Teaching Analytics: Towards Automatic Extraction of Orchestration Graphs Using Wearable Sensors

Luis P. Prieto
CHILI Lab, EPFL
RLC D1 740, Station 20
1015 Lausanne, Switzerland
luis.prieto@epfl.ch

Kshitij Sharma
CHILI Lab, EPFL
RLC D1 740, Station 20
1015 Lausanne, Switzerland
kshitij.sharma@epfl.ch

Pierre Dillenbourg
CHILI Lab, EPFL
RLC D1 740, Station 20
1015 Lausanne, Switzerland
pierre.dillenbourg@epfl.ch

María Jesús Rodríguez-Triana
REACT Lab, EPFL
ME A3 30, Station 9
1015 Lausanne, Switzerland
maria.rodrigueztriana@epfl.ch

ABSTRACT

'Teaching analytics' is the application of learning analytics techniques to understand teaching and learning processes, and eventually enable supportive interventions. However, in the case of (often, half-improvised) teaching in face-to-face classrooms, such interventions would require first an understanding of what the teacher actually did, as the starting point for teacher reflection and inquiry. Currently, such teacher enactment characterization requires costly manual coding by researchers. This paper presents a case study exploring the potential of machine learning techniques to automatically extract teaching actions during classroom enactment, from five data sources collected using wearable sensors (eye-tracking, EEG, accelerometer, audio and video). Our results highlight the feasibility of this approach, with high levels of accuracy in determining the social plane of interaction (90%, κ=0.8). The reliable detection of concrete teaching activity (e.g., explanation vs. questioning) accurately still remains challenging (67%, κ=0.56), a fact that will prompt further research on multimodal features and models for teaching activity extraction, as well as the collection of a larger multimodal dataset to improve the accuracy and generalizability of these methods.

Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: Computer Uses in Education—*Collaborative learning*; J.1 [**Computer Applications**]: Administrative Data Processing—*Education*

LAK '16, April 25 - 29, 2016, Edinburgh, United Kingdom

ISBN 978-1-4503-4190-5/16/04...$15.00

DOI: http://dx.doi.org/10.1145/2883851.2883927

Keywords

Teaching analytics, Multimodal learning analytics, Activity detection, Wearable sensors, Teacher reflection

1. INTRODUCTION

Aiding educators in understanding and improving teaching and learning processes is one of the main aims of learning analytics [41]. Such teacher-oriented learning analytics efforts, exemplified by tools such as eLAT [13] or LOCO-Analyst [23], so far have focused mostly on online or blended learning scenarios, using almost exclusively the traces and information available in such digital platforms. There is, however, an emergent trend within the community that also looks into capturing and modelling the physical interactions that make up the learning process in face-to-face situations, using a variety of data sources to complement the usual digital traces (multimodal learning analytics – MMLA [29])

These efforts into supporting teaching practice through analytics (also known as 'teaching analytics' [1, 44, 45]) are often portrayed as a cycle involving the gathering of data from the learning situation, analyzing it and performing an intervention as a result of the new understanding of the situation [14, 5]. Indeed, most of the existing efforts in teaching analytics focus on the first steps of this cycle, namely, the data gathering, analysis and visualization of learning processes (maybe due to the recent emergence of this community).

However, the crucial step of supporting teacher interventions based on learning analytics insights remains undersupported, starting from its very first step: knowing what the initial state of the teaching practice was, so as to know what needs to be changed. In the case of online learning, the teacher plan (either explicit or reified into the learning environment) is normally available, and the lack of enactment flexibility of most currently-used platforms somehow guarantees that such plan was 'executed'. However, in face-to-face (or blended) learning, even if the lesson plan is available, how do we know what (and how) the teacher *actually* did during the lesson, so that we can know how to intervene to improve it?

To answer this question, we need to go beyond technical,

low-level measures that are often hard to interpret, towards more 'sense-making' indicators [40], i.e., indicators that have a pedagogical meaning. The way this is normally achieved in research is through manual video coding by a human (see, e.g., [33]), which is costly in terms of time and effort.

Therefore, bringing together teaching analytics and MMLA, the overall question we try to explore in this paper is: *can we automatically characterize teaching practice in a face-to-face situation, in pedagogical terms?* Such automated characterization can be crucial to the wide and scalable application of learning analytics to teacher reflection and inquiry, an area of application of increasing importance for this community [26, 31].

To explore this question, the paper presents a case study in which we use data from multiple wearable sensors (including accelerometers, EEG or eye-trackers) and machine learning techniques to automatically characterize the teacher activity and the social plane of interaction of one teacher across 4 sessions of collaborative learning with primary school students. The next section introduces the main related work in teaching and multimodal learning analytics; later on, we outline how the general research question above has been operationalized, in the form of automatically generating an 'orchestration graph' of the teacher enactment. Afterwards, the context, methods, analyses and results of our case study are described, we discuss its main implications and limitations, and we outline the most likely avenues for further research opened by this study.

2. RELATED WORK

2.1 Teaching Analytics

Within learning analytics' general aim of improving our understanding of teaching and learning [43] and optimising learning and the environments in which it occurs [16], 'teaching analytics' is conceived as a sub-field that focuses on the design, development, evaluation of visual analytics methods and tools *for teachers*, to understand learning *and teaching* processes [45]. This particular use of learning analytics is often mentioned in connection with teacher inquiry (or teacher reflection) processes [26, 31].

Although the previous definition of teaching analytics considers both learning and teaching processes (and products) as subjects of analysis, so far teaching analytics research has been mostly focusing on analyzing student learning/behavior, and providing feedback to the teacher (see [48, 47]). Thus, many of these research works depict 'learning analytics for teachers' rather than 'teaching analytics' in a literal sense. Although the focus on student learning is certainly needed (as student learning is the main goal of any educational scenario), a complementary teacher-oriented view is also necessary to understand how and why some of the student learning processes take place, and assess the most adequate intervention.

Interestingly, most of the works in teaching analytics (both theoretical and implementations), that consider the teaching side, do so through the inclusion of teacher-generated artifacts, especially the teacher's plan for the lesson. This plan can be either explicit, as it often happens in proposals that combine learning design (LD) approaches and learning analytics [38, 21, 25, 15, 22, 42] or implicit in the resources and structure of the learning environment (e.g., in [36]).

However, how do we know whether the design of the lesson was actually followed, or what events not specified there could have an influence in the learning process? Despite the fact that teaching practice (especially, the face-to-face classroom) has often been seen as improvisational [39], few works attempt the characterization of the actual enactment of the lesson, often in very specific episodes: examining teachers' tool usage patterns (e.g., while using an educational digital library tool [51]); through explicit audience-provided feedback during lectures [37]; or through the visual analysis of the reasoning behind expert teacher assessments [19].

As it often happens elsewhere in learning analytics, most of current teaching analytics research is restricted to the analysis of easily-accessible data from digital platforms, thus creating a certain 'streetlight effect' [17] (in this context, analyzing learning only in places where there is an abundance of data, even if it is not the where most of the learning actually occurs). To address this well-known limitation of traditional learning analytics approaches (especially, in face-to-face learning scenarios), there is an emergent trend of complementing the easily-available digital traces from learning platforms with other data captured from the physical world: multimodal learning analytics.

2.2 Multimodal Learning and Teaching Analytics

A hidden assumption present in much of learning analytics research is that it is about the usage of pre-existing, machine-readable data [16], very often in the context of online learning. Noticing this blind spot, along with the realization that all learning is, almost by definition, blended [30], always containing some amount of physical embodiment (even if subdued by computer-mediated interaction), has prompted the proposal of multimodal approaches to learning analytics (MMLA, see [27, 29]). This flavor of learning analytics is used to aid in understanding and supporting more free-form, creative learning activities that are not as constrained as online ones [3], or that are more process-oriented in nature (such as project-based learning [49]).

Typical examples of MMLA include Worsley & Blikstein's work to understand creative construction activities using human annotations, speech, gesture and electro-dermal activation data [50]. Working on co-located, collaborative problem solving, Ochoa et al. [28] used video, audio and pen stroke information, extracting simple features to discriminate between experts and non-experts. More recently, other authors have proposed Feeler [12], a system that uses EEG in conjunction with application logs to promote student reflection about learning.

This kind of multimodal approaches to analyze learning processes in the physical world are not yet widespread in the sub-area of teaching analytics. Isolated examples include the iKlassroom conceptual proposal [46], which features a map of the classroom to help contextualize the real-time data about the learners in a lecture; Also in the context of university lectures, Raca and Dillenbourg [35] take an unobtrusive computer vision approach to assess student attention from their posture and other behavioral cues. Again, we can see a larger focus on modelling student actions and information, and a dearth of studies that characterize teacher practice in the classroom using such multimodal approaches.

There exists, nonetheless, a wealth of research in the field of sensors (especially using inertial sensors such as accelerometers), very often applied to the fields of health and enter-

tainment [32]. However, the large majority of the initiatives in this field target low-level, physical activities such as walking, running, sitting, etc. Such activities are unlikely to prompt interesting reflection from a teacher: we need novel ways of characterizing teacher practice in terms that "make sense" [40] for teaching practitioners. Only very recently, researchers in education are starting to turn to higher-level features and the modelling of pedagogically-meaningful interactions [2]; these efforts, however, are still confined to particular kinds of classroom episodes and pedagogical approaches (e.g., question turns in dialogic learning), and many challenges of the data gathering setup are in the process of being tackled (e.g., for accurate automated speech recognition [10]).

3. OPERATIONALIZING 'TEACHER PRACTICE': ORCHESTRATION

From the related work outlined above, we see that teaching analytics that look at actual (blended or face-to-face) teacher enactment of learning situations is an essential missing piece of support for the teacher reflection/inquiry cycle. However, general-purpose physical activities commonly used in wearable sensors (like walking and sitting) are bound to have little significance for a teacher. Hence, we need a generic but still pedagogically-meaningful way of characterizing the variety of pedagogical approaches that often coexist in everyday classroom practice.

One potential way of characterizing teacher practice (especially when using technology) is that of orchestration, defined as "the process of productively coordinating supportive interventions across multiple learning activities occurring at multiple social levels" [8]. In line with this definition of teaching practice as orchestration, graphical and computational representations of the orchestration of a lesson can be made (what may be called 'orchestration graphs' [7]). This kind of graph, representing time horizontally and social plane (individual, group, or whole-class) vertically, can be used to model the student learner activities, but also the teachers' supporting actions (such as explanation, monitoring, repairs, etc.) [34], and have been used extensively in computer-supported collaborative learning (CSCL), both to express the teacher plan (or 'script') [9] and the improvised actions during enactment [34].

So far, these kinds of representations are being generated post-hoc by researchers, on the basis of observations or the manual coding of videos of the lesson (see figure 1, middle). This kind of process, if done in a detailed manner, is very time-consuming, and the time required to do such analyses makes the feedback cycle of teacher practice and reflection unnecessarily long. Indeed, automated characterization of teacher practice along the lines of these orchestration graphs could be a great enabler for teacher inquiry processes in conjunction with other teaching analytics more focused on the learner.

With these elements in mind, and taking into account that other efforts are already looking into the multimodal characterization of student activities in physical classrooms [11, 50], we can operationalize our general research question about automatically characterizing teaching practice, into a more concrete one: *can we use multimodal teaching analytics to extract automatically the (teacher-side) orchestration graph of the enactment of a lesson?*. Towards this aim, we set ourselves to: 1) explore different sensors and modalities (as well as different features extracted from them) in order to assess their predictive power to build the teaching activity and social plane of the orchestration graph; and 2) generate predictive models that use those data sources and features to automatically characterize the teaching activity and social plane of a lesson's enactment. Below, we present an exploratory case study in which data from five different modalities are used to characterize the teaching activities and social planes of interaction of a single teacher, across four sessions with primary school students.

Figure 2: Partial view of the classroom in one of the case study sessions. In the center, the teacher wearing the eye-tracking and EEG devices

4. CASE STUDY

4.1 Context

The data for our study was gathered during an open doors day at our lab, in which entire classes of primary students from nearby schools are shown novel educational technologies. In this case, the visits were structured as simulated math lessons in a room equipped as a multi-tabletop classroom (see figure 2). Four sessions of 35-40 min were held with four different cohorts of 19-21 students per session, in which a researcher (wearing a number of sensors) acted as the main teacher-facilitator of the session. In two of the sessions the researcher had an assistant (to provide more variety in orchestration load in the otherwise very similar situations), and in all of the sessions 1-2 of the usual school teachers accompanied the children, acting as observers.

The four sessions had a similar lesson plan (see also figure 1) and similar usages of classroom technology, including the use of tangible, paper-based geometry exercises to be solved in small groups, and a whole-class collaborative and competitive game that used those same tabletops and the classroom projector, based around the same geometrical notions (rotation, translation, coordinate systems). Although the general plan for the sessions was the same (alternating phases of small-group student work and whole-class synchronization points to keep the whole class engaged), the activities and social interaction for each of the sessions were left fluid, improvised over the skeleton of the lesson plan (as it often happens in everyday teaching practice in primary schools). More concretely, the activities were organized in the following coarse sequence: 1) Explanation of the activity and technology involved; 2) Questioning of students about the mathematical concepts to be seen (to have an idea of their level of prior knowledge); 3) In small groups, use the

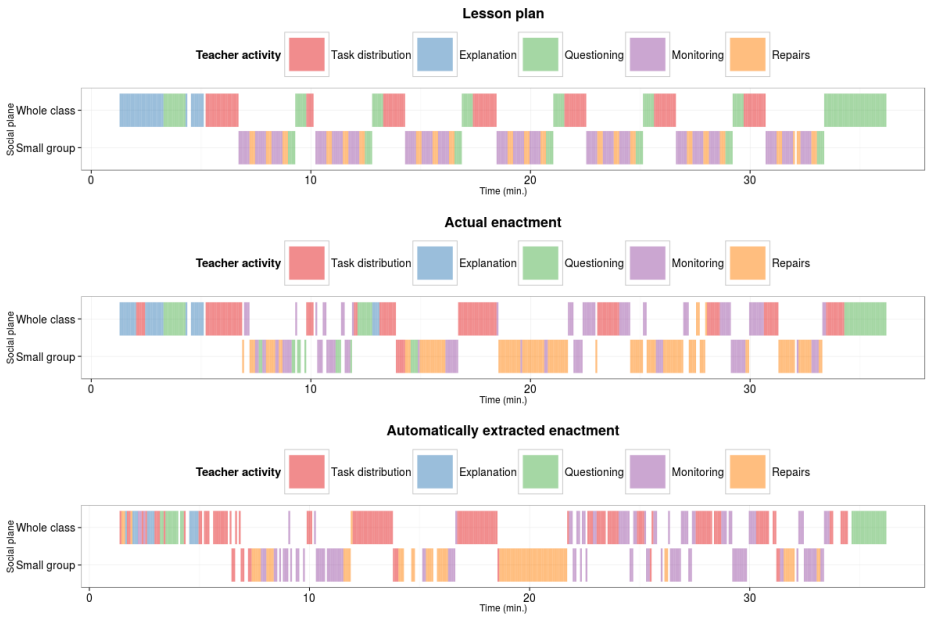

Figure 1: Example orchestration graphs, including the representation of the intended lesson plan (top), the actual teaching enactment as coded by a human researcher (middle), and the orchestration graph predicted by the best-performing models of our study (bottom)

tabletops to solve very basic geometry exercises, in order to get familiar with the concepts and technology presented; 4) Play a whole-class game in which students first collaborate in small groups to rotate/translate geometric figures, and then there is a whole-class resolution phase to see which team better protected their area; 5) The teacher does a final round of questions to assess the students' new understanding of the concepts.

4.2 Data Gathering and Feature Extraction

During the four sessions described above, the main teacher/facilitator wore several sensors in order to capture relevant teaching practice data (taking into account that teaching has an important cognitive aspect, but also a physical one): a single-electrode, portable electroencephalogram (EEG) device, mobile eye-tracker googgles (which recorded not only the eye movements, but also a subjective video+audio stream), plus a smartphone located in his pocket, set to record 3-axis accelerometer data as he moved around the classroom. From these sensors, five different data streams were considered: 1) Eye-tracking variables (e.g., saccades, fixations, pupil size); 2) EEG data (including the raw electrode reading, the usual EEG bands, attention, etc.); 3) Accelerometer readings; 4) Subjective video feed depicting the field of view of the teacher (taken from the eye-tracker's camera); and 5) Subjective audio feed (also obtained from the eye-tracker recording).

From these five data streams, up to 144 features were extracted (see Table 1). In general, we explored generic, relatively simple features used for multiple purposes (e.g., simple face detection in the video, audio energy or envelope, general fixation/saccade features), rather than going into more advanced, expensive techniques like speech segmentation and recognition, or the definition of areas of interest, visual object recognition, etc.

Given the disparity of sampling rates of the different devices, each data stream was divided into equal-length, rolling windows of 10 seconds, using a 5s. slide. Then, the different features were calculated for each 10s window[1] (e.g., by averages, deviations, maximum values, etc.), a technique commonly used in the activity detection field [32]. The feature extraction and data analysis pipeline was developed using SMI's BeGaze software (for eye-tracking feature extraction), Matlab® for audio processing (including the Matlab Audio Analysis Library [20]), and the DLib[2] library to perform basic face detection on the video stream. Then, the data from all the features were joined and analyzed using R.

The subjective audio/video stream has then been manually coded by a human researcher, assigning to each 10s window a value for the teaching activity being done (explanation, questioning, monitoring, repairs or task distribution/transition), as well as the social plane of the teacher's interaction at that moment (individual, group or whole-class). The different machine learning models described in the following section have been trained and validated against this ground truth.

Given the limits of this multimodal dataset (with only 1 subject, over 4 sessions), the models have been trained over the data of 3 of the sessions, with 25 iterations of bootstrap resampling, in order to tune the model parameters and have a first estimation of *in-session* performance. Finally, the models have been tested against the data of the remaining session, to give a more realistic estimation of the performance of the model when predicting about data from a session that the algorithm has never been trained against. This

[1]During exploratory data analysis, other window lengths (e.g., one second) were also used, with similar or worse predictive performance.
[2]http://dlib.net/

Data sources				
Eye-tracking	**EEG**	**Accelerometers**	**(Subjective) Video**	**(Subjective) Audio**
Pupil diameter (+sd)	Electrode	X value (+sd)	Image blurriness	Zero-cross rate
Nr. long fixations	Attention	Y value (+sd)	(+sd,median,max)	Energy
Saccade speed	Meditation	Z value (+sd)	Nr. blurry frames	Energy entropy
Fixation duration	Delta band	Jerk (+sd)	Nr. blurry episodes	Spectral centroid
Fixation dispersion	Theta band	Jerk FFT	Length blurry episodes	Spectral spread
Saccade duration	Low Alpha band	(30 coefs.)	(+sd,median,max)	Spectral entropy
Saccade amplitude	High Alpha band		Length clear episodes	Spectral flux
Saccade length	Low Beta band		(+sd,median,max)	Spectral rolloff
Saccade velocity	High Beta band		Nr. faces per frame	Mel-Frequency Cepstrum
	Low Gamma band		(+sd,median,max)	(MFCC, 13 coefs.)
	Mid Gamma band		Nr. of frames with faces	Harmonic ratio
	Blink strength		Nr. of face episodes	Fundamental frequency
			Length of face episodes	Chroma vector (12 coefs.)
			(+sd,median,max)	Auto-correlation (6 coefs.)
			Length of face-free episodes	Envelope (+sd)
			(+sd,median,max)	Envelope skew, kurtosis
			Faces per clear episode	Linear Predictor
			(+sd)	coefficients (6 coefs.)
				Line Spectral Frequency
				coefficients (6 coefs.)

Table 1: Overview of the features extracted for the different data sources in the study. In general, averages of the value over a 10s window are taken; (+sd,median,max) denotes that standard deviations, medians and maximum values were also extracted

process has been repeated holding up for testing each one of the sessions, and then averaging the performance in this 'out of session' testing. For model comparison, the Kappa (κ) statistic has been used, as it considers not only the accuracy of the model, but also how much better than a random predictor it is.

4.3 Results: Predicting Teaching Activity

Given our first goal of trying to understand the predictive value of each data source and its features (in this case, for detecting whether the teacher was explaining, vs. monitoring the work of students), we first tried predicting the teaching activity using data from a single data source at a time. During our prior exploratory analyses with different machine learning models from the families commonly used for classification problems, we had found that random forests (RF, [4]) performed the best (or close to the best) in almost all combinations of data in this dataset. The results of this *'mono-modal prediction'* using random forests are shown at the beginning of table 2. We can see that the eye-tracking, audio and video streams perform much better than EEG and accelerometer data, from which we get not much better accuracy than if we selected randomly. Audio features performs the best out of the five data sources, but still the predictions are not very accurate (55.9%, κ=0.41).

To understand the value of having such a rich multimodal dataset, we also trained a random forest on our whole dataset and features. We can see that the accuracy of this *'full multimodal'* prediction is larger, but the performance on out-of-session episodes is still not very high (63.8%, κ=0.52). However, in such a large set of features, it is clear that not all of them will be equally predictive, and many will be just noise. In order to understand which variables are the most informative (with the hope of increasing the performance of our teaching activity extraction), we have triangulated among two methods: a) We have calculated the effect size of each of the features with respect to teaching activity [6], in order to find those with the highest distinguishablity; b) We computed the permutation variable importance of each feature in the random forest (which estimates which variables are most important over all the decision trees of the RF). The results of this ranking, in table 3 (left), show that the pupil diameter mean size, taken from the eye-tracker, is the most important variable according to both measures. Aside from this, we can also note that the simple face detection features we calculated (e.g., the maximum number of faces that appeared in a frame of the episode) also feature among the top variables in distinguishing among different teacher activities. The fact that also a few audio features made it into this list (e.g., envelope skewness, which captures the overall asymmetry of the wave, and can be related with the teacher's voice volume), highlights the value of having multimodal data in order to predict not-so-obvious, higher-level activities such as these.

An important issue to note is that in all the aforementioned models, each episode is used for training and prediction in an independent manner (regardless of their order and position in time). However, teaching activities (as well as learning activities) form sequences over time, the same way that lesson plans are normally structured as sequences. As a first attempt to exploit the *temporal structure* of the sessions and the activities, we developed another random forest predictor using the most important multimodal features (which itself gave a slight increase in performance). Then, we corrected these RF prediction probabilities (i.e., how likely we are to be in a certain teaching activity, given the current input features) with those of the transitions between activities from a 10s window to the next (by building a discrete-time Markov chain –DTMC– from the training data). This 'multimodal Markov-improved random forest' predictor gave us a slight increase in accuracy (up to κ=0.56, see the last row

Data source	Features	Best model	In-session perf.		Out-of-session perf.	
			Accuracy	κ	Accuracy	κ
Eye-tracking only	All	Random Forest	50.2%	0.34	45.7%	0.28
EEG only	All	Random Forest	34.1%	0.11	29%	0.06
Accelerometer only	All	Random Forest	44%	0.25	31%	0.09
Audio only	All	Random Forest	58.2%	0.45	55.9%	0.41
Video only	All	Random Forest	50.8%	0.35	45.7%	0.28
All	All	Random Forest	67.4%	0.57	63.2%	0.51
Audio+video	All	(Random Forest)	64.2%	0.53	61.7%	0.49
All	Top 7	(SVM)			61%	0.49
All	**Top 80**	**RF+Markov Chain**			**67.3%**	**0.56**

Table 2: Performance of different models, data sources and features in predicting teacher activities

in table 2), but still did not manage to make the prediction overly reliable.

Another aspect worth exploring in such multimodal analyses, which often feature hundreds of different features from different data sources, is the *cost-benefit* analysis of: a) gathering such multimodal data (e.g., the cost of the device itself, and the effort of setting it up for recording a session); and b) the computational cost (in terms of time and computational power) required to pre-process, analyze and predict the multimodal data. In order to explore these aspects, we show in table 2 two additional predictors and their performance: we can see that a simpler random forest predictor using only the audio and video streams from the subjective head-mounted camera, already provides a performance that is not so different from that of the full multimodal dataset (without the Markov correction). On the other hand, a much faster and simpler model, using a support vector machine (SVM) and only the top seven variables in terms of importance, also provides a similar level of performance (accuracy of 61%, κ=0.49).

4.4 Results: Predicting Social Plane of Interaction

In general, we have followed the same sequence of analyses and modelling as described above for the automated extraction of teacher activities. In this case, however, we are trying to discriminate between the moments in which the teacher is interacting individually, in small group or with the whole classroom of students. Given that the lesson plan of the sessions (and their actual enactment) included very few interactions at the *individual* level, in the analyses below we will try to distinguish only among two social planes, small group and whole-class. Again, we have used the κ statistic when predicting out-of-session (i.e., on data of a session the model has not been trained with) as the main yardstick to compare performance of the predictive models. In this case, our exploratory modelling concluded that generalized boosted models (GBM, concretely, stochastic gradient-boosted decision trees [18]) performed better, and is thus used below for comparison purposes.

The performance of predictive models based on a *single data source* can be found at the beginning of table Table 4. There, we can see that again the predictive models based on eye-tracking, audio and video data perform much better than those based on EEG or accelerometer data. In this case, surprisingly, eye-tracking features perform the best out of the five data sources (achieving already 86.1% accuracy, κ=0.72).

Regarding the added value of having a *multimodal* dataset, a GBM model fed with all the features from the five data sources achieved better performance than the eye-tracking-only model (89.6% accuracy, κ=0.79). To understand which of the features in our multimodal dataset might have most predictive potential for the social level of interaction, we again triangulated among the effect size calculation of each features (to assess distinguishability) and the variable importance that can be extracted from the multimodal GBM predictor. In table 3 (right) we can see that, again, the pupil diameter mean is the most important feature by all accounts. The rest of this feature ranking is dominated by video-based features, both based on face-detection, as well as those based in blur of the image (indicating teacher moving the head and, hence, the field of view). Using the variables from this ranking to achieve a more efficient multimodal predictor afforded only marginal benefits over the full-featured one (89.9% accuracy, κ=0.8, using the first 81 variables). In the case of predicting social level of interaction, our attempts of incorporating the time structure of the session through Markov chains did not provide any additional increase in accuracy.

Regarding the *cost-benefit* (or simplicity vs. value) tradeoffs when building these multimodal predictors, again we found that a predictor based in the audiovisual information only (see table 4) provided a successful extraction of the social level around 85% of the time (κ=0.69), comparable to the one based on eye-tracking, and not much worse than the most accurate model. Similarly, a much simpler and faster (but still multimodal) support vector machine based on the top five features of our ranking (table 3, right) obtains an accuracy comparable to that of the full dataset (88.2%, κ=0.76).

5. DISCUSSION

The results presented above provide a first exploration of the multimodal data streams and the feature space available for researchers in teacher analytics, in order to achieve an automated characterization of teacher activity in pedagogical terms (rather than just physical ones), and show how reasonable accuracy can be achieved by using machine learning techniques, even with such simple and generic features. Indeed, our models were able to distinguish between two different social planes of interaction, close to 90% of the time, and the resulting orchestration graph so generated looked remarkably similar to the actual one, as coded by a human (see figure 1, bottom) and, certainly, was closer to what happened in the classroom than just using the lesson plan as a proxy.

Teacher activity			Social plane of interaction		
Feature	Rank ES	Rank RF	Feature	Rank ES	Rank GBM
Avg. pupil diameter (ET)	1	1	Avg. pupil diameter (ET)	1	1
Max. nr. faces/frame (VD)	2	4	Std. dev. blur in frame (VD)	2	3
Std. dev. faces per frame (VD)	3	–	Max. nr. faces/frame (VD)	3	5
Envelope skewness (AU)	4	2	Max. len. episode w/o faces (VD)	4	14
Std. dev. faces/clear frame (VD)	5	6	Std. dev. faces/frame (VD)	5	–
Total nr. faces/window (VD)	6	3	Std. dev. faces/clear frame (VD)	6	9
3rd MFCC coefficient (AU)	7	16	Med. len. episode w/o faces (VD)	7	51

Table 3: Most predictive variables in the multimodal feature set, according to the ranking obtained by measuring effect size (ES) and importance in the best-performing whole dataset model (RF/GBM). Along with the features, in parentheses, the data source to which it belongs (ET=eye-tracking, AU=audio, VD=video)

			In-session perf.		Out-of-session perf.	
Data source	**Features**	**Best model**	Accuracy	κ	Accuracy	κ
Eye-tracking only	All	Gradient Boosted T.	87.5%	0.75	86.1%	0.72
EEG only	All	Gradient Boosted T.	55.1%	0.08	50.9%	–0.02
Accelerometer only	All	Gradient Boosted T.	67.6%	0.34	61.2%	0.19
Audio only	All	Gradient Boosted T.	81.4%	0.62	79.3%	0.58
Video only	All	Gradient Boosted T.	81.7%	0.63	81.9%	0.63
All	All	Gradient Boosted T.	90.6%	0.81	89.6%	0.79
Audio+video	All	Gradient Boosted T.	86.1%	0.72	84.8%	0.69
All	Top 5	(SVM)			88.2%	0.76
All	**Top 81**	**Gradient Boosted T.**	**90.6%**	**0.81**	**89.9%**	**0.80**

Table 4: Performance of different models, data sources and features in predicting social plane of interaction

Distinguishing among the different teaching activities, however, still remains difficult. Looking at the erroneously-predicted episodes, we find that certain kinds of error were more common than others (e.g., the activities of monitoring students' work and providing repairs when one of them asks a question, which flow very fluidly into each other[3]). Such results hint at the necessity of developing new sets of features (e.g., based on basic automated speech analysis), but also of developing further our coding schemes so that they provide as much pedagogical value as possible, while remaining distinguishable.

In our exploration of the different multimodal data sources, we have found that basic features based on accelerometer and EEG signals provide very poor information to distinguish teaching activity and social plane (which is to be expected, as they are rather noisy and contaminated by even minimal physical movement, which can be irrelevant for this kind of characterization). On the other hand, we found that eye-tracking data had a surprising amount of useful information, especially the mean pupil diameter of each 10s window. This finding can be interpreted in the sense that such measures are known to be related to emotional response and cognitive load factors, that is, this measure may capture the different levels of cognitive load elicited by the different teaching activities at different social planes (a hypothesis supported by our previous research in measuring cognitive load in the classroom [33]). However, it is also worth noting that eye-tracking measures (and the features extracted from them) are also most prone to be subject-dependent, which may pose a limitation if we are looking for models that are generalizable across teachers.

[3]For instance, the same random forest models, if applied to a coding scheme in which monitoring and repair are joined into a single category, achieve an accuracy of 75% (κ=0.6).

This exploratory case study also enabled us to uncover interesting trade-offs between the accuracy of the machine learning models and the cost, effort and convenience of gathering and analyzing the different data sources. For instance, we found that using just the subjective audiovisual feed (easily attainable using a small camera such as those used for sports, head-mounted cameras like Google Glass, or even a simple mobile phone) provided already quite good accuracy, even if the eye-tracking measures provided an additional edge (and have other advantages from the point of view of research, such as providing access to the subject's cognitive load). Even more convenient (and cheaper) would be the use of fixed cameras (e.g., such as the ones used to assess student attention in [35]), although the subjective feeds are more likely to capture the teacher's experience and activities (e.g., occlusions, dead angles), and has side benefits for teacher inquiry processes (as it is easier to remember or understand a situation, or empathize with it – as we found during the coding of the videos). Regarding the cost in computing power we found that, although the best performance was attained by rather complex, black-box models like random forests, much simpler and faster models like support vector machines, using only a handful of multimodal features, also provided reasonable results (which can be useful, for instance, if real-time extraction of the orchestration graph were needed).

Despite the interest of these findings, this exploratory case study also suffers from a number of limitations, the most important of which is the data set used (featuring only one subject, across four sessions that were rather similar to each other). This fact make the accuracies and performance reported in this paper very tied to the concrete subject and situation of our case study (i.e., low potential generalizability to other teachers, or to very different classroom situa-

tions). This is especially true of some features we found very important (like pupil diameter), although it may be less so for many of the audiovisual features that helped distinguish between activities and social planes (e.g., faces in the field of view, or certain audio features). On the positive side, our finding that audio/video data may provide adequate performance, along with the increasing affordability of such means of data gathering, make us hope for more scalable approaches that can be widely deployed, as other recent work in this area [2] also demonstrate.

Another limitation of this study is the feature set and machine learning methods explored. Even if we used more than a hundred features from five data streams, the features used were rather simplistic and generic, in many cases not tuned specifically for such teaching practice discovery. Although our results already provide a first step into separating useful features in each of the kinds of data, more work is still needed to distill more targeted features (maybe from other sensor data streams) that can help us distinguish more clearly between certain similar activities. Furthermore, the teaching practice categorization used in this study is only one example, and other characterizations are also possible, especially for researchers or practitioners interested in concrete pedagogical approaches such as collaborative learning, or inquiry-based learning (which will also prompt new exploration efforts into different sets of useful multimodal features). Also, further exploration is needed in applying more complex algorithms (e.g., deep/recurrent neural networks), which have recently shown promising capabilities in dealing with rich multimodal data (e.g., [24]).

Finally, it should be noted that one important aspect of orchestration, of which this study has barely scratched the surface, is that of the *time structure* of the teaching activities, and of the different signals used, as teaching is by nature a sequence of actions over time. We anticipate that looking into the rhythm and pulse of each classroom situation, and exploiting time-series tools such as semi-Markov models (which model the time spent in one state before jumping to the next one), will probably provide additional increases in the accuracy of our automated teaching activity extraction.

6. CONCLUSIONS AND FUTURE WORK

We started this paper by noting the scarcity of teaching analytics research that actually studies teacher practice, and especially teacher actions in the face-to-face classroom, beyond their lesson plans. As a first step into incorporating analyses of teaching activities to existing learning analytics and teacher reflection processes, we have explored a multimodal approach to automatically extracting the orchestration graph of a face-to-face collaborative learning session, based on the data of several wearable sensors.

The results of our study show that the approach is feasible and achieves reasonably good accuracy (especially to discriminate the for social plane of interaction). Also, the different models and data sources used show how this kind of approach is feasible, not only for researcher teams with advanced equipment, but also for mass deployment by practitioners themselves (by using a simple subjective audiovisual feed). However, our study also demonstrates the difficulty in discriminating between certain teaching activities, solely on the basis of a set of generic (and rather simple) features. More efforts are needed into extracting new discriminating features from these and other data sources.

These favorable results, and the hope that even a personal camera or a mobile phone could achieve useful results, open the door for larger-scale, wearable-based studies into teaching practice, either based on the generic characterization of practice as 'orchestration' (presented in this paper), or through different ones based on more concrete pedagogical approaches. Regarding the categorization of teaching practice, we are currently validating the interest of the teacher community in this kind of approach to reflection and teaching analytics through example visualizations extracted from this dataset[4].

In our current and future work we are also continuing the exploration of more accurate statistical models (especially, those aimed at the time structure of the problem), and of more discriminating features for automatic activity and social plane extraction. However, our most prominent current work relates to the recording of a larger multimodal dataset of teaching sessions by multiple teachers, especially in primary schools but also at other educational levels. We hope this expanded dataset will help us overcome the largest limitation of the work presented here, namely, the limited generalizability of the results beyond a concrete classroom. Once this dataset is gathered, we expect to share it, along with the processing and analytical code used, with the teaching analytics community, as we believe openness and collaboration will be crucial in using this kind of multimodal analytics to study orchestration (since expertise from pedagogy, psychology, and a wide array of data and signal processing fields is required to make sense of the heterogeneous data), and to overcome the current scalability limitations of research into face-to-face orchestration.

7. ACKNOWLEDGMENTS

This research was supported by a Marie Curie Fellowship within the 7th European Community Framework Programme (MIOCTI, FP7-PEOPLE-2012-IEF project no. 327384).

8. REFERENCES

[1] M. Bienkowski, M. Feng, and B. Means. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. Technical report, U.S. Department of Education, Office of Educational Technology, 2012.

[2] N. Blanchard, S. D'Mello, M. Nystrand, and A. M. Olney. Automatic classification of question & answer discourse segments from teacher's speech in classrooms. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015), International Educational Data Mining Society*, 2015.

[3] P. Blikstein. Multimodal learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 102–106. ACM, 2013.

[4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] D. Clow. The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 134–138. ACM, 2012.

[6] J. Cohen. *Statistical power analysis for the behavioral sciences (rev.* Lawrence Erlbaum Associates, Inc, 1977.

[4]See http://classroom-mirror.meteor.com.

[7] P. Dillenbourg. *Orchestration Graphs: Modeling Scalable Education*. EPFL Press, Lausanne, Switzerland, 1 edition, 2015.

[8] P. Dillenbourg, S. Järvelä, and F. Fischer. The evolution of research on computer-supported collaborative learning. In *Technology-enhanced learning*, pages 3–19. Springer, 2009.

[9] P. Dillenbourg and P. Tchounikine. Flexibility in macro-scripts for computer-supported collaborative learning. *Journal of computer assisted learning*, 23(1):1–13, 2007.

[10] S. K. D'Mello, A. M. Olney, N. Blanchard, B. Samei, X. Sun, B. Ward, and S. Kelly. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 557–566. ACM, 2015.

[11] F. Domínguez, K. Chiluiza, V. Echeverria, and X. Ochoa. Multimodal selfies: Designing a multimodal recording device for students in traditional classrooms. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 567–574. ACM, 2015.

[12] E. Durall and T. Leinonen. Feeler: supporting awareness and reflection about learning through EEG data. In *Proceedings of the 5th Workshop on Awareness and Reflection in Technology Enhanced Learning In conjunction with the 10th European Conference on Technology Enhanced Learning*, pages 67–73, 2015.

[13] A. L. Dyckhoff, D. Zielke, M. Bültmann, M. A. Chatti, and U. Schroeder. Design and implementation of a learning analytics toolkit for teachers. *Journal of Educational Technology & Society*, 15(3):58–76, 2012.

[14] T. Elias. Learning analytics: The definitions, the processes, and the potential. 2011.

[15] V. Emin-Martinez, C. Hansen, M. J. Rodríguez-Triana, B. Wasson, Y. Mor, R. Ferguson, and J.-P. Pernin. Towards teacher-led design inquiry of learning. *E-learning papers. Special issue on Learning Analytics and Assessment*, 36:3–14, 2014.

[16] R. Ferguson. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5-6):304–317, 2012.

[17] D. H. Freedman. Why scientific studies are so often wrong: The streetlight effect. *Discover Magazine*, 26, 2010.

[18] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[19] G. Gauthier. Using teaching analytics to inform assessment practices in technology mediated problem solving tasks. In Vatrapu et al. [48].

[20] T. Giannakopoulos and A. Pikrakis. *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press, 2014.

[21] Y. Hayashi and R. Mizoguchi. Articulation of scenario construction of lessons based on ontological engineering. In Vatrapu et al. [47].

[22] D. Hernández-Leo, J. I. Asensio-Pérez, M. Derntl, L. P. Prieto, and J. Chacón. Ilde: Community environment for conceptualizing, authoring and deploying learning activities. In C. Rensing, S. de Freitas, T. Ley, and P. Muñoz-Merino, editors, *Open Learning and Teaching in Educational Communities*, volume 8719 of *Lecture Notes in Computer Science*, pages 490–493. Springer International Publishing, 2014.

[23] J. Jovanović, D. Gašević, C. Brooks, V. Devedžić, and M. Hatala. Loco-analyst: A tool for raising teachers' awareness in online learning environments. In *Creating New Learning Experiences on a Global Scale*, pages 112–126. Springer, 2007.

[24] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 427–434, New York, NY, USA, 2015. ACM.

[25] D. Laurillard. *Teaching as a design science: Building pedagogical patterns for learning and technology*. Routledge, 2012.

[26] Y. Mor, R. Ferguson, and B. Wasson. Editorial: Learning design, teacher inquiry into student learning and learning analytics: A call for action. *British Journal of Educational Technology*, 46(2):221–229, 2015.

[27] L.-P. Morency, S. Oviatt, S. Scherer, N. Weibel, and M. Worsley. Icmi 2013 grand challenge workshop on multimodal learning analytics. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 373–378. ACM, 2013.

[28] X. Ochoa, K. Chiluiza, G. Méndez, G. Luzardo, B. Guamán, and J. Castells. Expertise estimation based on simple multimodal features. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 583–590. ACM, 2013.

[29] X. Ochoa, M. Worsley, K. Chiluiza, and S. Luz. Mla'14: Third multimodal learning analytics workshop and grand challenges. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 531–532. ACM, 2014.

[30] M. Oliver and K. Trigwell. Can 'blended learning'be redeemed? *E-learning and Digital Media*, 2(1):17–26, 2005.

[31] D. Persico and F. Pozzi. Informing learning design with learning analytics to improve teacher inquiry. *British Journal of Educational Technology*, 46(2):230–248, 2015.

[32] S. J. Preece, J. Y. Goulermas, L. P. Kenney, D. Howard, K. Meijer, and R. Crompton. Activity identification using body-mounted sensors—a review of classification techniques. *Physiological measurement*, 30(4):R1, 2009.

[33] L. P. Prieto, K. Sharma, and P. Dillenbourg. Studying teacher orchestration load in technology-enhanced classrooms. In *Design for Teaching and Learning in a Networked World*, pages 268–281. Springer, 2015.

[34] L. P. Prieto, S. Villagrá-Sobrino, I. M. Jorrín-Abellán, A. Martínez-Monés, and Y. Dimitriadis. Recurrent routines: Analyzing and supporting orchestration in technology-enhanced primary classrooms. *Computers & Education*, 57(1):1214–1227, 2011.

[35] M. Raca and P. Dillenbourg. System for assessing

classroom attention. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 265–269. ACM, 2013.

[36] S. Rebholz, P. Libbrecht, and W. Müller. Learning analytics as an investigation tool for teaching practitioners. In Vatrapu et al. [47].

[37] V. Rivera-Pelayo, E. Lacić, V. Zacharias, and R. Studer. LIM App: Reflecting on Audience Feedback for Improving Presentation Skills. In D. Hernández-Leo, T. Ley, R. Klamma, and A. Harrer, editors, *Scaling up Learning for Sustained Impact*, number 8095 in Lecture Notes in Computer Science, pages 514–519. Springer Berlin Heidelberg, Sept. 2013.

[38] M. J. Rodríguez-Triana, A. Martínez-Monés, J. I. Asensio-Pérez, and Y. Dimitriadis. Script-aware monitoring model: Using teachers' pedagogical intentions to guide learning analytics. In Vatrapu et al. [47].

[39] R. K. Sawyer. Creative teaching: Collaborative discussion as disciplined improvisation. *Educational researcher*, 33(2):12–20, 2004.

[40] G. Siemens. Sensemaking: Beyond analytics as a technical activity. Presentation at the EDUCAUSE ELI 2012 Online Spring Focus Session.

[41] G. Siemens and P. Long. Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, 46(5):30, 2011.

[42] K. Thompson, R. Martinez-Maldonado, D. Wardak, P. Goodyear, and L. Carvalho. Analysing F2F Collaborative Design and Learning : Experiences in a Design Studio A Place for Learning. In L. P. Prieto, Y. Dimitriadis, A. Harrer, M. Milrad, and J. D. Slotta, editors, *Proceedings of the Orchestrated Collaborative Classroom Workshop 2015 co-located with 11th International Conference on Computer Supported Collaborative Learning*, pages 25–29, Gothenburg, Sweden, 2015.

[43] M. van Harmelen and D. Workman. Analytics for learning and teaching. *JISC CETIS Analytics Series*, 1(3):1–41, 2012.

[44] R. Vatrapu, C. Teplovs, N. Fujita, and S. Bull. Towards visual analytics for teachers' dynamic diagnostic pedagogical decision-making. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, pages 93–98. ACM, 2011.

[45] R. K. Vatrapu. Towards semiology of Teaching Analytics. In *Workshop Towards Theory and Practice of Teaching Analytics, at the European Conference on Technology Enhanced Learning*, TAPTA'12, Saarbrücken, Germany, 2012.

[46] R. K. Vatrapu, K. Kocherla, and K. Pantazos. iklassroom: Real-time, real-place teaching analytics. In Vatrapu et al. [48].

[47] R. K. Vatrapu, P. Reimann, W. Halb, and S. Bull, editors. *Proceedings of the Workshop Towards Theory and Practice of Teaching Analytics (TaPTA 2012)*, number 894 in CEUR Workshop Proceedings, Aachen, 2012.

[48] R. K. Vatrapu, P. Reimann, W. Halb, and S. Bull, editors. *Proceedings of the 2nd International Workshop on Teaching Analytics (IWTA 2013)*, number 985 in CEUR Workshop Proceedings, Aachen, 2013.

[49] M. Worsley. Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 353–356. ACM, 2012.

[50] M. Worsley and P. Blikstein. Leveraging multimodal learning analytics to differentiate student learning strategies. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 360–367. ACM, 2015.

[51] B. Xu and M. Recker. Teaching Analytics: A clustering and triangulation study of digital library user data. *Educational Technology & Society*, 15(3):103–115, 2012.