Agents Behavior Semi-automatic Analysis through Their Comparison to Human Behavior Clustering

Kévin Darty, Julien Saunier, Nicolas Sabouret

▶ **To cite this version:**

Kévin Darty, Julien Saunier, Nicolas Sabouret. Agents Behavior Semi-automatic Analysis through Their Comparison to Human Behavior Clustering. 14th International Conference on Intelligent Virtual Agents (IVA 2014), Aug 2014, Boston, MA, United States. pp 154-163, 10.1007/978-3-319-09767-1_18 . hal-01062385v2

HAL Id: hal-01062385
https://hal.science/hal-01062385v2

Submitted on 23 Sep 2014

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

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

Agents behavior semi-automatic analysis through their comparison to human behavior clustering

Kévin Darty[1], Julien Saunier[2], and Nicolas Sabouret[3]

[1] Laboratory for Road Operations, Perception, Simulators and Simulations, French Institute of Science and Technology for Transport, Development and Networks
[2] Computer Science, Information Processing and Systems Laboratory, INSA of Rouen
[3] Computer Sciences Laboratory for Mechanics and Engineering Sciences, CNRS

Abstract. This paper presents a generic method to evaluate virtual agents that aim at reproducing humans behaviors in an immersive virtual environment. We first use automated clustering of simulation logs to extract humans behaviors. We then propose an aggregation of the agents logs into those clusters to analyze the credibility of agents behaviors in terms of capacities, lacks, and errors by comparing them to humans ones. We complete this analysis with a subjective evaluation based on a questionnaire filled by human annotators to draw categories of users, making their behaviors explicit. We illustrate this method in the context of immersive driving simulation.

Keywords: Virtual autonomous agent, virtual environment, behavior analysis, clustering and aggregation, logs explicitation

1 Introduction

Intelligent virtual agents (*IVAs*) are used in several fields such as crowd simulation [1] and virtual human listener [12]. In these simulations, agents have to produce realistic behaviors. The notion of behavior can cover different views, from low level actions (*e.g.* action units on human face [6]) to complex emerging movements in crowds [1]. One specific aspect of *IVAs* is that they interact directly with human users in virtual environments (*VEs*). In this context, providing realistic behavior is a key issue to avoid breaking immersion in the *VE* [10].

In the domain of *IVAs*, several studies have already addressed the questions of believability or credibility of *IVAs* behaviors. For instance, Campano et al. [4] proposed evaluation methods for affective models. Pelachaud et al. [14] proposed a credibility evaluation of the agent affective behavior model. These methods rely on evaluation studies using participants judgment of the agent behavior credibility. Only few research rely on "objective" analysis of simulation data, and are mostly coming from the multi-agent systems (*MAS*) domain (*e.g.* [2]) in which the interaction context is very different.

We propose a method for the analysis of the agents credibility that combines human expertise and simulation logs analysis. We consider the specific case of

agents aiming at reproducing human behaviors in an immersive *VE*. We propose to analyze the agents behaviors in terms of capacities, lacks, and errors with respect to humans. First, human participants act in the *VE*. Their behavior is logged and analyzed using objective methods from *AI*. *IVAs* are then evaluated by comparing their behaviors with the human participants ones in the same situation. We complete this analysis with a Human Sciences evaluation.

The next section presents related works in the domain of objective and subjective evaluation that was used in our research. Section 3 presents our method based on data clustering and aggregation algorithms. Section 4 illustrates the potential of this method in the context of a driving simulation.

2 Related works

In our work, we want to evaluate the agents behavior at a strategic level: we consider that the behavior is based on a choice of tactics and that it evolves according to the dynamics of the environment, and to the mental state of the person [9]. For this reason, we will distinguish action logs, which are only traces of the behavior, from the behavior itself as it can be analyzed by a human. The work presented in this section relate to this level of behavior.

2.1 Objective Approach

Analysis of simulation data for the evaluation of the behavior credibility is widely used in the field of *MAS*. It consists in verifying through quantitative data that agents behave as in a "real" situation. This validation method is generally used at the macroscopic level [1]. However, having a valid collective behavior does not imply that the individuals behaviors are realistic. This is the reason why other researchers proposed to focus on the validation at the microscopic level. Caillou [2] showed that data analysis is more complex at this scale and cannot be done directly on the simulation logs due to the semantic gap between the noisy raw data and the sought behaviors. Field experts are generally consulted to determine high-level variables that describe the behavior to be analyzed through the data. An automated clustering algorithm can then be used to classify the agents behaviors [2]. For generic methods, as one does not have any information on the domain-specific behaviors, they are unpredefined. Therefore, the clustering method has to be unsupervised with a free number of clusters.

It is also worth noting that, in the domain of interaction, Delaherche and Chetouani proposed behavior traces clustering methods for the study of synchrony [8]. However, their goal is not to evaluate the realism of an *IVA*.

The main limitation of this approach is that while it allows to see the difference between categories of behaviors, extracted from the logs (*i.e. behavior log clusters*), it does not provide information beyond the used variables: it cannot give a meaning to the obtained clusters. On the contrary, the subjective approach, which relies on a higher-level analysis, offers this possibility.

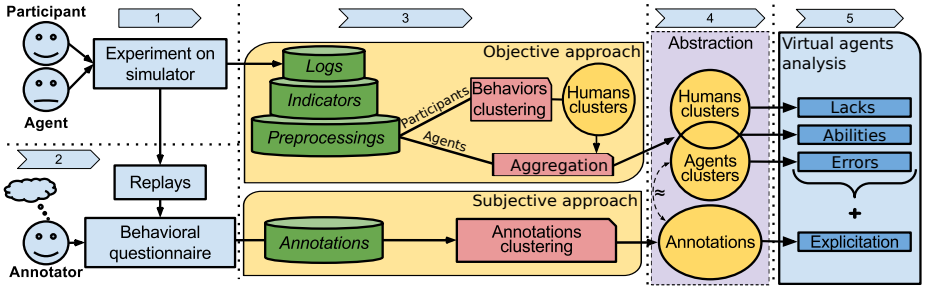


Fig. 1: Behavior analysis and evaluation method.

2.2 Subjective approach

The subjective approach for the evaluation of behavior similarity with human beings has been widely used in the domain of *IVAs* [11,14]. It consists in evaluating immersion quality through questionnaires. When it comes to *IVAs* and behavior analysis, the studies focus on the behavioral credibility [13], *i.e.* the evaluation of how human and *IVA* behaviors appear similar.

In this approach, the *IVAs* are observed by the immersed participant. The strength of this subjective approach is to characterize the adopted behaviors via questionnaires and to catch high-level behaviors through human participants annotations. One can regroup these behaviors based on these high-level descriptions. We shall then speak of *annotated behavior clusters*.

However, it is difficult to process hundreds of agents with such a method, since it requires a strong involvement of human participants.

Both logs clusters and annotations clusters aim at evaluating the adopted behaviors. For this reason, objective and subjective approaches through simulation data analysis and human expertise, complement each other. In our method, we propose to combine them: we use automated data analysis and aggregation method to build behavior log clusters, and human observers fill out a questionnaire about the adopted behaviors to build annotated behavior clusters.

3 Behavior analysis and evaluation method

The method we propose is based on the combination of simulation logs analysis (objective part) and answers to a behavior questionnaire (subjective part). The simulation data are classified into behavior logs clusters. The behavior questionnaire allows us to define situation-specific users categories for both participants and agents. We then evaluate the agents by analyzing the behaviors logs clustering composition and make the behaviors explicit via the annotation clusters.

The general method is described in the Figure 1. It consists of 5 main steps: 1) collection of data in simulation and 2) annotation of these data, 3) data preprocessing and automatic clustering, then 4) clusters comparison, and finally 5) composition analysis and explicitation.

In the following subsections, we present these different steps.

Steps 1 and 2: experiments

We use the human behaviors as the reference behaviors to analyze agents ones. This is why the first step of our method is the collection of quantitative data about human participants from an immersive simulation in a *VE*. We also produce new simulations in which the participant is replaced by an agent (*i.e.* placed in an identical situation to that presented to participants) and then collect the very same logs as for the participants. Different types of agents are generated by exploring the parameter space such as normativity, experience, decision parameters... We call *main actors* both humans and *IVAs* gathered together. The raw data from main actors experiments in the simulator are called *simulation logs*.

The second step is the subjective evaluation of the main actors behaviors. A different set of participants annotates the video replays of all main actors simulations via the behavior questionnaire. This step produces a set of annotations.

Step 3a: logs clustering

The first objective of the third step is to compare participants behaviors and agents behaviors so as to report on the capability of the agents to reproduce human behaviors. Our goal is to compute behavior categories that serve as abstractions to the logs: each cluster shall regroup different logs representative of the same type of high level behavior (see Figure 1).

To begin with, field experts are consulted to identify important domain-specific indicators. The values for these indicators are computed from the simulation logs and then turned into scalars within a series of preprocessing as described on the left part of figure 2.

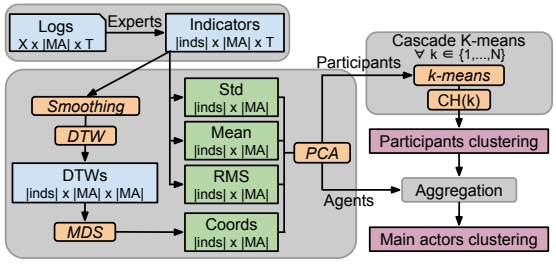


Fig. 2: Logs preprocessing, clustering, and aggregation with the time (T), the number of variables (X), of indicators ($|inds|$), and of main actors ($|MA|$).

The reason for this pre-processing it that most of the indicators are temporal and an identical behavior adopted by several main actors can occur with a temporal offset. In order to take this into account, we use the Dynamic Time

Warping [16] algorithm (DTW) which computes mutual distances. The K-means algorithm need the data to be in a dimensional space for which axes are perpendicular. So as to include DTW similarities as new variables describing the main actors, we use a Multi-Dimensional Scaling algorithm (MDS) to place each main actor in a dimensional space. We then process a *Principal Component Analysis* to project the data on a hyperplane with non-correlated axes. The outcome of this process is a set of indicators. In order to draw the behavior categories, we use an automatic unsupervised clustering algorithm on these indicators. The number of clusters is not defined a priori: we apply the *K-means* clustering algorithm on the participants indicators, and use the *Variance Ratio Criterion* [3] to determine the appropriate number of clusters K.

This first part of step 3 has already been published and more information can be found in [7]. In this paper, we add a further building block to this method: the aggregation of agents to the participants clusters.

Step 3b: cluster aggregation

During the clustering process, the addition of a main actor modifies the clusters shape and may change the affectations. However, agents and humans should not be considered with the same view, since the humans behaviors represent the target to which we want to compare our agents behaviors. For this reason, we do not want agents to modify the humans clustering. In order to keep the human clusters intact, the k-means algorithm is applied on participants only. The agents are then aggregated to the fixed human clusters if close enough or classified into new agents clusters.

Our method works as follows. We define for each participants cluster $C_i \in \mathcal{C}$ a threshold t_i above which the agent a is considered as being too far from the centroid m_i to be aggregated. This threshold t_i is defined on each dimension (*i.e.* on each indicator ind) as the distance between the centroid m_i and the farthest participant p: $t_i^{ind} = \forall\ p,\ max(|m_i^{ind} - p^{ind}|)$.

In order to allow the aggregation of the near neighbors, we enlarge t_i by a percentage of the mean of all thresholds t_i, based on a tolerance rate ϵ: this allows to have singleton clusters (for which $t_i = 0$) attracting other main actors.

Each agent a is aggregated to the participants cluster C_i of which the centroid is the closest among those under the threshold t_i for each dimension. If some agent(s) did not aggregate to any participants cluster due to the thresholds, the first "remaining" agent creates its own cluster C_{k+1} which is added to the clusters set $\mathcal{C}$ so that remaining agents can aggregate to it. Similarly, each remaining agent tries to aggregate to one of the remaining agents clusters, following the same threshold rule as for human clusters, or creates its own cluster if this is not possible. Thus, as shown in Figure 2, we obtain a clustering composed of all the main actors.

Step 3c: annotation clustering

The second objective is to analyze the behaviors through annotations, following the subjective approach. We use the same methodology as for logs clustering: identification of key indicators, unsupervised clustering on those indicators for the human participants and aggregation of *IVAs* to the human clusters.

The subjective approach requires manual annotation of replays: when the number of *IVAs* increases, it becomes impracticable to annotate them all. Yet, under the hypothesis that agents aggregated to a logs cluster of participants should have been annotated in the same way as the participants of this cluster, it is still possible to make these agents behaviors explicit via the participants annotation. In this case, the combination of the objective approach and the subjective one allows us to compare any number of agents and any number of agent models with human participants. However, clusters composed only of agents can no longer be explicited. For this reason, in the experiment presented hereafter, and considering that we had only a limited number of agents, we used manual annotation of all main actors.

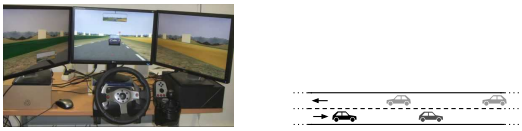
The second difficulty for annotation clustering is to choose the right set of indicators. We want these indicators to be both field-specific and situation-specific. In general, behavior questionnaires allow to characterize the participant general behavior, while the participant can adopt a specific (and different) behavior in a local situation. For this reason, as will be shown in the next section, we adapt domain-specific behaviors questionnaires to define the indicators for the situation annotation. Scale scores of questionnaires are calculated by adding the scale-related questions, and normalized between 0 and 1.

We then classify the participants scores using the *K-means* algorithm with the agents aggregation, which builds our annotation clusters.

Steps 4 and 5: clusters analysis

The fourth step of our method is the comparison of the two clusterings (logs clusters with annotations clusters). As both evaluate behaviors, having a strong similarity between them in terms of composition is a partial verification that the logs clustering is meaningful in terms of situation-specific user categories, and thus corresponds to task-related high-level behaviors. We evaluate the similarity between them (dashed arrow) with the *Rand Index* (*RI*).

The fifth and final step consists in analyzing the *IVAs*. It is possible to distinguish three cluster types in terms of participant and agent composition: 1) The clusters containing both humans and agents : They correspond to high-level behaviors that are correctly reproduced by the agents. 2) Those consisting of simulated agents only: They correspond to behaviors that were produced only by the agents. In most cases it reflects simulation errors, but it can also be due to a too small participants sample. 3) Those consisting of participants only: They correspond to behaviors that have not been replicated by the agents. Thus, they are either lacks in the agent model or due to a too small agents sample in the parameter space. We then combine this human-agent comparison with the



(a) The driving simulator device.

(b) Scenario of the driving experiment.

Fig. 3: Application to the study of driving simulation

annotation analysis to give explicit information (*i.e.* high-level characteristics) about those agents behaviors and about the missing behaviors if any.

4 Evaluation

This section illustrates our method with an application to the study of driving simulation, and then presents the data analysis and discusses the results.

4.1 Case study

We used the *ARCHISIM* road traffic simulator [5]. We want to evaluate the realism and credibility of the agents driving behaviors. To do this, the participants drive a car on a road containing simulated vehicles. The circuit (shown in Figure 3b) provides a driving situation on a single carriage way with two opposing lanes. It corresponds to about 1 minute of driving. The main actor encounters a vehicle at low speed on the right lane and oncoming vehicles on the left lane with increasing distances between them.

The *Driving Behavior Questionnaire* (*DBQ*) [15] collects data about drivers habits. In order to have a situation-specific questionnaire, we chose to base the annotation on the *DBQ* scales. The annotation questionnaire provides 5 *Likert*-type scales: *slips*, *lapses*, *mistakes*, *unintended* and *deliberate violations*. In addition, it supplies a scale related to the accident risk.

During the simulation we collect the main actors logs. We collect each 300 ms several variables such as the cap to the lane axis, the speed, and the topology. The road traffic experts chose both high-level indicators (*e.g.* the inter-vehicles distance, the time to collision, and the number of lane changings) and low-level variables (like speed, acceleration, and lateral distance).

The 22 participants of our driving simulation experiment are regular drivers aged from 24 to 59. Our experiment is carried out on a device comprising a steering wheel, a set of pedals, a gearbox and 3 screens (see Figure 3a).

Firstly, a test without simulated traffic is performed for the participant to get accustomed to the *VE*. Then, the participant performs the scenario, this time in interaction with simulated vehicles. It should be noted that as the behavior of simulated vehicles is not scripted, situations differ more or less depending on the main actors behavior. The data are then recorded for the processing phase and

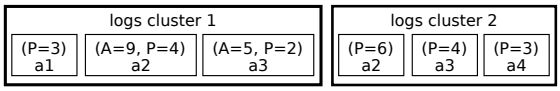


Fig. 4: Comparison of main actors between logs clustering (with participants P and agents A), and annotations clustering grouped together with the cluster numbers ($a\#$) being written just bellow the composition.

a video is made for the replay. Finally, 6 other participants fills the annotations questionnaire after viewing the video replays (22 participants and 14 agents).

4.2 Results

We have compared the logs clustering and the annotations clustering (see Figure 4). The Rand Index between the two clusterings is 0.51. There are 2 behavior log clusters and 4 behavior annotation clusters.

Logs cluster 1 contains 9 participants and all the 14 agents. The *number of lane changings* indicator value is 0 meaning that these main actors did not try to overtake the vehicle at low speed and preferred to follow it. As this cluster is composed of both agents and participants, it is therefore a capacity of the agent model to reproduce a human behavior which is to choose not to overtake.

Logs cluster 2 is only composed of participants. These 13 participants overtake the vehicle at low speed after the 2^{nd} or the 3^{rd} oncoming vehicle. As there are only participants, this human behavior can be considered as a lack in the agent model: the agents cannot choose to overtake as some human do.

Annotation cluster 1 contains 3 participants and no agent. The annotators consider that they are the more dangerous drivers with very high scores on each scale and especially on the *judgment* scale. Since no agent was considered that dangerous, and as the aim of the agents is to reproduce the most complete panel of human behaviors, there is a lack of unsafe behaviors in the model.

Annotation cluster 2 is composed of both participants and agents. They are annotated as very cautious drivers with the smallest scores on each scale. The space parameter of these agents ensure more respect for the highway code and smoother driving. The normative human behavior can therefore be considered as partly reproduced.

Annotation cluster 3 is a smaller cluster composed of participants and agents. The annotators considered them as ordinary drivers with medium scores. As for the previous cluster, the average behavior is reproduced in this situation.

Annotation cluster 4 is only composed of participants. It has some high scores on the specific *memory* and *judgment* scales. This behavior considered as slightly dangerous is also not reproduced by agents.

4.3 Discussion

In the *logs cluster 1*, the indicators were not able to distinguish the *annotations cluster 1* from the rest of the main actors. The judgment scale is very high

and a video replay shows that these participants tried to overtake several times unsuccessfully. Likewise, the participants of the *annotations cluster 4* were not separated from *logs cluster 2*. This might be due to the fact that they dared to overtake just after the second oncoming vehicle, which requires to pull back in a short time frame. However, our indicators did not detect that difference.

Several behaviors can be annotated in the same way. This is an issue to analyze the similarity between the annotations clustering and the logs clustering with the RI measure. A solution could be compute the RI on logs clustering for which logs clusters annotated in the same way will be merged into one cluster.

We have a significant similarity between annotations and logs clusterings, meaning that we are able to classify our logs data into high-level behavior clusters which are meaningful in terms of driving annotations. Nevertheless the two clusterings are not identical with regard to the clusters composition. This could be due to the few number of annotators. This problem may also come from the clustering algorithm which is a classic but basic one. However, we already tried other algorithms such as EM and HAC without better results. We have to test with time-series based algorithms. Also, the experts of the domain have to be consulted to understand what missing indicator could catch these differences.

The third type of cluster (which does not appear in this experiment) is composed of agents only. In that case, we can consider - as no participant adopted this behavior - that the agents behavior is inaccurate (*i.e.* is an error) and should be investigated further. The method has to be applied in other situations in order to verify this particular case.

5 Conclusion & perspectives

This paper presents a method to study *IVAs* behaviors through an experiment in a *VE*. This validation is original in coupling an objective analysis of the agents behaviors through simulation logs, with a subjective analysis, coming from Human Sciences, of the situation-specific user categories through an annotation done by participants. The objective analysis uses an unsupervised clustering algorithm applied on simulation logs in order to classify participants behaviors, and an aggregation method to compare agents behaviors to humans ones. This comparison allows us to evaluate the agents behaviors credibility in terms of capacities, lacks, and errors. It also provides an analysis of which *IVA* parameter space produces which perceived behavior. The method is generic for *VEs* where agents aim at reproducing human behaviors. When applied to a new domain, some of the tools have to be adapted, such as the choice of the behavior questionnaire which is domain-specific.

Our validation method was applied to the road traffic simulation. This experiment showed that the methodology is usable for mixed and complex *VEs* and that it is possible to obtain high-level behaviors from the logs via our abstraction.

Several tracks for further work remain to explore. On the clustering part, the evaluation of multiple time series based algorithms should help classifying

the temporal data. On the aggregation part, the automation of the parameter calibration will be beneficial to the agents aggregation accuracy.

Another research open issue is how the behaviors clustering evolve through multiple situations of a longer scenario, whether the participants clusters remain stable or change in number or composition.

References

1. Bosse, T., Hoogendoorn, M., Klein, M.C., Treur, J., Van Der Wal, C.N.: Agent-based analysis of patterns in crowd behaviour involving contagion of mental states. In: Modern Approaches in Applied Intelligence, pp. 566–577. Springer (2011)
2. Caillou, P., Gil-Quijano, J.: Simanalyzer: Automated description of groups dynamics in agent-based simulations. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3. pp. 1353–1354. International Foundation for Autonomous Agents and Multiagent Systems (2012)
3. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3(1), 1–27 (1974)
4. Campano, S., Sabouret, N., de Sevin, E., Corruble, V.: An evaluation of the cor-e computational model for affective behaviors. In: Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems. pp. 745–752. International Foundation for Autonomous Agents and Multiagent Systems (2013)
5. Champion, A., Éspié, S., Auberlet, J.M.: Behavioral road traffic simulation with archisim. In: Summer Computer Simulation Conference. pp. 359–364. Society for Computer Simulation International; 1998 (2001)
6. Courgeon, M., Martin, J.C., Jacquemin, C.: Multimodal affective and reactive character. In: 1st Workshop on Affective Interaction in Natural Environments (2008)
7. Darty, K., Saunier, J., Sabouret, N.: A method for semi-automatic explicitation of agent's behavior: application to the study of an immersive driving simulator. In: The 6$^{\text{th}}$ International Conference on Agents and Artificial Intelligence (ICAART 2014). pp. 81–91. SciTePress (2014)
8. Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: A survey of evaluation methods across disciplines. Affective Computing, IEEE Transactions on 3(3), 349–365 (2012)
9. Fisher, D., Rizzo, M., Caird, J.: Handbook of driving simulation for engineering, medicine, and psychology. CRC Press (2011)
10. Fontaine, G.: The experience of a sense of presence in intercultural and international encounters. Presence: Teleoperators and VEs 1(4), 482–490 (1992)
11. Gratch, J., Marsella, S.: Evaluating a computational model of emotion. Autonomous Agents and Multi-Agent Systems 11(1), 23–43 (2005)
12. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Intelligent Virtual Agents. pp. 125–138. Springer (2007)
13. Lester, J.C., Converse, S.A., et al.: The persona effect: affective impact of animated pedagogical agents. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 359–366. ACM (1997)
14. Pelachaud, C.: Modelling multimodal expression of emotion in a virtual agent. Phil. Trans. R. Soc. B: Biological Sciences 364(1535), 3539–3548 (2009)
15. Reason, J., Manstead, A., Stradling, S., Baxter, J., Campbell, K.: Errors and violations on the roads: a real distinction? Ergonomics 33(10-11), 1315–1332 (1990)
16. Salvador, S., Chan, P.: Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis 11(5), 561–580 (2007)