequences of 3D joints, by proposing an approach to human action recognition from 3D skeleton sequences extracted from depth data. Relationship between joint movement and time is captured using multiple covariance matrices over sub-sequences and organizing them hierarchically [4]. This paper focusses however more on motor intention than social intention, which is the aim of our study.

In summary, the main contributions of this paper are:

- We introduce a new problem of social intention from motion.
- We propose a new 3D dataset. To the best of our knowledge this is the first dataset designed for social intention from motion.
- We propose a new classification algorithm which discriminates between social intentions and improve the communication between human and avatar.

This paper proposes a new approach to understand the social intention vs personal intention, and the behavior of people in real life. This work is conducted with a closest collaboration with cognitive scientist experts, and researchers in computer science. An overview of the method is drawn in Fig. 1. The rest of the paper is organized as follows. The section II describes the proposed method of the shape analysis trajectories. The section III presents the shape analysis framework of the landmarks motion. The section IV describes the dataset and the result obtained with the proposed approach. We will conclude in section V.

II. Shape Analysis of Trajectories of Human Social intention

Our hypothesis was that we can characterize the social intention using image data by studying the shape of markers trajectories corresponding to the human movements. This requires proper mathematical representations of these trajectories and statistical models for studying their variability. In the last few years, many approaches have been developed to analyze shapes of 2D curves [8]. We can cite approaches based on Fourier descriptors, moments or the median axis.

This work has been partially supported by PIA, ANR (grant ANR-11-EQPX-0023), European Founds for the Regional Development (Grant FEDER-Presage 41779).

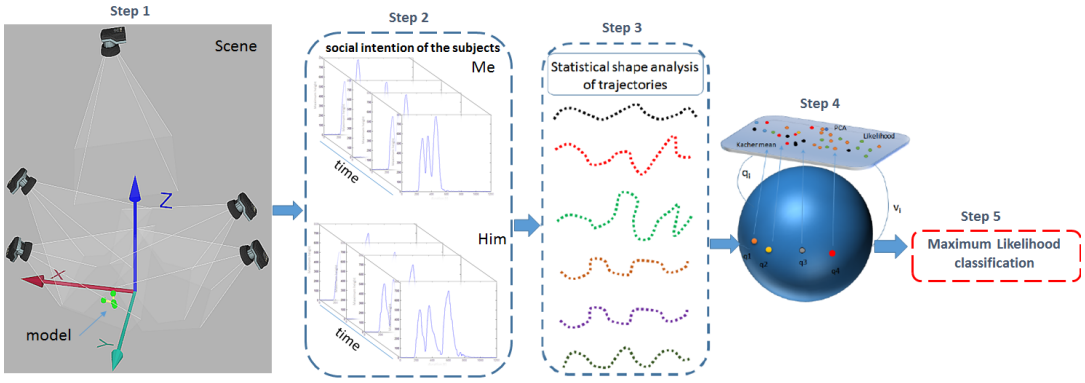

Fig. 1. Overview of the proposed method

More recent works in this area consider a formal definition of shape spaces as a Riemannian manifold on which they can use the classic tools for statistical analysis on tangent spaces. Motivated by the promising results obtained in 3D facial recognition [3], human action recognition [2] and 3D facial expression recognition [1], we propose to use the shape analysis framework proposed by [12]. Each motion of markers is represented by a single trajectory. Formally, we start by considering a given trajectory as a continuous parameterized function $\beta(t) \in \mathbb{R}^3$, $t \in [0,1]$. β is first represented by its *Square-Root Velocity Function* (SRVF), q, according to :

$$q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}, t \in [0,1].$$

Then, with the $\mathbb{L}^2$-norm of the q functions scaled to 1 ($\|q\| = 1$), the space of such representation: $\mathcal{C} = \{q : [0,1] \to \mathbb{R}^3, \|q\| = 1\}$ becomes a Riemannian manifold and have a spherical structure in the Hilbert space $\mathbb{L}^2([0,1], \mathbb{R}^3)$. Given two curves β_1 and β_2 represented by their SRVFs q_1 and q_2 on the manifold, the geodesic path connecting q_1, q_2 is given analytically by the minor arc of the great circle connecting them on $\mathcal{C}$ (see [12] for further details).

It has been proved in [12] that under the $\mathbb{L}^2$-norm, the quantities $\|q_1 - q_2\|$ and $\|q_1 \circ \gamma - q_2 \circ \gamma\|$ are same, where the composition $(q \circ \gamma)$ denotes the function q with a new parameterization dictated by a non-linear function $\gamma : [0,1] \to [0,1]$. This important property allows curves registration by re-parameterization, and thus makes the curves registration easier. In fact, it allows to consider one of the curves as reference and search for a $\gamma^* = \text{argmin}_{\gamma \in \Gamma}(\|q_1 - \sqrt{\dot{\gamma}} q_2 \circ \gamma\|)$ which optimally registers the two curves. This optimization is resolved by Dynamic Programming, as described in [12]. The distance between two elements q_1 and q_2 is defined as $d_{\mathcal{C}}(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle)$. Such distance represents the similarity between the shape of two curves in $\mathbb{R}^3$. Basically, it quantifies the amount of deformation between two shapes. This distance called also elastic distance, is invariant to rotation, scaling and it takes into account the stretching and the bending of the curves [12].

III. Statistical Shape Analysis of Trajectories

The main goal of our study is to categorize the user intention among two classes c_k denote $\{personal, social\}$. For that we propose to learn representative distributions of trajectories for each class.

Since manifolds lack a vector space structure and other Euclidean structures such as norm and inner product, machine learning algorithms including principal component analysis (PCA) and Maximum Likelihood clustering algorithm cannot be applied in their original forms on the manifold $\mathcal{C}$. A common approach used to cope with its non-linearity consists in approximating the manifold valued data with its projection to a tangent space at a particular point on the manifold, for example, the mean of the data μ. Then, each sample shape q_i is mapped in the tangent space at the mean shape $T_\mu \mathcal{S}$ using the inverse exponential map [5] defined as:

$$v_i = exp_\mu^{-1}(q_i) = \frac{\theta}{sin\theta}(q_i - cos(\theta)\mu) \ , \qquad (1)$$

where $\theta = d_{\mathcal{C}}(\mu, q_i)$. The original shape q_i can be retrieved from the velocity vector v_i by using the exponential map operator [5] defined as:

$$q_i = exp_\mu(v_i) = cos(\|v_i\|)\mu + sin(\|v_i\|)\frac{v_i}{\|v_i\|}. \qquad (2)$$

Shapes are projected in the tangent space of the mean using the inverse exponential map (eq. 3).

$$v_2^* = exp_{q_1}^{-1}(q_2^*) = \frac{\theta}{sin\theta}(q_2^* - cos(\theta)q_1) \qquad (3)$$

Such tangent space is a linear vector space which is more convenient to compute statistics. Hence, in order to learn the distribution of tangent vectors on the tangent space, we can first perform PCA to learn a principal subspace denoted $\mathcal{B}$. Then, the covariance matrix on this principal basis is computed as $\Sigma = \sum_{i=1}^{N} v_i v_i^T$, where v_i are the tangent vectors projected into the principal subspace $\mathcal{B}$.

Finally, the multivariate normal distribution of trajectory c_k, $p(v|c_k; \Sigma)$ is learned using the covariance matrix Σ computed from the set of v_i where $|\Sigma|$ is the determinant of the covariance matrix Σ.

$$p(v|c_k; \Sigma) = \frac{1}{(2\pi)^{n/2} \ |\Sigma|^{1/2}} e^{-\frac{1}{2} v^T \Sigma^{-1} v} \qquad (4)$$

In addition, the previous learned distributions can be employed so as to generate random trajectory shapes representing random trajectories.

IV. Experimentations and results

This experimentation contains 3 parts: a) data acquisitions and a learning step; b) classification; c) Kinematic analysis of the evolution of subjects to interact with the avatar.

A. Data collection

It is important to remind here, that the main goal is to propose a new method that is based on gesture recognition to anticipate the social intention of a person. In order to successfully drive this study, all the using scripts are writing under Matlab and C/C++. Then the using equipments are :
1) Qualisys motion capture camera (qualisys system). The qualisys system is delivered with a desk computer with 8 GB, a processor Intel core i7-4770k (8 CPUs) at 3.5 GHz. The frequency of those cameras can varies from 100 to 500 Hz.

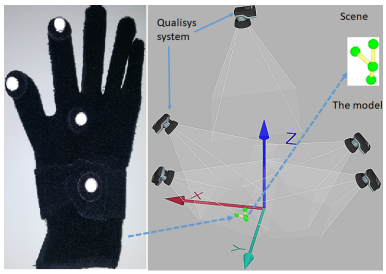

Fig. 2. Left picture shows the black glove that contains 4 infrared reflective markers, the right picture summarizes the whole scene with the emplacement of the motion capture cameras, and the model in green color.

2) A black glove equipped with infrared reflective markers, all those equipments are also provided by qualisys system.
3) A Matlab software (version R2014a) installed on a the desk computer (qualisys system); the Qualisys system provide a specific driver that allow to couple all the Matlab scripts with their system. Thus, it is possible to command all the cameras directly from Matlab for real time analysis. To work with those cameras, it is very important to definite a model like in Fig. 2. The green color means that the model is well known by the system, otherwise the model is in red color and not recognized.

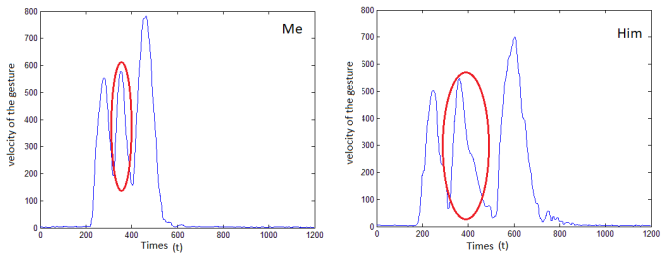

Fig. 3. Left picture shows the gesture of ”Me”, and the right picture the gesture of ”him”. It is clear to observe the difference between the two curves by considering the middle segment or curve.

To evaluate the effectiveness of the method, we collect data from 15 selected subjects with their age varying from 20 to 50 years old. Subjects were invited to seat in front of a table, puts a black glove that contains 4 infrared reflective markers, and a red mug was located at a particular position on the table. We have placed the 4 markers at a specific position on the glove: the index (tip), the thumb (tip), the hand, and the wrist position as shown in Fig. 2.

Following the broadcast of a specific instruction (”Him”, ”Me”), the subject had to move the red mug on the table from the point A to the point B, this 2 positions being visible on the table.
1) When the word ”Me” was broadcasted, the subject moves the red mug on the table from the point A to the point B as unwanted object.
2) When the word ”Him” was broadcasted, the subjects moves the red mug from the point A to the point B with an intention to give it to another person.
In each trial, the participants have to perform the motor action by endorsing either a personal intention (move the glass to a preferred location) or a social intention (move the glass to have it filled by the other person). In each trial, the participants have to comply with the instructions given by the computer. The sequences of actions was randomized.

For a quick analysis of the speed motion, we plot in real time the gestures performed by each subject. Fig. 3 shows the speed motion when the subject displaces the red mug with different intentions. The first bell-shape curve or segment corresponds to the reaching movement where the participant take the mug from a point A, the second segment corresponds to the transport movement where the participant displaces the red mug from point A to point B on the table. And, the last segment is the return to the initial position. The second segment provides more information about the social intention of the subject. The subject repeats the actions with the two intentions 50 times successively (25 times for ”Him”, and 25 times for ”Me”). The whole scene was covered and recorded with 5 Qualisys camera with a frequency of 200 Hz, and a spatial accuracy of less than 0.2 mm. The consuming time for a subject to react at specific sound was set at maximum 6 seconds.

A preprocessing step is performed in order to eliminate any incorrect gestures, trembling, hesitation, bad signal, or missing data. A 3D median filter was used to remove noise like an exaggerate value when the data are lacking. The speed curve of each successful gesture contains 4 minimum and 3 maximum, this is the good condition required for good gesture.

The interesting part in the Fig. 3 is the second bell-shape curve (curve segment) that is in the red circle in the graph of the velocity (Him, Me). This bell-shape curve is very important because it represents the way the participant puts the red mug on the table.

Thus, the trajectories of the different markers are defined as shape spaces and subsequently analysed as a Riemannian manifold for the purposes of characterizing the intention. With the method proposed in III, we obtain a distribution

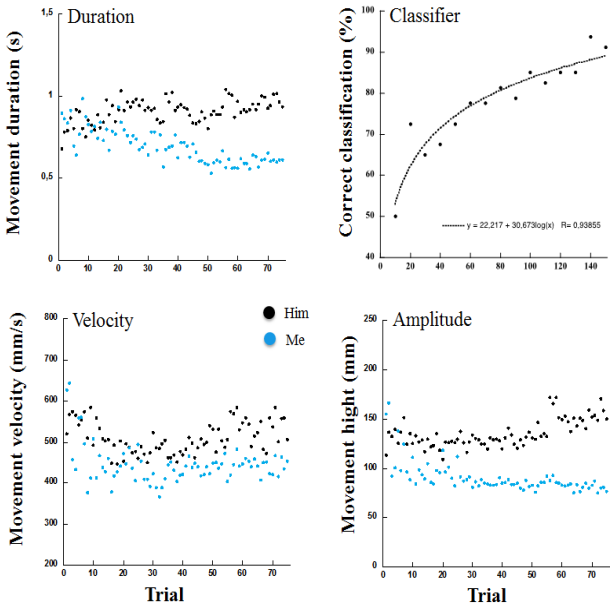

Fig. 4. Performance of the classifier (upper right) with participants (N=8) performing movements with the individual (Him) and social (Me) intention. Variation of movement duration (upper left), movement velocity (lower left), and movement amplitude (lower right) across the 150 trials (75 for each intention) for the movements (object transport phase) performed with the individual (black circle) and social (blue circle) intention.

(classifier) that contains two classes corresponding to the gesture of ”Him” and ”Me”. Let’s talk about the testing classifier.

Fig. 5. Left picture shows the gesture of ”Him”, the avatar looks at the subject. Right picture represents a move of ”Me” because the avatar serves some drink to the subject.

B. Classification

5 subjects are invited in the scene to do the following experiments. To be in the same state as the learning step, We keep the same previous configuration of the scene. An avatar was projected on the TV which simulates a behavior of a barman in his own environment of working. The selected subjects have no information about the experiments. The participant seats in front of the table with the TV, with an animated barman avatar. At a specific broadcast word (Him, Me) like the previous step, the participant moves the red mug from the position A to the position B with either the social or individual intention. The participant should perform the good social intention to trigger the appropriate action by the virtual barman, see Fig. 5. For the gesture of ”Me” the avatar look at you, and for the gesture of ”Him” the avatar serves you some virtual drink. The virtual agent detects the social intention of other natural agents in 68% of cases. This experiments is in real time. This score is comparable with the average score produced by human evaluation. The execution time that includes the reaction of the subject, and the avatar is about 9 seconds. As there is no standard benchmark, it is difficult to compare with other approaches. We will make our dataset available to Face and gesture community.

C. Interaction Human and Avatar

We propose in this step, to help people to be more communicative between them and with a avatar. Thus, 8 new subjects are invited to do the same experiments, in the same condition. In this step, subject repeats the same actions with the two intentions 150 times successively (75 times for ”Me”, and 75 times for ”Him”).

The figure 4 shows the performance of the classifier with participants (N=8) performing movements with the individual ”Me” and social ”Him” intention. Variation of movement duration, movement velocity, and movement amplitude across the 150 trials for the movements performed with the individual and social intentions. As expected by the cognitive scientist experts, this figure helps to understand the difference between the 2 gestures, and the people behavior. The gesture of ”Me” has a lower amplitude, shorter duration, and slower velocity than gesture of ”Him”, which becomes more obvious after several movement rehearsals. The classification step shows the number of trials required for a subject to communicate, and to interact efficiently with the avatar (Fig. 4). The kinematic performances according to the intention (individual, social) start diverging after the rehearsal of 50 trials, which corresponds to a performance of the classifier of 73%. The classification rate increases thus from 50% to 73%, and then reaches close to 100% after 150 movement rehearsals. These experiments show that the subjects are initially less efficient when interacting with an avatar (50% of detection) than with a human (68% of detection), but improve when they receive feedback on their performances using the classifier.

V. Conclusion

In this paper, we propose a geometrical approach to analyze the motion of the human arm in the context of social intention (personal or social intention) prediction. We propose to analyze the motion of different markers as trajectories, and then to analyze them in shape space of curves. Results obtained in the experiments on a new dataset show an average recognition of about 68% for the proposed method, which is comparable with the average score produced by human evaluation [6]. The experimental results show also that the classification rate could be used to improve social communication between human and virtual agents.

REFERENCES

[1] B. Ben Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava. 4-D facial expression recognition by learning geometric deformations. *IEEE Trans. Cybernetics*, 44(12):2443–2457, 2014.

[2] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo. 3D human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Trans. on Cybernetics*, 45(7):1340–1352, 2014.

[3] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama. 3D face recognition under expressions, occlusions, and pose variations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(9):2270–2283, 2013.

[4] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2466–2472. AAAI Press, 2013.

[5] S. Kurtek, A. Srivastava, E. Klassen, and Z. Ding. Statistical modeling of curves using shapes and related features. *Journal of the American Statistical Association*, 107(499):1152–1165, 2012.

[6] D. Lewkowicz, Y. C. F. Quesque, and Y. Delevoye-Turrell. Individual differences in reading social intentions from motor deviants. frontiers in psychology. *Frontiers in Psychology*.

[7] R. Marteniuk, C. MacKenzie, M. Jeannerod, S. Athenes, and C. Dugas. Constraints on human arm movement trajectories. *Canadian Journal of Psychology*, 41:365–378, 1987.

[8] A. E. Oirrak, M. Daoudi, and D. Aboutajdine. Estimation of general 2D affine motion using fourier descriptors. *Pattern Recognition*, 35(1):223–228, 2002.

[9] F. Quesque and Y. Coello. Perceiving what you intend to do from what you do: Evidence for embodiment in social interactions. *Socioaffective Neuroscience and Psychology*, 5:365–378, 2015.

[10] F. Quesque, Y. Delevoye-Turrell, and Y. Coello. Facilitation effect of observed motor deviants in a cooperative motor task: Evidence for direct perception of social intention in action. *Quarterly Journal of Experimental Psychology*.

[11] F. Quesque, D. Lewkowicz, Y. Delevoye-Turrell, and Y. Coello. Effects of social intention on movement kinematics in cooperative actions. *Frontiers in Neurorobotics*, 7(14):1–10, 2013.

[12] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(7):1415–1428, 2011.

[13] A. Zunino, J. Cavazza, A. Koul, A. Cavallo, C. Becchio, and V. Murino. Intention from motion. *CoRR*, abs/1605.09526, 2016.