Technology Enhanced Learning of Motions Based on a Clustering Approach

Quentin Couland, Ludovic Hamon, Sébastien George

Chapter # - will be assigend by editors

TECHNOLOGY ENHANCED LEARNING OF MOTIONS BASED ON A CLUSTERING APPROACH

Quentin Couland[1], Ludovic Hamon[1], Sébastien George[1]

[1]LIUM - EA 4023, Le Mans Université, 72085 Le Mans, Cedex 9, France

{quentin.couland / ludovic.hamon / sebastien.george}@univ-lemans.fr

Abstract: The analysis of user motions can be useful in many fields to observe human behavior, follow and predict its action, intention and emotion, to interact with computer systems and enhance user experience in Virtual (VR) and Augmented Reality (AR). These analyses can be empirically made by the expert or with the help of a Technology Enhanced Learning (TEL) system, allowing the extraction of relevant information from the motion in a pedagogical context. Such analyses are rarely made from 3D captured motions. This can be explained by several factors: the complexity and high dimensionality of the data, and the difficulty to correlate the observation and analysis needs of the expert to the extracted data. Machine learning techniques could be used to address some of these problems. In particular, the use of unsupervised learning techniques could help in giving advice according to the analysis of clusters, representing user profiles. During a learning situation, the expert will be assisted in their evaluation task. This work presents two main contributions: (i) the use of clustering techniques to separate motions, into different categories according to a set of well-chosen features, and (ii) the development of a TEL environment using clustering techniques in order to assist the expert in its motion-based evaluation task.

Key words: Human motion, human learning, technology enhanced learning, machine learning, clustering

1. INTRODUCTION

Motion capture is increasingly used in multiple domains such as video-game, animation movies, Virtual Reality (VR), sport, medicine, industry and education. Thanks to breakthroughs made in electronics, Human-Computer Interface (HCI) and data processing, it is reasonable to assume that capturing, editing and sharing human gestures will be soon generalized. This assumption has a strong impact on education and on every domain implying human movements. Indeed, different kinds of information can be extracted from human motion analysis. One can easily generate low-level descriptors such as kinematic and dynamic data (Nunes & Moreira, 2016)(Larboulette & Gibet, 2015). Gestures may have a meaning for verbal (Huang, *et al.*, 2015) or non-verbal communication (Chang, *et al.*, 2013). In addition, high-level data linked to human emotion (Kobayashi, 2007), intention (Yu & Lee, 2015) and action (Kapsouras & Nikolaidis, 2014) can be reified and built. Monitoring learner activities can imply the generation of a large amount of motion data that cannot be manually analyzed (Gu & Sosnovsky, 2014). Automatic methods, such as machine learning techniques, can ease such a task. This set of techniques can process high-dimensional data for classification purposes, feature extraction, regression problems, *etc*. (Ng, 2016). In an educational context, these algorithms are used to study and classify learner actions (Lokaiczyk, *et al.*, 2007) and/or behavior (Markowska-Kaczmar, *et al.*, 2010), from motions thanks to supervised learning. However, this kind of algorithms implies: (i) the existence of a large database of specific annotated motions for each task to learn (ii), the knowledge of the different classes names in advance. There is a lack of work regarding the automatic extraction of relevant information in pedagogical situations from learner motions. This can be explained by several technical and scientific issues: the heterogeneity, the complexity, the high-dimensional nature of such data, and the need to correlate this information with the observation needs of the teacher. Some of these issues could be overcome by the use of clustering algorithms, in order to avoid the requirements related to supervised ones (database size and labeling), and by using morphology-invariant descriptors relevant in the given context. The goal of this work is to use kinematic descriptors along with clustering techniques in order to: (i) make and visualize well-separated clusters representing user profiles, (ii) use clusters features and inter-clusters distance to lead him to the cluster of acceptable motions and (iii) create a TEL environment using clustering techniques usable by an expert in order to assist them in their motion evaluation task.

The goal of this work is to create a new TEL environment dedicated to motion learning, using clustering techniques, in order to assist the expert in

their evaluation task. The creation of such a system requires the solving of some scientific and technic challenges: (i) the separation of motions in well-defined cluster, based on their properties, (ii) obtaining a separation corresponding to the degree of success of the task and (iii) the validation of this approach and the use of this system in a real learning case. These challenges were addressed through 3 experimentations, all related to throwing motions. Results show that while it was possible to achieve a good separation of the motions in different clusters corresponding to different strategies of throwing, it was not possible to obtain a separation corresponding to the degree of success of the motions. Furthermore, the experimentation conducted in order to validate the usability and the usefulness of this system showed an improvement in the quality of the learners motion.

The remainder of the chapter is structured as follows: section 2 presents a review of motion-based analysis methods with a focus on educational-based work, showing the lack of unsupervised and generic approach for motion analysis. Our new approach using clustering techniques is shown in section 3. The unsupervised approach allows using few unlabeled data in order to assist the expert. The experimentation on the separation of user profiles and its related protocol, results and discussion are detailed in section 4. Section 5 presents the TEL developed in order to validate the use of such a system in a real learning situation, showing an improvement in the learner motions. Finally, perspectives and future work ends this study.

2. RELATED WORK

Human learning motion can use captured motions, in order to assist the student in their learning task. In this context, the motion is mainly represented as a sequential evolution of human postures through time. Usually, a fixed time-step separates each posture (called "frame"). One way to represent a posture is to build a set of joints, hierarchically structured thanks to a graph, each node describing a joint. This set of joints is organized according to a skeleton model, *i.e.* a tree data structure, in which the root represents the low body part of the torso (*i.e.* the hip bone) and the nodes represent the body joints. Each node contains the position and the orientation, related to its parent node. It is possible to extract kinematic and dynamic descriptors from this structure such as the speed of the joint, its acceleration, its displacement through time, *etc.* (Nunes & Moreira, 2016) (Larboulette & Gibet, 2015). Zhu and Hu worked on the learning of specific

motions for reeducation (Zhou & Hu, 2008). The skeleton model was not systematically considered, because different kinds of sensors were used to gather motion data, depending on the observed movement. The data were used in order to analyze the patient's gait. No automatic analyses of the recorded movements were made, the observations and deductions of information were always made by a human expert. For Japanese archery learning, Yoshinaga and Soga developed a system based on a Kinect sensor to capture learner skeletons and its variations through time (Yoshinaga & Soga, 2015). Expert movements were also recorded and learners could compare their motions with the expert ones. The analysis was empirically made by humans. Le Naour *et al.* proposed a superimposition of the expert model and the student one in order to learn the throwing motion in American football (Le Naour, et al., 2019). The quality of the motion was assessed with the help of the Dynamic Time Warping algorithm computed between the expert and the learner motion, and the regularity of the learner motion between different sessions. The superimposition allowed for a better motion reproduction. Chan *et al.* used a TEL environment in order to learn dance motions (Chan, et al., 2011). Expert motions were recorded and showed to the learner. The learner motions were then compared to the expert ones, highlighting the parts of the student's body that were not synchronized with the expert, by using a distance threshold. A score was given to the student, to evaluate their performance. Maes *et al.* also worked on the visualization of the expert movements, the learning of the motion step-by-step, and the evaluation of the learner motion through a score in the same context (Maes, et al., 2012). In the last case, the score gives no hints about which part of the motion must be corrected , and thus the system gave no pedagogical feedbacks. Xu *et al.* developed a TEL environment in order to help children learning specific motions (Mingliang, et al., 2019). These motions were related to the Chinese culture: operating looms, shooting arrow, riding horses, etc. The system makes use of a database of sub-motions and two Hidden Markov Models to achieve this goal. The first one allowed the segmentation of the motions, and the second one adapted the learning process to the student if the motions were not correct. While using automatic methods, a database containing motions related to the considered study case must be captured beforehand, in order to cover the widest range of possibilities. Furthermore, no pedagogic feedback was inferred from the performed motions.

There are a lot of TEL environments dedicated to motion learning. These systems can be used in various contexts such as rehabilitation (Zhou & Hu, 2008), surgical procedure learning (Pepley, et al., 2017) and sports motion learning (Yoshinaga & Soga, 2015) (Le Naour, et al., 2019). An evaluation of the learner's motion was sometimes proposed, whether as a score, motion

data visualization (expert and learner 3D avatar in VR, sometimes superimposed), or as a visualization of the learning path (in case the motion was segmented). A lot of these systems were not based on generic models allowing to consider different tasks as well as observation and analysis needs. Indeed, the expert's knowledge is often hard-coded during the design phase. Consequently, a heavy re-engineering is required in order to adapt them to other contexts. The generic aspect of such a system must be considered during the design phase, in order to be reusable in other contexts.

Studies using supervised and unsupervised algorithms to analyze facial expressions, gestures and actions exist and some of them were based on 3D captured data. Patrona *et al.* presented a framework for action recognition and evaluation based on extreme machine learning (Patrona, *et al.*, 2018). Using fuzzy-logic, a feedback (depending on the activity context) is given to the learner, such as the velocity at specific frames, in order to improve the realized motion. This feedback requires a reference motion and a large corpus of existing motions, as the goal is to classify the motion into predefined categories from different datasets (CVD exercise, MSRC-12 and MSR-Action3D). Hachaj and Marek used a set of expert rules relating to the learner displacements (*e.g.* the distance covered by the learner in a time step), in order to classify motions (Hachaj & Marek R., 2015). Although these approaches were efficient, the motions were related to simple and everyday activities (*e.g.* walking, jogging, running) that did not require a cognitive effort or strong motor skills to learn. Furthermore, the goal was not to evaluate the success degree of the motion and the descriptors could not be used to give a pedagogical feedback. Lui *et al.* worked on video databases from which two sets of descriptors were extracted (Lui, *et al.*, 2011). These descriptors were, on the one hand, spatial and temporal localized features that were used with a Bag Of Features approach, and a manifold product on the other hand. The results showed an acceptable data partitioning, especially with the set of descriptors dedicated to the manifold product. The performed motions were also trivial in terms of cognitive effort, and the descriptors could not be used to give feedbacks to the learner. Due to the nature of the motions, the degree of success of the task was not evaluated. Pirsiavash *et al.* assessed the quality of motions without any *a priori* on the considered methods (Pirsiavash, et al., 2014). The data are extracted from videos and consist of pixels gradients, joints trajectory and successive postures of the performer. The considered motions were related to Olympic diving and figure skating. The motion were associated with the expert judge scores, and fed to a SVM algorithm, allowing to extract the most relevant motions features linked to the scores. The system then gives a score,

assessing the quality of the new motions. The results of this work show that while the scores are still far from the expert ones, they are better than the scores given by non-expert humans.

With a sufficient amount of data for the training phase, supervised machine learning algorithms are efficient when the searched and estimated hypothesis is well designed for the problem complexity. However, these kinds of algorithms need a large amount of labeled data related to the given context. The data labeling is usually a costly task in terms of time and resources. Furthermore, some pre-processing steps can change the nature of the data (*e.g.* PCA), and some decision/separation frontier cannot be easily interpreted by humans (*i.e.* such as those built by SVN or Neural Networks). Consequently, analyzing and giving feedbacks to the learner can be a hard or impossible task. Unsupervised learning approaches do not need labeling data to group them into different clusters. However, there is a lack of studies using unsupervised machine learning algorithms to automatically extract useful pedagogical information from 3D motion data in a pedagogical context. This approach could allow the automatic detection of the most distinguishing features of a set of motions, to group them as learner profiles according to the observation needs of the teachers. In addition, a more efficient help could be provided to the expert in advising the learner by observation of : (i) the features of the acceptable motion groups and (ii), the current distance separating the current performed motion from these groups. The development of this kind of system must take into account the motion variation, in order to achieve the same desired goal or task (whether it is a set of postures in space and time or the position of an object).

The presented work is based on the three following hypothesis: (i) for one identified task to learn, it is possible to group motions in separable clusters, with each cluster made of motions with common features, (ii), it is possible to automatically group gestures according to the degree of success of the motion-based task and (iii) it is possible to use clustering methods in order to create an interactive TEL system assisting the expert in its evaluation and advising task. This approach, as well as three experiments conducted to validate these hypotheses are detailed in the next sections.

3. A CLUSTERING APPROACH FOR MOTION ANALYSIS

For a manual task to learn, there is usually not a unique and perfect motion to achieve it. In most of the cases, the features of a targeted gesture are defined by one or several experts. Establishing which of those features

are relevant, allowing to tell if the motion is successful or not, depends on the context and the expectations of the professionals, which can vary from one expert to another. This means that, for a given a learning situation, the set of discriminant features is not the same for every expert. Using supervised learning algorithms implies that a database containing labeled motions exists. The degree of success of the task must be stored within the labels of each sample. In practice, most of the databases focus on trivial motions, such as sitting, running, walking, etc. The chosen approach relies on the automatic analysis of motions through clustering techniques to avoid most of the drawbacks of the supervised approach. The overall and implemented method can be seen in Fig. 1. From a captured motion-based corpus, a first pre-processing step applies several filters to clean the data if needed (*e.g.* frames loss or corrupted, framerate variation, etc.). The next step allows the extraction of well-chosen kinematic, dynamic and geometric descriptors (Larboulette & Gibet, 2015). One should be careful about them, as some descriptors are morphology-dependent (*e.g.* those related to the distance between two joints), and some are not (*e.g.* the joint rotation). The data are then analyzed through the descriptors linked to the observation needs of the teacher. These descriptors are then used in a clustering process, using the k-means algorithm, from which several metrics are computed to assess its quality. The use of an IT environment allows observing the state of the current motion in terms of: (i) features compared to those of the acceptable motion groups and (ii) distance between this motion and these groups. From the observation of this state, the expert can give feedback to the learner, while refining their observation needs.

This chapter focuses on two parts of Fig. 1 automatic analysis block, namely the clustering process and the feedback system (system advices), implying that clean data are available. An example of such data can be seen in Fig. 2c. The goal is to find a set of descriptors, algorithms and metrics to: (i) separate the motion corpus in different groups, (ii) give an indication of the degree of success for each group and (iii), automatically give advices to the learner from the group features and the current motion state through a visualization of the data. Such separation-based system would allow analyzing the unperceived or hard-perceived properties of the motion clusters, giving information related to the characteristics of different and acceptable motion profiles, and thus giving a more accurate advice for the improvement of the learner motion. The next section presents the experimentation conducted, in order to validate the presented hypotheses.

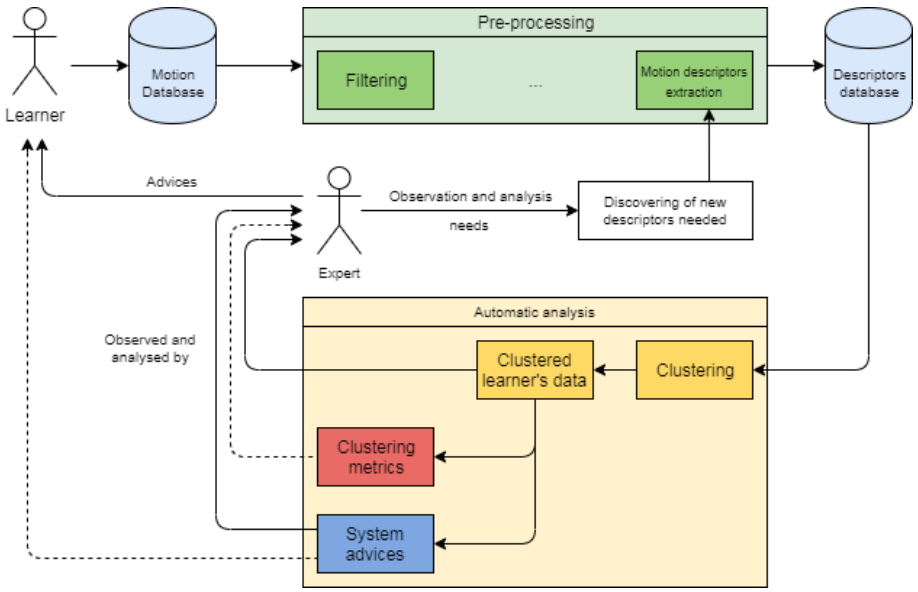


Figure 1: The Motion Learning Analytics (MLA) system, dedicated to human motion learning.

4. EXPERIMENTATION ON CLUSTERING WITH KINEMATIC DESCRIPTORS

This section is dedicated to an experimentation for the validation of the first two previous hypotheses. As a reminder, these assumptions are : (i) it is possible to separate the data into well-defined clusters, and (ii) it is possible to obtain a separation corresponding to the degree of success of the motion. The validation of the first hypothesis would prove that it is possible to obtain different learners profiles regarding the considered task, allowing the expert to adapt their advices for each group. The validation of the second hypothesis would prove that it is possible to obtain various degrees of success of the motion regarding their characteristics, allowing to determine a threshold of what is considered an acceptable motion and a better understanding of how to improve a motion, going from one profile to another.

4.1 Protocol

For this experimentation, a database made of motions requiring some dexterity was created. The Bottle Flip Challenge was the chosen task. The

goal is to throw a bottle, such as it completely rotates once on the horizontal axis, and lands correctly on a table. The distance from the person performing the gesture to the table was empirically set to 70cm (27.5 inches), indicated by a mark on the floor. The MOCAP Perception Neuron suit made of Inertial Measurement Units (IMU) was used for the capture (https://neuronmocap.com/). It allows capturing 72 joints (some of which are interpolated) at the rate of 60 frames per second. The skeleton of the subject was measured according to the official measuring guide provided with the suit, in order to have data skeletons made in accordance with the user morphology. Due to the nature of the sensors, the experimental protocol ensures that (i) no device generating electromagnetic perturbations was close to the user, and (ii) all metallic accessories were removed (including rings, bracelets, watches, belt with metallic buckle, etc.). During the experiment, the MOCAP suit had to be regularly recalibrated, due to the inherent drift of the sensors. Each subject had to perform the motion a hundred times and for every throw, the success (or not) of the task was recorded.

Fig 2.a shows the artifacts of the suit sensors, on the hand's speed data. Such data are not usable, as the original signal is distorted by the noise. In order to compensate these errors, a Savitsky-Golay filter was applied on

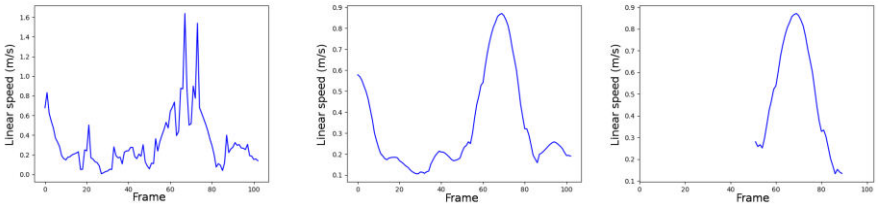


Figure 2: speed of the captured motion through time of the right-hand of a user (b): initial speed filtered (c): extracted throwing part (Couland, et al., 2018).

each motion (Fig.2.b). Then, the throwing part of the motion was automatically segmented to extract the motion part of interest (Fig.2.c). This method is based on the detection of one or more local minimums to the left and the right of the global maximum value of the speed values for the dominant hand. It is particularly suited for throwing motions, as the characteristic of such a motion implies having the highest speed value at the moment the object is released. From those cleaned data, some descriptors were computed. Since the subjects have different morphologies, morphology-invariant descriptors were chosen: speed and acceleration (vector norm and direction, components along each axis in both cases). The descriptors were computed from three moments of each cleaned motion: the beginning of the throw, the maximum value of the speed norm for the dominant hand (corresponding to the release of the bottle), and the end of the

throw. The chosen clustering method is the k-means algorithm to give a first insight of the possible separation. In addition, this algorithm is faster than other clustering methods (*i.e.* execution time scales linearly with data size) and has easily explainable results. The *k* values ranged from 2 to 10 for this experimentation.

In order to analyze the clustering results, the following metrics suited to our approaches were chosen.

To evaluate the separation quality of the obtained clusters, the Average Silhouette Score (*ASS*) was computed (Rousseeuw, 1987). The Silhouette Score (SS) is a metric indicating if a sample belongs well to its assigned cluster (compared to other clusters). The Average Silhouette Score (ASS) is the mean of every sample SS. It gives an indication about the clusters homogeneity: the highest this value is, the better the clusters are separated. This value ranges from -1 to 1, with 1 meaning that every sample is close to the others in the same clusters (the clusters are well separated), and 0 indicating that the clusters are overlapping. In this last case, a possible explanation is that the number of clusters is either too low or too high. An ASS between 0 and 0.25 means that no structure is found in the data, a value between 0.25 and 0.5 indicates that a weak structure is found (potentially artificial), an ASS above 0.5 suggests that an acceptable structure is found, while an ASS value above 0.7 means that a strong structure is found (Struyf, et al., 1997). In this context, the metric allows verifying the separation quality of the clusters, thus giving an indication about the relevancy of the computed descriptors and clustering algorithm in terms of separation.

To assess the separation quality in terms of motion groups representing the same degree of success of the task (in our case a successful, or failed throw), a metric such as the accuracy of the clustering seems to not be a relevant indicator. For example, if the k-means algorithm is considered, this metric, based on the computation of a Euclidian distance, is relative to the measured data, the required accuracy of the measuring system and the learning situation. This accuracy is often ascertained by an advanced expert both in the application domain and in computer sciences. In order to verify the difference between the ground truth and the obtained labeling (*i.e.* failed/success motion), the precision, the recall, the F1-score and the Adjusted Rand Index (ARI) were chosen. These metrics were only computed for $k=2$, as the ground truth is defined for $k=2$ (successful/failed). As a reminder, the F1-score is a combination of two metrics (recall and precision) representing the labeling accuracy. This value ranges from 0 to 1, with 1 indicating a perfect matching. The ARI is a measure of the similarity between two data partitioning. This index maximum value is 1, corresponding to a perfect matching between the two labeled clusters and

their labeled data. 0 corresponds to a random cluster assignment, and negative values are obtained if the clustering is orthogonal to an extent.

4.2 Results

The recorded data consisted of 1300 motions, performed by 13 different subjects. 11 subjects were right-handed, and 2 were left-handed. For the clustering, different sets of joints have been considered: hand (H), forearm (FA), arm (A), these body parts being the most solicited during the movement. The computed descriptors were: Speed Norm (SN), Speed value in x, y and z (Sxyz), Speed directions in x, y, and z (SDxyz), and Speed Norm and directions in x, y and z (SNDxyz). The precision (*P*), recall (*R*), F1-score (*F1*) and Adjusted Rand Index (*ARI)* are given for *k=2*, as it corresponds to the ground truth. The Average Silhouette Score (*ASS*) is also given for *k=2*, as it is the k value that gives the best value in most of the case (the *ASS* values show non-significant variations for other k values when k=2 does not give the best *ASS* values). The clustering was performed on: (i) the mixed data (left and right-handed together) (ii) left-handed data only and (iii) right-handed data only. Table 1 shows the obtained results. F1-score, *ASS* and *ARI* values slightly decreased when joints were added to the dominant hand, meaning that the dominant hand was the most important joint for this case. The highest *ASS* scores were obtained for speed values along the three axes, in the right-handed (0.73) and mixed (left and right-handed) data (0.54). Left-handed best *ASS* values are for the speed norm values (0.41), yet they are lower than the right-handed and mixed data *ASS* values for the same data (0.42 and 0.48). The *ARI* stayed close to 0, regardless of the joints and descriptors combination (ranging between 0.05 and 0).

Table 1: clustering metrics for various joint combinations for the Bottle Flip Challenge experiment.

Joints	H					H, FA					H, FA, A				
Metric	ASS	P	R	F1	ARI	ASS	P	R	F1	ARI	ASS	P	R	F1	ARI
	Left and Right-Handed														
SN	0.48	0.25	0.33	0.29	0.04	0.44	0.25	0.33	0.29	0.04	0.43	0.25	0.33	0.29	0.04
Sxyz	**0.54**	0.18	0.67	0.3	0.05	**0.52**	0.27	0.32	0.29	0.05	**0.51**	0.18	0.68	0.29	0.05
Sdxyz	0.24	0.21	0.53	0.3	0	0.27	0.25	0.25	0.25	0.04	0.22	0.18	0.72	0.27	0.04
SNDxyz	0.21	0.18	0.47	0.3	0	0.27	0.25	0.26	0.26	0.04	0.22	0.26	0.28	0.27	0.04
	Left Handed														
SN	**0.41**	0.39	0.39	0.39	0.02	**0.42**	0.38	0.39	0.39	0.01	**0.41**	0.31	0.61	0.39	0.01
Sxyz	0.35	0.32	0.57	0.39	0	0.34	0.32	0.57	0.39	0	0.33	0.35	0.43	0.39	0
Sdxyz	0.31	0.34	0.48	0.4	0	0.27	0.34	0.54	0.39	0	0.23	0.34	0.48	0.4	0
SNDxyz	0.27	0.34	0.49	0.4	0	0.25	0.33	0.48	0.39	0	0.22	0.34	0.52	0.41	0
	Right Handed														
SN	0.42	0.18	0.29	0.22	0	0.36	0.17	0.28	0.21	0	0.34	0.17	0.28	0.21	0
Sxyz	**0.73**	0.19	0.12	0.15	0.01	**0.71**	0.19	0.12	0.15	0.01	**0.71**	0.19	0.12	0.15	0.01

Sdxyz	0.28	0.15	0.45	0.28	0	0.2	0.16	0.49	0.27	0	0.26	0.19	0.13	0.15	0.01
SNDxyz	0.26	0.16	0.45	0.28	0	0.19	0.19	0.52	0.27	0	0.26	0.17	0.87	0.15	0.01

4.3 Discussion

The combination of the speed vectors in each axis is a good separation criterion, as suggested by results shown in section 4.2. The best ASS values were obtained for the descriptors extracted from the dominant hand, suggesting that other body parts only add noise. This can be partially explained by the fact that every joint motion is related to the other, and that the hand movement is the one with the widest range of values (in terms of speed).

While the *ASS* had an acceptable value (*ASS* ≈ 0.5) for the mixed data, better results were obtained when right-handed and left-handed people are separated (*ASS* ≈ 0.75). The acquisition problems of the suite can explain this phenomenon (and are discussed below in this section). In terms of relative distance, the most discriminant features were the maximum speed value, in both Z (forward) and Y (upward) directions (regarding to the subject), as seen in Table 2.

Table 2 : Relative distance of the clusters centroids, for the right hand, with the speed directions in x, y, and z, for *k=2*.

	Beginning	Maximum	End
X (Side)	0.0398	0.5071	0.0110
Y (Upward)	0.0415	1.7497	0.0998
Z (Forward)	0.0847	2.0477	0.0536

The clusters were indeed separable; however, the ARI stayed close to 0 for every case (*max(ARI)* ≈ *0.05*), indicating a random cluster assignment. That means that the obtained clusters cannot be related to the outcome of the throw. Consequently, the current descriptors (speed, acceleration and direction) with the proposed separation model are uncorrelated from the degree of success of the task. One can argue that, the considered task itself does not present a significant variation from one throw to another, in terms of speed and acceleration. Furthermore, the computed descriptors all rely on speed or acceleration, and that can possibly limit the variability of the results. Other high-level descriptors exist (Larboulette & Gibet, 2015), and could be used to analyze the motions. For example, the jerk (rate of change of the acceleration during the motion) can give an indication on how smooth the motion is, and the curvature, which is a measure of how fast a curve is changing through time, can give a more accurate information about the wrist rotation. The geometric descriptors, such as the rotation of joints through

time, and the displacement of the center of mass are also interesting values to consider.

In this experimentation, several problems arose. First, the distance between the subject and the table was not constant, as some people took a small step back before throwing. The table was also slippery, and the bottle slid on the table. Thus the distance between the subject and the impact point of the bottle cannot be measured with consistency regarding the throws of all subjects.

The MOCAP suit limits the experiment to its sensors accuracy and their constraints for a good use, opposed to, for example, an infrared camera system. Having accurate rotation data of the wrist would be interesting, as it represents a crucial part of the motion. Furthermore, a frame-by-frame analysis showed that the data flow was not constant. The mandatory software, for getting the data, used some undocumented method to counterbalance the data loss, that creates the artifacts seen in Fig. 1.a. While

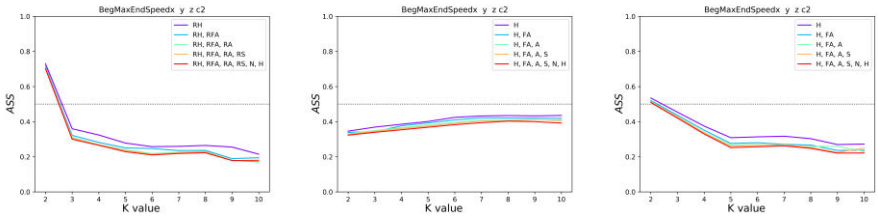


Figure 3: ASS score for various joint combinations and k ranging from 2 to 10 of (a) the right-handed subjects (b) the left-handed subjects (c) left and right-handed subjects together.

the pre-processing steps took care of these problems, nothing can ensure that the used method did not alter the initial data. Furthermore, the left side of the suit (from the shoulder to the hand) outputted noisy data. When the clustering was performed, mixing left-handed and right-handed data gave worse results than keeping only the right-handed subjects, due to noisy nature of the left-handed data (Fig.3). This noise was visible on the captured data, and it seems that the suit has difficulties to handle a capture of the full body.

4.4 Ball throwing

As the motion variability of the previous task can be discussed, another experiment was conducted to verify if the computed descriptors, combined with the k-means algorithm, can separate the motions according to the ground truth. In this experiment, a subject must throw a ball in one of two bins, placed in a line front on him (one placed 2m (6,56 ft) from them,

another one placed 3.5m (11,48 ft) from them). The subject has to perform 100 throws, without any constraints about the throwing motion. For each throw, the degree of success of the throw, the bin aimed at, and the type of throw (*i.e.* basket type launch or bowling type launch) were recorded. In this experiment, only right-handed data and a sub-set of the suit sensors were used, to limit the artifacts. Having multiple labeling for each motion allows working on the degree of success, as well as the descriptors ability to discriminate different throw strategies. The same joints combinations, as well as the same metrics were used to evaluate the results. For each metric, the results are given for $k=2$, as it is the number of clusters that gives the best results. The results have shown that the *ASS* and *ARS* values stay the same as the first experimentation for the successful/failed labeling, indicating that the separation of the motions was still not feasible with the proposed method. However, the clustering gives good *ASS* and *ARI* scores (0.59 and 0.83 respectively) for the throwing type, with the norm, and "norm + directions" descriptors (Table 3), suggesting that the data were separable regarding this criterion. Adding joints other than the dominant hand does not (or marginally) improve the results.

Table 3: clustering metrics for various joint combinations for the ball throwing experiment.

Joints	H					H, FA					H, FA, A				
Metric	ASS	P	R	F1	ARI	ASS	P	R	F1	ARI	ASS	P	R	F1	ARI
	All data (ground truth = success / fail)														
SN	0.59	0.51	0.89	0.64	0.01	0.6	0.5	0.86	0.63	0	0.6	0.5	0.86	0.63	0
Sxyz	0.56	0.48	0.89	0.62	-0.01	0.54	0.48	0.89	0.62	-0.01	0.54	0.48	0.89	0.62	-0.01
Sdxyz	0.23	0.41	0.25	0.31	-0.01	0.23	0.54	0.77	0.64	0.04	0.24	0.55	0.77	0.64	0.05
SNDxyz	0.26	0.53	0.89	0.67	0.04	0.27	0.55	0.77	0.64	0.05	0.26	0.53	0.77	0.63	0.03
	All data (ground truth = closest / farthest bin)														
SN	0.59	0.65	1	0.79	0.21	0.6	0.64	0.98	0.78	0.19	0.6	0.64	0.98	0.78	0.19
Sxyz	0.56	0.54	0.88	0.67	0.01	0.54	0.54	0.88	0.67	0.01	0.54	0.54	0.88	0.67	0.01
Sdxyz	0.23	0.65	0.34	0.45	0.02	0.23	0.68	0.86	0.76	0.2	0.24	0.69	0.86	0.77	0.22
SNDxyz	0.26	0.68	0.98	0.8	0.26	0.27	0.71	0.88	0.79	0.26	0.26	0.69	0.88	0.77	0.22
	All data (ground truth = throwing type)														
SN	0.59	0.87	0.95	0.91	0.83	0.6	0.83	0.95	0.89	0.79	0.6	0.83	0.95	0.89	0.79
Sxyz	0.56	0.06	0.05	0.05	-0.09	0.54	0.06	0.05	0.05	-0.09	0.54	0.06	0.05	0.05	-0.09
Sdxyz	0.23	0.08	0.1	0.09	-0.07	0.23	0.54	0.95	0.69	0.39	0.24	0.53	0.95	0.68	0.37
SNDxyz	0.26	0.74	0.95	0.83	0.68	0.27	0.55	1	0.71	0.42	0.26	0.56	0.95	0.7	0.42

4.5 Discussion

These two experiments allowed us to evaluate a clustering approach for the automatic analysis of motions with the MLA platform. With the

considered set of joints, kinematic descriptors and k-means algorithm, it was possible to obtain a good separation of motions regarding some observable properties of these motions. In our case, the throwing type could be detected thanks to the kinematic properties. It means that, for the considered task, it is possible to obtain multiple clusters corresponding to different throwing strategies. Analyzing which descriptors are the most discriminant for each cluster can give a hint about the different strategies used by the learner, thus leading to the determination of multiple learners profiles. However, it was not possible to achieve an acceptable separation corresponding to the degree of success of the task. Having such a separation would have allowed to automatically determine if a motion was within the acceptable range or not. The lack of experts in the Bottle Flip Challenge field, and thus the lack of evaluation criterion of the gesture itself, was also a hindrance for the choice of descriptors and the analysis of the motion.

In both of these experiments, the analysis is binary: the motion is either in one group or the other (successful/failed, throwing from above/from below, etc.). In a real motion learning context, it is not always possible to separate the results with only two categories. Moreover, the expert does not take part in the analysis process: their knowledge is only used in the selection of the relevant descriptors. While an autonomous system can be useful, the goal is to provide a set of tools to help the experts in their motion analysis task. The feedback system developed to answer these requirements is presented in the next section.

5. FEEDBACK SYSTEM

The next step of this work consisted in developing a TEL environment able to assist the expert in their motion analysis and advise task. Since multiple experts can have different view points about the properties of the targeted motion to learn, the term " targeted motion " will be used in this section to designate the learning objective.

The system must give: (i) advices to the learner about specific modalities of their gestures and (ii) a visualization of the main flaws or lacks of the learner motion. This system requires the expert to record (i) some targeted motions (at least a dozen), and (ii) some non-acceptable motions for each identified flaw. The more data for each group, the better the identification of acceptable motion groups and their features in the next step will be. The expert is then asked to designate one or more motion descriptors for each mistake, in order to be able to extract these motion descriptors from the expert data. The descriptor specification is made as follows:

- The name of the motion's flaw
- The used descriptor(s) for this flaw
- The joint(s) on which this/these descriptors(s) will be computed, along with the dominant hand of the learner if relevant (e.g. if "right" must be replaced with "left" in the joints name if the person is left-handed)

An example of such a specification can look as follows:

- Flaw: leaning forward when throwing
    - Descriptor: mean speed
        - Joint : left shoulder, dominant hand side: no
        - Joint: right shoulder, dominant hand side: no
- Flaw: elbow moving during the throw
    - Descriptor: mean speed
        - Joint: left arm, dominant hand side: yes
        - Joint: left shoulder, dominant hand side: yes

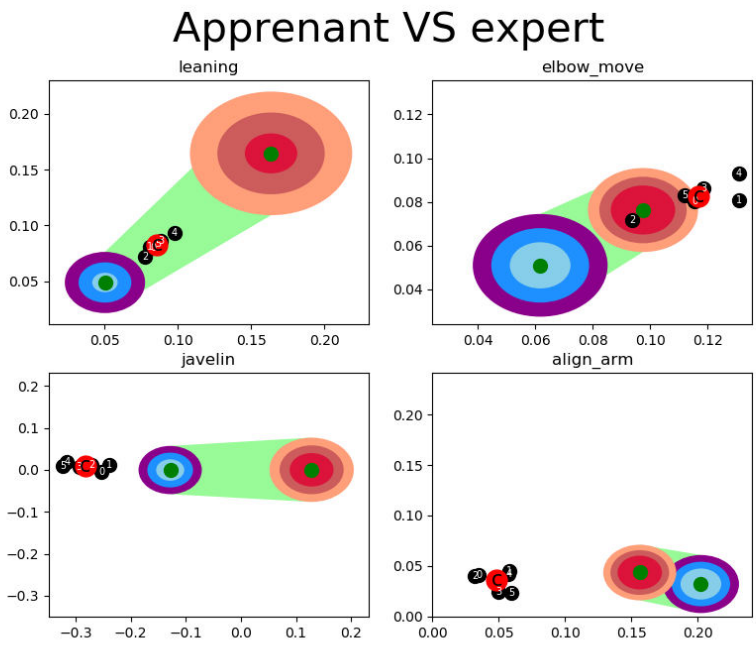
Descriptors values can be normalized, in order to have a consistent scale when evaluating the importance of the fault compared to another one. These data are then used in a clustering process, in order to obtain two groups for each identified flaw, one corresponding to the targeted motions, the other to the non-acceptable motions. A naïve approach would assign each data to the corresponding label, *i.e.* acceptable or not. However, this makes the assumption that the expert data are separable regarding this labeling. In practice, when the expert is recording multiple motions with mistakes made deliberately, a self-correction can appear. Indeed, the expert tends to unconsciously correct their motion. This can lead to outliers, *i.e.* flawed motions being more similar to acceptable ones. The clustering phase allows putting these motions into the group they truly belong to, without manually delete those outliers. Consequently, this method can produce overlapping groups. In this case, the acceptable motions are not sufficiently different from the non-acceptable ones, or the used descriptors are not significant to represent the motion flaw.

The learner data are then compared to the expert ones. In order to give relevant feedback regarding the most important flaws of the learner motion, this comparison is made by computing the Euclidean distance of the projection of the mean of the learner data point on the line that goes through each cluster center. Let c_g be the centroid of the targeted motion cluster, c_b

the centroid of the non-acceptable motion cluster, and c_a the centroid of the learner motions. We define

$$A = y_{c_b} - y_{c_g}$$
$$B = x_{c_b} - x_{c_g}$$
$$C = (x_{c_g} * y_{c_b}) - (x_{c_b} * y_{c_g})$$

The distance *D* is then defined as:



$$D = \frac{A * x_{c_a} + B * y_{c_a} + C}{\sqrt{A^2 + B^2}}$$

The use of this distance is based on the hypothesis that if the projection of the learner mean data on the aforementioned line is located inside the trapezoid linking the two expert motions groups (fig. 4), this flaw is more easily correctable than if it is outside of this trapezoid. The system then takes

Figure 4 : an example of the visual feedback proposed by the system. There is one visualization for each flaw. The blue group consists of the acceptable motions, while the red one consists of the flawed motions. The red point labeled "C" is the centroid of the learner motion.

the two most prominent flaws (in terms of the distance *D*) and highlights the

two mistakes that the learner must correct before anything else. This requires that the expert write down, for each flaw, at least one relevant advice to correct the gesture.

An experiment has been conducted in order to validate the proposed evaluation process. The goal was to throw darts , aiming at the center of a target, with respects to the sport official rules, including the target size, the darts length and mass, and the throwing distance (2.37 meters). An interview conducted with an expert of the darts game allowed us to find 4 majors flaws usually found in the beginner motions:

- *Leaning*: the learner leans towards the target when throwing a dart, resulting in the dart landing lower than expected
- *Elbow move*: when throwing a dart, the elbow moves instead of only rotating, which leads to a less controlled motion
- *Javelin*: the learner's arm goes next to (or even behind) their head, instead of staying in front of the head during the throw
- *Align arm*: the arm tends to go toward the center of the body (to the left for a right-handed person, and *vice versa*)

These 4 flaws can be detected in the motion data. Each of these flaws is not exclusive, *i.e.* a beginner can perform a motion with several flaws. In addition, other flaws exist. Nevertheless, they would require other capture devices and data to detect them. For example, the moment the dart was released by the learner can be detected with an appropriate infrared capture system following the dart motion, thanks to some reflectors on it. Since we aim to study an evaluation process only based on body-motion data, we only considered flaws that can be detected with the analysis of human movement with the above-mentioned capture suit in this study.

45 subjects were separated into 3 different groups, according to the use of the advice system: (group 1) the advices were given by the expert based only on their observation, (group 2) the advices were given only by the system and (group 3) the advices were given by the expert using their observation and the evaluation of the system to refine their analysis. The distance to the center of the target, as well as the distance between each motion to the centroid of the cluster for each flaw were noted. Each subject had to throw 36 darts, divided into 4 series of 9 throws each. Between each series, the system can give two feedbacks (groups 2 and 3) and can be used to visualize the learner data (group 3). The preliminary results show that there is an improvement in the motion shape of the learner (*i.e.* correction of the 4 flaws) for each group, without getting a significant difference from one group to another. However, no significant improvement was obtained for the distance of the darts to the center of the target. This can be explained by the

fact that the user will focus on the imitation of the targeted motion shape to the detriment of the throw accuracy. In addition, this shape can strongly differ from the initial motion of the user. Consequently, it seems that not enough throws are made by the learners to both improve the accuracy of the throw and the motion shape. Furthermore, a real learning situation would last longer.

6. CONCLUSION AND PERSPECTIVES

A new approach regarding the analysis of 3D motions was presented in this chapter. The goal is to give a method to analyze the motion, through explainable descriptors extracted from it, leading to personalized feedback given to the learner in order to improve their motion. After acquiring and processing the motion data, some descriptors based on speed, acceleration and direction were extracted. These descriptors were then used in a clustering process, in order to find different explainable types of motions. This approach relied on three hypotheses: (i) it is possible to separate the motions into explainable clusters, (ii) it is possible to obtain partitions corresponding to the degree of success of the task and (iii) it is possible to use clustering methods in order to create an interactive TEL system assisting the expert in its evaluation and advising task. While the second objective did not reach the expectations, the results of the first objective showed that the separation of the clusters is indeed possible, validating this hypothesis, and the used descriptors (with the proposed method for the first two tasks presented in this study) in terms of discriminant features.

The computation of more descriptors is planned, as the current ones may be limited, regardless of the application context. Most of the high-level descriptors used in various studies about human motions are a combination of multiple low-level ones based on kinematic, dynamic and geometric properties (Larboulette & Gibet, 2015). It would be possible to propose a template language in order to allow the user to specify their own descriptors. This would allow computing a predefined set of descriptors for every motion from a combination of low-level descriptors. As the data are made of time series, the use of the Dynamic Time Warping (DTW) algorithm, computing a similarity distance between the trajectory of two motions (Morel, 2017), would provide another measure, giving inter and intra-clusters information about the motions. Future work will also focus on performing recursive clustering on obtained clusters, in order to find if the motions, in each

cluster, are separable according to the degree of success of the task or other features.

The feedback system can be improved in multiple ways. As an engineering perspective, a Graphical User Interface (GUI) will be developed to tweak the parameters of the different phases and algorithms (*e.g.* motion segmentation, clustering parameters, etc.). Regarding the clustering phase on the expert data, it would be possible to automatically detect which data points are the furthest from the cluster center its assigned to, in order to delete them to reduce the overlapping effect between motion groups. The system selects the advice to give (*i.e.* the most important mistakes to correct). However, this choice does not consider: (i) the severity of the flaw and (ii) the specificity of each flaw (*e.g*. if the arm is not aligned towards the target, there is no indication about the side causing the problem). A comparison of the differences of each descriptor between the expert and the learner data could provide a hint and lead to a more precise feedback.

REFERENCES

Chang, C.-Y., Chang, C.-W., Zheng, J.-Y. & Chung, P.-C., 2013. Physiological emotion analysis using support vector regression. *Neurocomputing,* Issue 122, pp. 79-87.

Chan, J. C. P., Leung, H., Tang, J. K. T. & Komura, T., 2011. A Virtual Reality Dance Training System Using Motion Capture Technology. *IEEE Transactions on Learning Technologies,* Volume 4, pp. 187-195.

Couland, Q., Hamon, L. & George, S., 2018. Enhancing Human Learning of Motions: An Approach Through Clustering. *European Conference on Technology Enhanced Learning*.

Gu, Y. & Sosnovsky, S., 2014. Recognition of Student Intentions in a Virtual Reality Training Environment. *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces*, pp. 69-72.

Hachaj, T. & Marek R., O., 2015. Full Body Movements Recognition - Unsupervised Learning Approach with Heuristic R-GDL Method. *Digital Signal Processing,* Volume 46, pp. 239-252.

Huang, J., Zhou, W., Li, H. & Li, W., 2015. Sign Language Recognition using 3D convolutional neural networks. *Multimedia and Expo (ICME), 2015 IEEE International Conference on,* pp. 1-6.

Kapsouras, I. & Nikolaidis, N., 2014. Action recognition on motion capture data using a dynemes and forward differences representation. *Journal of Visual Communication and Image Representation,* 25(6), pp. 1432-1445.

Kobayashi, Y., 2007. The EMOSIGN - analyzing the emotion signature in human motion. *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on,* pp. 1171-1176.

Larboulette, C. & Gibet, S., 2015. A Review of Computable Expressive Descriptors of Human Motion. *Proceedings of the 2Nd International Workshop on Movement and Computing,* pp. 21--28.

Lokaiczyk, R., Faatz, A., Beckhaus, A. & Goertz, M., 2007. Enhancing Just-in-Time E-Learning Through Machine Learning on Desktop Context Sensors. *Modeling and Using Context: 6th International and Interdisciplinary Conference, CONTEXT 2007, Roskilde, Denmark, August 20-24, 2007. Proceedings,* pp. 330-341.

Lui, Y. M., O'Hara, S. & Draper, B. A., 2011. Unsupervised Learning of Humain Expressions, Gestures, and Actions. *Face and Gesture 2011,* pp. 1-8.

Maes, P.-J., Amelynck, D. & Leman, M., 2012. Dance-the-Music: an educational platform for the modeling, recognition and audiovisual monitoring of dance steps using spatiotemporal motion templates. *EURASIP Journal on Advances in Signal Processing,* Issue 1, p. 35.

Markowska-Kaczmar, U., Kwasnicka, H. & Paradowski, M., 2010. Computational Intelligence for Technology Enhanced Learning. *Intelligent Techniques in Personalization of Learning in e-Learning Systems,* pp. 1-23.

Mingliang, X. et al., 2019. Personalized training through Kinect-based games for physical education. *Journal of Visual Communication and Image Representation,* Volume 62, pp. 394-401.

Morel, M., 2017. *Multidimensional time-series averaging: application to automatic and generic evaluation of sport gestures,* s.l.: s.n.

Ng, A., 2016. *CS229 - Machine Learning course, Lecture N 19: Stanford Engineering Everywhere, Stanford University.* [Online] Available at: https://see.stanford.edu/Course/CS229 [Last access: 2016].

Nunes, J. F. & Moreira, P. M., 2016. *Handbook of Research on Computational Simulation and Modeling in Engineering.* s.l.:s.n.

Patrona, F., Chatzitofis, A., Zarpalas, D. & Daras, P., 2018. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognition,* Volume 76, pp. 612-622.

Rousseeuw, P. J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* Volume 20, pp. 53-65.

Struyf, A., Hubert, M. & Rousseeuw, P., 1997. Clustering in an Object-Oriented Environment. *Journal of Statistical Software, Articles,* 1(4), pp. 1--30.

Yoshinaga, T. & Soga, M., 2015. Development of a Motion Learning Support System Arranging and Showing Several Coaches' Motion Data. *Procedia Computer Science,* Volume 60, pp. 1497-1505.

Yu, Z. & Lee, M., 2015. Human motion based intent recognition using a deep dynamic neural model. *Emerging Spatial Competences: From Machine Perception to Sensorimotor Intelligence*, Septembre, pp. 134-149.

Zhou, H. & Hu, H., 2008. Human motion tracking for rehabilitation - A survey. *Biomedical Signal Processing and Control,* pp. 1-18.

AUTHOR INFORMATION

Quentin Couland
Le Mans University, LIUM - EA 4023
Avenue Olivier Messiaen, 72085 Le Mans, Cedex 9, France

Telephone number: 33 2 43 59 21 38
Email address: quentin.couland@univ-lemans.fr
Website: https://lium.univ-lemans.fr/en/team/quentin-couland/

Quentin Couland is a PhD student and junior lecturer at the LIUM (IEIAH team) since the 1st September 2015. He studied machine learning (supervised/unsupervised learning), signal and data processing (advanced signal processing, data compression, image compression), and documents processing. His thesis' goal is to develop a Technology Enhanced Learning (TEL) environment allowing a student to improve their motion learning, by using their motion data, with machine learning techniques (clustering).

Ludovic Hamon
Le Mans University, LIUM - EA 4023
Avenue Olivier Messiaen, 72085 Le Mans, Cedex 9, France
Telephone number: 33 2 43 59 49 01
Email address: ludovic.hamon @univ-lemans.fr
Website: https://lium.univ-lemans.fr/en/team/ludovic-hamon/

Since 2014, Ludovic HAMON is associate professor at "IUT de Laval", Le Mans Université, France. In July 2011, he received his PhD in automation and applied computer science at the LARIS laboratory (EA 7315). He currently conducts his research at the LIUM laboratory (computer science laboratory of Le Mans Université). His main fields of interest are: the manual or automatic analysis of data coming from new interaction paradigms in VEHL (Virtual Environment for Human Learning); modelling and analysis of human motions in learning situations; and contributions of virtual and augmented environments in the same context.

Sébastien George
Le Mans University, LIUM - EA 4023
Avenue Olivier Messiaen, 72085 Le Mans, Cedex 9, France
Telephone number: 33 2 43 59 49 16
Email address: sebastien.george@univ-lemans.fr
Website: https://lium.univ-lemans.fr/en/team/sebastien-george/

Sébastien George is Full Professor of Computer Science since 2013 at Le Mans University, in France. He teaches at the Institute of Technology of Laval. He is at the head of the team IEIAH (working in the field of Technology Enhanced Learning) at LIUM Laboratory. He received the PhD degree in computer science in 2001. Then he did a postdoctoral fellowship at the TeleUniversity of Quebec in Canada, before joining INSA Lyon in 2002.

He is the co-author of more than 180 publications in scientific books, journals and conferences. He is the editor-in-chief of the journal STICEF (Sciences and Technologies of Information and Communication for Education and Training). His major fields of interest are computer supported collaborative learning, authoring tools and assistance to human tutoring. He is particularly interested in applications integrating innovative Human Machine Interactions in the context of education and training (serious games, mixed reality).