Towards Automated Analysis of Joint Music Performance in the Orchestra

Giorgio Gnecco[1], Leonardo Badino[2], Antonio Camurri[1], Alessandro D'Ausilio[2], Luciano Fadiga[2], Donald Glowinski[1], Marcello Sanguineti[1], Giovanna Varni[1], and Gualtiero Volpe[1]

[1] DIBRIS Department, University of Genoa, Genoa, Italy
{giorgio.gnecco,antonio.camurri,donald.glowinski,marcello.sanguineti, gualtiero.volpe}@unige.it,giovanna@infomus.org
[2] IIT - Italian Institute of Technology - Genoa, Italy
{Leonardo.Badino,Alessandro.Dausilio,Luciano.Fadiga}@iit.it

Abstract. Preliminary results from a study of expressivity and of non-verbal social signals in small groups of users are presented. Music is selected as experimental test-bed since it is a clear example of interactive and social activity, where affective non-verbal communication plays a fundamental role. In this experiment the orchestra is adopted as a social group characterized by a clear leader (the conductor) of two groups of musicians (the first and second violin sections). It is shown how a reduced set of simple movement features - heads movements - can be sufficient to explain the difference in the behavior of the first violin section between two performance conditions, characterized by different eye contact between the two violin sections and between the first section and the conductor.

Key words: automated analysis of non-verbal behavior, expressive gesture analysis, computational models of joint music action.

1 Introduction

Music is a well-known example of interactive and social activity where affective non-verbal communication plays a fundamental role. Several works have already shown how a player can convey expressive intentions by his/her movements.

Among visual features, in this paper we focus on the so-called *ancillary* or *accompanist gestures* [7], i.e., movements of the body of a music player or of a music instrument, which are not directly related to the production of the sound (in contrast to *instrumental* or *effective gestures*, which are directly involved in sound production). For instance, the movements of the heads of string players during a performance are ancillary gestures, whereas the movements of their bows are (mainly) instrumental gestures. Instrumental gestures are obviously informative since, without them, musicians would not be able to express the different musical ideas they want to communicate. Ancillary gestures are informative, too, since often they allow one to recognize different expressive intentions, without looking at the instrumental gestures/listening to the performance.

For instance, Davidson claimed that visual information alone is sufficient to discriminate among performances of the same piece of music played with different expressive intentions (inexpressive, normal and exaggerated) [3], and that the larger the amplitude of the movement, the deep the expressive intention [4]. This finding was also confirmed by other studies, e.g., Castellano et al. investigated the discriminatory power of several movement-related features for the case of a piano player [1], and Palmer et al. showed how the movement made by the *bell* of a clarinet is larger when the player performs more expressive interpretations of the same piece [5]. However, these works focus on a performance by one player only. More recent studies address non-verbal communication in larger musical ensembles such as a string quartet [6] and a section of an orchestra [2].

The present study is aimed at investigating how the behavior of a group of players spontaneously changes, concerning the head ancillary movement, when changing the way it can interact with the rest of the orchestra.

The paper is organized as follows. In Sections 2 and 3, the experimental methodology and data analysis are described, resp.. In Section 4, the obtained results are presented and discussed. Section 5 contains some conclusive remarks.

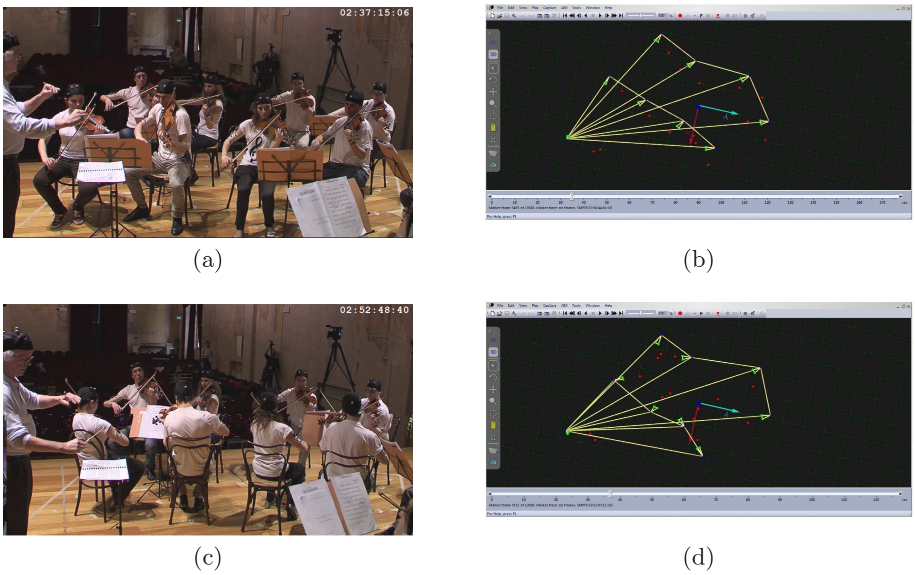
2 Experimental methodology

Two violin sections of an orchestra and two orchestra's professional conductors were involved in the study. Each section counted 4 players and was equipped with passive markers of the Qualisys motion capture system. More specifically, for each player one marker was placed on the head, two markers were placed above the eyes, and one on the nape (back of the neck). For the conductors, one marker was placed on the head. Additional markers were placed on the bows of the players and on the baton of the conductors. Two experimental conditions were tested, which only differ by the way a section (called from here on the *first section*) interact with the conductor and the other section (called from here on the *second section*). In one condition (condition A) the violinists from the first section - disposed in a single row - were able to see the conductor, but not the violinists from the second section. In the second condition (condition B) the violinists from the first section - still disposed in a single row - were able to see the second violin section, but not the conductor (since they were faced backwards with respect to him). Both conditions A and B were experimented with the two conductors. All the other variables (groups of violinists/piece) were the same for the two conditions[1] and each of them was repeated six times (three times with a conductor and three other times with another one). Figure 2 shows the two settings. Each recording consisted of about 1 minute of music excerpts from the Overture to the opera “Il signor Bruschino” by G. Rossini.

Concretely, the study focuses on measuring how much the movements of the heads of the musicians change when moving from condition A to condition

[1] Additional data were collected by varying the piece, but in this paper we present only the results obtained for a fixed piece.

B. Likely, each musician has his/her individual behavior, but the hope behind the experiments (later confirmed by the results) is that a common pattern of behavior (different in each condition) can be extracted.



(a) (b)

(c) (d)

Fig. 1. Panels (a) and (c) show the players and the conductor when condition A and condition B are tested, resp.. Panels (b) and (d) show a snapshot of the head's markers positions of the players and the conductor when condition A and condition B are tested, resp.. Triangles correspond to positions and directions of heads. The red points are the unlabeled markers (mainly associated with the bows).

3 Data Analysis

Movement data were collected by using a Qualisys motion capture system equipped with 7 cameras, integrated with the EyesWeb XMI platform (see `www.eyesweb.org`) for obtaining synchronized multimodal data, including audio and physiological signals (not used in the work described in this paper). A reduced data set, made of 12 recordings, was extracted from the collected data, and movement features due to ancillary gestures were automatically computed.

The remaining of the data analysis has been performed in MATLAB 7.7 by computing the means and covariance matrices of the extracted features in order to find significant differences in such quantities between Conditions A and B.

Choice of the features: the following features were computed in the data analysis. Their computation was made possible by the QTM representation of each marker, which provides its position in each frame.

1. For each musician i ($i = 1$: the conductor, $i = 2, \ldots, 5$: the violinists in the first section, starting from the concertmaster, $i = 6, \ldots, 9$: the violinists in the second section) and each frame j ($j = 1, \ldots, N_{\text{frames}}$) of a same recording[2], we evaluated the current direction $\mathbf{d}_i^{(j)}$ in the horizontal plane of the head of the musician, then the corresponding sample mean direction $\bar{\mathbf{d}}_i$ with respect to all such frames. Each $\mathbf{d}_i^{(j)}$ is defined as the unit vector connecting the marker on the nape of the musician i to the point located in the middle of the line between the two other markers above his/her eyes, whereas $\bar{\mathbf{d}}_i$ is obtained by averaging each component of $\bar{\mathbf{d}}_i^{(j)}$ with respect to j and normalizing the obtained vector.
2. For each musician i and each frame j of a same recording, we evaluated the components $t_i^{(j)}$, $n_i^{(j)}$, and $z_i^{(j)}$ (parallel to $\bar{\mathbf{d}}_i$, orthogonal to $\bar{\mathbf{d}}_i$ in the horizontal plane, and orthogonal to the horizontal plane, resp.) of the position vector $\mathbf{p}_i^{(j)}$ of his/her head. Such a position vector is defined with respect to a fixed Cartesian coordinate system with the origin at the center of the stage (see Figures 1(b) and 1(d)).
3. For each frame, we have defined the vector $\mathbf{a}^{(j)} \in \mathbb{R}^{27}$ with components $t_1^{(j)}$,$n_1^{(j)}$,$z_1^{(j)}$, $t_2^{(j)}$,$n_2^{(j)}$,$z_2^{(j)} \ldots$, $t_9^{(j)}$,$n_9^{(j)}$,$z_9^{(j)}$, then we have computed its sample mean $\bar{\mathbf{a}}$ with respect to the frames, and its sample covariance matrix
$$\texttt{sample cov}(\mathbf{a}) := \frac{1}{N_{\text{frames}} - 1} \sum_{j=1}^{N_{\text{frames}}} \left(\mathbf{a}^{(j)} - \bar{\mathbf{a}}\right) \left(\mathbf{a}^{(j)} - \bar{\mathbf{a}}\right)^T \in \mathbb{R}^{27 \times 27} .$$
Here, $\mathbf{a}$ denotes the random variable and $\mathbf{a}^{(j)}$ its realization.
4. For each musician i and each frame j of the same recording, we computed the oriented angle $\theta_i^{(j)}$ between the two vectors $\mathbf{d}_i$ and $\mathbf{d}_i^{(j)}$. Then, we defined the vector $\mathbf{b}^{(j)} \in \mathbb{R}^9$ with components $\theta_1^{(j)}$, $\theta_2^{(j)}$, $\ldots$, $\theta_9^{(j)}$ and we computed its sample mean $\bar{\mathbf{b}}$ with respect to the frames and its sample covariance matrix
$$\texttt{sample cov}(\mathbf{b}) := \frac{1}{N_{\text{frames}} - 1} \sum_{j=1}^{N_{\text{frames}}} \left(\mathbf{b}^{(j)} - \bar{\mathbf{b}}\right) \left(\mathbf{b}^{(j)} - \bar{\mathbf{b}}\right)^T \in \mathbb{R}^{9 \times 9} .$$
Here, $\mathbf{b}$ denotes the random variable and $\mathbf{b}^{(j)}$ its realization.

Finally, for each of the two conditions, all the sample means and sample covariance matrices were averaged over the six repetitions of the same music piece (three for each conductor). Apart from $z_i^{(j)}$ and the related features, all the

[2] To simplify the notation, we do use indices to distinguish among the three repetitions of the same experimental condition, between the two experimental conditions, and between the two conductors.

other features listed above can be extracted from the projections of the motion-capture data on the horizontal plane only. The reason for which in the definition of the directions of the heads we considered only such projections is that, for each musician, the two frontal markers were positioned much above his/her eyes, so the vertical components of the positions of such markers may be misleading in determining the direction of the head.

Before presenting the results, we describe some guidelines that were used in the data analysis.

– **Choice of the data**: we considered only the three markers associated with the heads of the musicians. Since all the features considered in the following have been calculated at a global scale (i.e., on the entire video, excluding only the frames preceding the performances and some noisy frames, e.g., frames with missing or unlabeled markers associated with the heads) the movement of the baton of the conductor was not taken into account.
– **Missing data**: in case of missing data (e.g., an unlabeled head marker or an undetected one), the corresponding frames were discarded and the features associated with such data were evaluated using only the remaining frames (thus reducing the value of N_{frames}). In particular, features defined as temporal means of certain measurements were computed by summing those measurements over all the available frames (with the exception of the ones containing missing data) and dividing by the number of such frames.
– **Segmentation of the video**: in order to reduce the noise in the data, the first frames in each video (the ones before the beginning of the music piece) were not been considered in the analysis. At least in principle, also the frames in which some musician is turning the page of his/her score should not be considered (or one should not take into account the movements of that musician in such frames only). However, due to the small number of such frames with respect to the total number of frames and the relatively slow movements involved in such “noisy” frames, we did not take into account this issue. So, we performed the analysis on the whole video (excluding only its beginning, which consists of several frames).

4 Results of the analysis

In this section, we show the obtained average values of the features defined in Section 3 for the available recordings, resp. under condition A and under condition B. For simplicity of exposition, instead of considering all the elements of the vectors and matrices defined above, we vary the index i from 2 to 5 (i.e., we present only the values of the features associated with the violinists of the first section). For both conditions A and B, Table 1 and Table 2 show, resp., the average (with respect to six executions of the same piece) sample means of the components of the head positions of the violinists in the first section and the corresponding entries in the covariance matrix of the vector $\mathbf{a}$ of head positions. Similarly, for both conditions A and B, Table 3 and Table 4 show, resp., for the violinists in the first section, the average (with respect to six executions of the same piece) sample means of the oriented angles between the mean head directions and the current head directions and the corresponding entries in the covariance matrix of the vector **b** of oriented angles. Let us make

Avg. sample mean of the feature vector a (cm)												
	t_2	n_2	z_2	t_3	n_3	z_3	t_4	n_4	z_4	t_5	n_5	z_5
Condition A	198.6	8.8	114.6	76.1	70.9	125.0	-14.2	32.4	117.2	-82.6	-40.2	119.8
Condition B	-74.6	-159.3	113.6	-91.7	-55.2	123.3	12.6	-47.1	114.4	75.2	-38.4	117.5

Table 1.

Avg. sample covariance matrix of the feature vector a (cm^2)												
Condition A												
	t_2	n_2	z_2	t_3	n_3	z_3	t_4	n_4	z_4	t_5	n_5	z_5
t_2	5.1	2.9	0.5	-2.9	3.9	-0.2	4.2	-2.1	-0.6	4.1	-2.5	1.1
n_2	2.9	3.9	0.3	-2.0	5.7	-0.2	3.4	0.8	-0.4	2.6	-1.7	1.2
z_2	0.5	0.3	0.6	-0.8	-0.1	-0.0	0.7	-0.6	-0.1	1.7	-1.0	0.4
t_3	-2.9	-2.0	-0.8	9.7	-2.0	0.9	-2.9	1.5	-0.2	-3.7	3.2	-1.5
n_3	3.9	5.7	-0.1	-2.0	24.6	4.3	2.2	2.0	1.0	2.4	-7.3	3.1
z_3	-0.2	-0.2	-0.0	0.9	4.3	3.7	-0.2	-1.3	0.4	1.0	-6.0	1.9
t_4	4.2	3.4	0.7	-2.9	2.2	-0.2	14.3	-2.1	-0.9	4.7	-6.1	2.7
n_4	-2.1	0.8	-0.6	1.5	2.0	-1.3	-2.1	22.7	0.7	-3.9	6.5	-2.0
z_4	-0.6	-0.4	-0.1	-0.2	1.0	0.4	-0.9	0.7	0.8	0.1	-0.8	0.2
t_5	4.1	2.6	1.7	-3.7	2.4	1.0	4.7	-3.9	0.1	18.4	-14.1	5.1
n_5	-2.5	-1.7	-1.0	3.2	-7.3	-6.0	-6.1	6.5	-0.8	-14.1	40.0	-8.7
z_5	1.1	1.2	0.4	-1.5	3.1	1.9	2.7	-2.0	0.2	5.1	-8.7	3.9
Condition B												
	t_2	n_2	z_2	t_3	n_3	z_3	t_4	n_4	z_4	t_5	n_5	z_5
t_2	29.6	1.0	-3.60	0.5	-4.4	-0.6	5.8	9.8	-0.6	-0.8	4.4	-2.7
n_2	1.0	5.8	0.6	-0.8	1.4	0.8	-5.0	0.7	0.6	-1.7	1.6	1.2
z_2	-3.6	0.6	2.1	0.1	2.1	0.4	-3.1	-2.2	1.0	-2.3	-2.8	1.1
t_3	0.5	-0.8	0.1	4.5	1.6	-0.2	-3.1	2.1	1.0	-2.3	-2.1	0.1
n_3	-4.4	1.4	2.1	1.6	9.8	2.2	-7.3	0.3	0.7	-5.3	8.9	0.4
z_3	-0.6	0.8	0.4	-0.2	2.2	1.2	-0.8	0.3	0.2	-1.7	1.8	0.8
t_4	5.8	-5.0	-3.1	-3.1	-7.3	-0.8	36.0	-6.5	-3.7	10.7	-13.3	-1.7
n_4	9.8	0.7	-2.2	2.1	0.3	0.3	-6.5	14.2	-0.1	-2.3	10.5	-2.2
z_4	-0.6	0.6	1.0	1.0	0.7	0.2	-3.7	-0.1	1.9	-2.5	-3.8	2.0
t_5	-0.8	-1.7	-2.3	-2.3	-5.3	-1.7	10.7	-2.3	-2.5	35.3	-12.6	-3.2
n_5	4.4	1.6	-2.8	-2.1	8.9	1.8	-13.3	10.5	-3.8	-12.6	56.3	-6.2
z_5	-2.7	1.2	1.1	0.1	0.4	0.8	-1.7	-2.2	2.0	-3.2	-6.2	5.4

Table 2.

Avg. sample mean of the feature vector b (rad)				
	θ_2	θ_3	θ_4	θ_5
Condition A	0.0012	0.0013	0.0012	0.0007
Condition B	−0.0014	0.0048	0.0019	−0.0273

Table 3.

Avg. sample covariance matrix of the feature vector b (rad^2)				
Condition A				
	θ_2	θ_3	θ_4	θ_5
θ_2	0.0188	−0.0050	0.0103	0.0069
θ_3	−0.0050	0.0695	−0.0051	0.0338
θ_4	0.0103	−0.0051	0.0547	0.0143
θ_5	0.0069	0.0338	0.0143	0.0829
Condition B				
	θ_2	θ_3	θ_4	θ_5
θ_2	0.2156	0.0221	0.0103	−0.0402
θ_3	0.0221	0.1407	−0.0908	−0.0077
θ_4	0.0103	−0.0908	0.2547	0.0425
θ_5	−0.0402	−0.0077	0.0425	0.2680

Table 4.

some comments on the discriminatory power of the chosen features. Of course, as shown by Table 1, the average sample mean of the feature vector **a** for condition A is different from the one for condition B, but this depends only on the slightly different positions of the violinists in the two settings, and - above all - on the different sample mean directions of their heads in the two situations. Similarly, Table 3 basically confirms only that the four vectors $\bar{\mathbf{d}}_i$ $(i = 2, \dots, 5)$ are the sample mean directions of the heads of the four violinists. It is more interesting to compare the average sample covariance matrices of the two feature vectors $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ for each of the conditions. In particular, inspection of the main diagonals of such matrices shows that the average - with respect to six executions of the same piece - trace of the matrix `sample cov`(**a**) (and its empirical standard deviation) is about 147.7 cm^2 (57.9 cm^2, resp.) for condition A and 202.1 cm^2 (99.6 cm^2, resp.) for condition B, whereas the average trace of the matrix `sample cov`(**b**) (and its empirical standard deviation) is about 0.2259 rad^2 (0.2191 rad^2, resp.) for condition A and 0.8790 rad^2 (0.3916 rad^2, resp.) for condition B, which is a more statistically significant difference. So, in a sense, larger deviations from the mean directions seem to be associated with condition B (in which the violinists of the first section are not able to see the conductor) with respect to condition A (in which they can see the conductor). This behavior in condition B may be motivated by the absence of a reference point (the conductor) to look at in such a situation. According to a further inspection of the available data - not shown here - this change in the relative size of the movements of the heads when passing from condition A to condition B seems not to depend on the conductor, although different absolute sizes of the movements of the heads of the musicians was observed for the two conductors.

5 Discussion

In this study we have considered all the features at a global scale (i.e., on the entire video). We plan to extend the analysis by computing features at a local scale, too. In this way we'll be able to take into account features such as the movement of the baton of the conductor, not examined in the present work. By concentrating on single musical phrases we'll have the possibility of addressing the issues related to turning pages (this can be done, e.g., by selecting musical phrases in which none of the musicians has to turn a page).

In particular, ongoing work focuses on automated analysis techniques based on temporal features, to measure synchronization and social roles within and between the two groups, and the influence of the conductor, with different conductors and different music performance conditions.

Acknowledgments

The EU ICT SIEMPRE project acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 250026-2. We thank the conductors P. Borgonovo and S. Tokay, the Orchestra of the Music Conservatory of Genoa directed by A. Tappero Merlo, and our colleagues at Casa Paganini - InfoMus Research Centre.

References

1. G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. Scherer. Automated analysis of body movement in emotionally expressive piano performances. *Music Perception*, 26:103–120, 2008.
2. A. D'Ausilio, L. Badino, Y. Li, S. Tokay, L. Craighero, R. Canto, Y. Aloimonos, and L. Fadiga. Leadership in orchestra emerges from the casual relationships of movement kinematics. *PLoS one*, 7:e35757, 1–6, 2012.
3. J. W. Davidson. Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, 21:103–113, 1993.
4. J. W. Davidson. What type of information is conveyed in the body movements of solo musician performers? *J. of Human Movement Studies*, 6:279–301, 1994.
5. C. Palmer, E. Koopmans, C. Carter, J.D. Loehr, and M. Wanderley. Synchronization of motion and timing in clarinet performance. In *Proc. 2nd Int. Symposium on Performance Science*, 2009.
6. G. Varni, G. Volpe, and A. Camurri. A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. *IEEE Trans. on Multimedia*, 12:576–590, 2010.
7. M. M. Wanderley. Quantitative analysis of non-obvious performer gestures. In *Proc. Gesture Workshop 2001*, volume 2, pages 241–253, 2002.