Marcello Pelillo
Edwin R. Hancock (Eds.)

LNCS 7005

Similarity-Based Pattern Recognition

First International Workshop, SIMBAD 2011
Venice, Italy, September 2011
Proceedings

Lecture Notes in Computer Science 7005

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Marcello Pelillo Edwin R. Hancock (Eds.)

Similarity-Based Pattern Recognition

First International Workshop, SIMBAD 2011
Venice, Italy, September 28-30, 2011
Proceedings

Springer

Volume Editors

Marcello Pelillo
Università Ca' Foscari, DAIS
Via Torino 155, 30172 Venice, Italy
E-mail: pelillo@dsi.unive.it

Edwin R. Hancock
The University of York
Heslington, York YO10 5DD, UK
E-mail: erh@cs.york.ac.uk

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-24470-4 e-ISBN 978-3-642-24471-1
DOI 10.1007/978-3-642-24471-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011937058

CR Subject Classification (1998): I.4, I.5, I.2.10, H.3, F.1, J.3

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Traditional pattern recognition techniques are intimately linked to the notion of "feature spaces." Adopting this view, each object is described in terms of a vector of numerical attributes and is therefore mapped to a point in a Euclidean (geometric) vector space so that the distances between the points reflect the observed (dis)similarities between the respective objects. This kind of representation is attractive because geometric spaces offer powerful analytical as well as computational tools that are simply not available in other representations. Indeed, classical pattern recognition methods are tightly related to geometrical concepts and numerous powerful tools have been developed during the last few decades, starting from the maximal likelihood method in the 1920's, to perceptrons in the 1960's, to kernel machines in the 1990's.

However, the geometric approach suffers from a major intrinsic limitation, which concerns the representational power of vectorial, feature-based descriptions. In fact, there are numerous application domains where either it is not possible to find satisfactory features or they are inefficient for learning purposes. This modeling difficulty typically occurs in cases when experts cannot define features in a straightforward way (e.g., protein descriptors vs. alignments), when data are high dimensional (e.g., images), when features consist of both numerical and categorical variables (e.g., person data, like weight, sex, eye color, etc.), and in the presence of missing or inhomogeneous data. But, probably, this situation arises most commonly when objects are described in terms of structural properties, such as parts and relations between parts, as is the case in shape recognition.

In the last few years, interest around purely similarity-based techniques has grown considerably. For example, within the supervised learning paradigm (where expert-labeled training data is assumed to be available) the well-established kernel-based methods shift the focus from the choice of an appropriate set of features to the choice of a suitable kernel, which is related to object similarities. However, this shift of focus is only partial, as the classical interpretation of the notion of a kernel is that it provides an implicit transformation of the feature space rather than a purely similarity-based representation. Similarly, in the unsupervised domain, there has been an increasing interest around pairwise or even multiway algorithms, such as spectral and graph-theoretic clustering methods, which avoid the use of features altogether.

By departing from vector-space representations one is confronted with the challenging problem of dealing with (dis)similarities that do not necessarily possess the Euclidean behavior or not even obey the requirements of a metric. The lack of the Euclidean and/or metric properties undermines the very foundations

of traditional pattern recognition theories and algorithms, and poses totally new theoretical/computational questions and challenges.

This volume contains the papers presented at the First International Workshop on Similarity-Based Pattern Recognition (SIMBAD 2011), held in Venice, Italy, September 28–30, 2011. The aim of this workshop was to consolidate research efforts in the area of similarity-based pattern recognition and machine learning and to provide an informal discussion forum for researchers and practitioners interested in this important yet diverse subject. The workshop marks the end of the EU FP7 Projects SIMBAD (http://simbad-fp7.eu) and is a follow-up of the ICML 2010 Workshop on Learning in non-(geo)metric spaces.

We believe that there are two main themes underpinning this research topic, which correspond to the two fundamental questions that arise when abandoning the realm of vectorial, feature-based representations. These are:

- How can one *obtain* suitable similarity information from data representations that are more powerful than, or simply different from, the vectorial?
- How can one *use* similarity information in order to perform learning and classification tasks?

The call for papers produced 35 submissions, resulting in the 23 papers appearing in this volume, 16 of which presented orally at the workshop and 7 in a poster session. The papers cover a wide range of problems and perspectives, from supervised to unsupervised learning, from generative to discriminative models, and from theoretical issues to real-world practical applications. In addition to the contributed papers, the workshop featured invited keynote talks by Marco Gori, from the University of Siena, Italy, Ulrike Hahn, from Cardiff University, UK, and John Shawe-Taylor, from University College London, UK. All oral presentations were filmed by Videolectures, and will be freely available on-line in due course.

We gratefully acknowledge generous financial support from the PASCAL2 network of excellence, and thank the International Association for Pattern Recognition (IAPR) for its sponsorship. We also acknowledge the Future and Emerging Technology (FET) Programme, of the 7th Framework Programme for Research of the European Commission, which funded the SIMBAD project, within which this workshop was conceived and of which was an outgrowth.

We would also like to take this opportunity to express our gratitude to all those who helped to organize the workshop. First of all, thanks are due to the members of the Scientific Committees and to the additional reviewers. Special thanks are due to the members of the Organizing Committee. In particular, Samuel Rota Bulò and Nicola Rebagliati managed the online review system and were webmasters, Furqan Aziz assembled the proceedings, and Veronica Giove provided administrative support.

Finally, we offer our appreciation to the editorial staff at Springer in producing this book, and for supporting the event through publication in the LNCS series. Finally, we thank all the authors and the invited speakers for helping to make this event a success, and producing a high-quality publication to document the event.

August 2011

Marcello Pelillo
Edwin Hancock

Organization

Program Chairs

Marcello Pelillo	University of Venice, Italy pelillo@dsi.unive.it
Edwin R. Hancock	University of York, UK erh@cs.york.ac.uk

Steering Committee

Joachim Buhmann	ETH Zurich, Switzerland
Robert Duin	Delft University of Technology, The Netherlands
Mario Figueiredo	Technical University of Lisbon, Portugal
Edwin Hancock	University of York, UK
Vittorio Murino	University of Verona, Italy
Marcello Pelillo (Chair)	University of Venice, Italy

Program Committee

Maria-Florina Balcan	Georgia Institute of Technology, USA
Manuele Bicego	Univeristy of Verona, Italy
Joachim Buhmann	ETH Zurich, Switzerland
Horst Bunke	University of Bern, Switzerland
Tiberio Caetano	NICTA, Australia
Umberto Castellani	University of Verona, Italy
Luca Cazzanti	University of Washington, Seattle, USA
Nicol Cesa-Bianchi	University of Milan, Italy
Robert Duin	Delft University of Technology, The Netherlands
Francisco Escolano	University of Alicante, Spain
Mario Figueiredo	Technical University of Lisbon, Portugal
Ana Fred	Technical University of Lisbon, Portugal
Bernard Haasdonk	University of Stuttgart, Germany
Edwin Hancock	University of York, UK
Anil Jain	Michigan State University, USA
Robert Krauthgamer	Weizmann Institute of Science, Israel
Marco Loog	Delft University of Technology, The Netherlands
Vittorio Murino	University of Verona, Italy

Elzbieta Pekalska	University of Manchester, UK
Marcello Pelillo	University of Venice, Italy
Massimiliano Pontil	University College London, UK
Antonio Robles-Kelly	NICTA, Australia
Volker Roth	University of Basel, Switzerland
Amnon Shashua	The Hebrew University of Jerusalem, Israel
Andrea Torsello	University of Venice, Italy
Richard Wilson	University of York, UK

Additional Reviewers

Marco San Biagio
Jaume Gibert Paola Piro
Nicola Rebagliati
Samuel Rota Bulò
Simona Ullo

Organization Committee

Samuel Rota Bulò (Chair)	University of Venice, Italy
Nicola Rebagliati	University of Venice, Italy
Furqan Aziz	University of York, UK
Teresa Scantamburlo	University of Venice, Italy
Luca Rossi	University of Venice, Italy

Table of Contents

Dissimilarity Characterization and Analysis

Generative Models of Similarity Data

Graph Based and Relational Models

Clustering and Dissimilarity Data

Applications

Spectral Methods and Embedding

Section-Wise Similarities for Clustering and Outlier Detection of Subjective Sequential Data

Oscar S. Siordia, Isaac Martín de Diego, Cristina Conde, and Enrique Cabello

Face Recognition and Artificial Vision Group, Universidad Rey Juan Carlos,
C. Tulipán, S/N, 28934, Móstoles, España
{oscar.siordia,isaac.martin,cristina.conde,enrique.cabello}@urjc.es

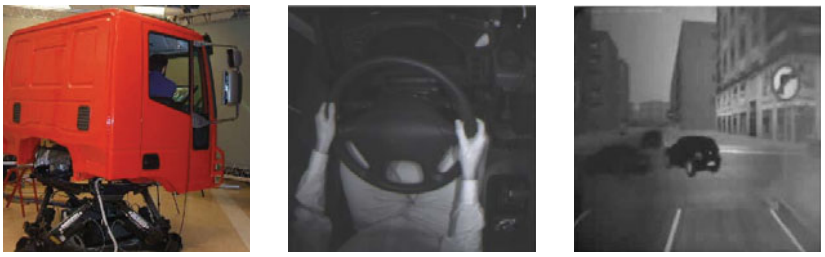
Abstract. In this paper, a novelty methodology for the representation and similarity measurement of sequential data is presented. First, a linear segmentation algorithm based on feature points is proposed. Then, two similarity measures are defined from the differences between the behavior and the mean level of the sequential data. These similarities are calculated for clustering and outlier detection of subjective sequential data generated through the evaluation of the driving risk obtained from a group of traffic safety experts. Finally, a novel dissimilarity measure for outlier detection of paired sequential data is proposed. The results of the experiments show that both similarities contain complementary and relevant information about the dataset. The methodology results useful to find patterns on subjective data related with the behavior and the level of the data.

Keywords: Subjective sequential data, Similarity, Clustering, Outlier.

1 Introduction

In the last few years, several representations of sequential data have been proposed, including Fourier Transforms [1], Wavelets [2], Symbolic Mappings [3] and, the most frequently used representation, Piecewise Linear Representation (see, for instance, [4,5,6,7]). Alternatively, the design of similarity measures for sequential data is addressed from a model-based perspective (see, for instance, [8,9]). In any case, the representation of the sequential data is the key to efficient and effective solutions. However, most of these representations imply sensitivity to noise, lack of intuitiveness, and the need to fine-tune many parameters [4]. In the present work, an alternative piecewise linear representation based on feature points is proposed. Similarity measures between sequential data is a common issue that has been treated in several ways. Usually, the statistical models fitted to the data are compared. Nevertheless, subjective sequential data are rarely considered. This kind of data corresponds to information collected from human opinions over a period of time. Although, it is not possible to successfully fit a unique model to all the data set since the changes on the level of the series usually respond to a great variety of factors, different model based approaches overcome this problem by employing one model per sequence [10].

M. Pelillo and E.R. Hancock (Eds.): SIMBAD 2011, LNCS 7005, pp. 61–76, 2011.



(a) Truck simulator (b) Driver top view (c) Road main view

Fig. 1. Truck simulator and sample frames of visual information acquired

The piecewise linear representation proposed in this work allows the definition of two similarity measures considering the behavior and the level of the sequential data, respectively. The proposed similarity measures were applied for clustering and outlier detection of a group of traffic safety experts' driving risk evaluations. Each expert provide two sequential risk evaluations of a simulated driving exercise. The data acquisition process was made as follows: a driving simulation exercise of ten minutes was recorded from a truck cabin simulator. Then, a group of 38 traffic safety experts were asked to evaluate the driving risk of the simulated exercise. One of the main objectives behind this project is to identify drivers' unsuitable behavior and lacks of attention.

The rest of the paper is organized as follows. Section 2 presents the process for the acquisition of the experts' evaluations. In Section 3, the piecewise linear representation algorithm for the linear segmentation of sequential data is developed. In Section 4, the section-wise similarities are defined. The evaluation of the performance of the proposed similarities on the experts' evaluations is presented in Section 5. Finally, Section 6 concludes.

2 CABINTEC Database

2.1 Data Acquisition

CABINTEC (âĂIJIntelligent cabin truck for road transportâĂİ) is an ongoing project focused on risk reduction for traffic safety [11]. The CABINTEC project is being developed in a highly realistic truck simulator (shown in Fig. 1(a)). The simulator was made using a real truck cockpit mounted over a stewart-platform to provide a natural driving sensation. Further, the driver's visual field is covered by a detailed simulated 3D scene. The data acquisition process consisted on the detailed monitoring of a simulated driving session of ten minutes in an interurban scenario that simulates a light traffic highway near San Sebastian (Spain). The driving exercise was carried out by professional drivers with more than 20 years of experience. The data acquired at the acquisition process consist of data registers of the vehicle dynamics and road characteristics, and visual information from two video sources: image sequences of driver's top view (Fig. 1(b)) and image sequences of the main view of the road (Fig. 1(c)). The data acquired in the truck simulator was used to make a detailed reproduction of the driving session to a

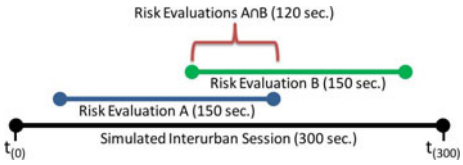


Fig. 2. Time line of the risk evaluations made by each traffic safety expert

group of traffic safety experts in a knowledge acquisition process. The knowledge acquisition process consisted on the risk evaluation, by a group of 38 traffic safety experts, of two partially overlapped sections of the simulated driving session (see Fig. 2). Each traffic safety expert was asked to evaluate the driving risk of the simulated session in the two different time periods in a randomly selected order. For that purpose, the simulation reproduction and knowledge acquisition tool called Virtual Co driver was used [12]. The Virtual Co driver system allows the evaluation of the driving risk through a Visual Analog Scale (VAS) in a range from 0 to 100, where 100 refers to the highest driving risk level. This method has been considered the best for subjective measurements (see, for instance, [13]). The main screen of the Virtual Co driver tool is shown in Fig. 3. The data considered for our CABINTEC database (shown in Fig. 4) consist on 76 risk evaluations (two for each traffic safety expert) obtained from the intersected time lapse between the two risk evaluations. Similar evaluations are expected for each expert. Hence, the capability of our acquisition methodology and the robustness of the subjective risk evaluations will be analyzed. In addition, wrong evaluations could be detected by comparing the two evaluations of the same expert. At first sight, given the high heterogeneity of the experts' VAS evaluations (see Fig. 4), it is hard to identify similar behavior between the curves. Further, given the subjectivity implied on the driving risk evaluation, small oscillations out of the main trend appear. These oscillations make it difficult to analyze subjective phenomena where a linear behavior along a temporary period of time is expected. To get a proper representation of sequential data, a piecewise representation is proposed in the next section.

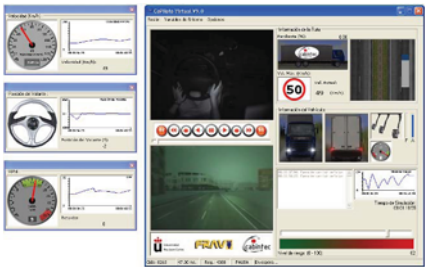


Fig. 3. Simulation reproduction and knowledge acquisition tool (Virtual Co driver)

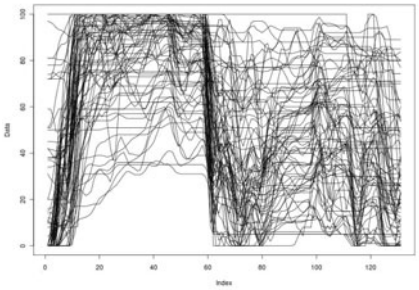


Fig. 4. Subjective sequential data acquired from the traffic safety experts

3 Trend Segmentation Algorithm

One of the main tasks of the present work is to define a proper similarity measure between subjective sequential data. Given the characteristics of sequential data (where sudden changes occur and where the key information is given by its trend), a piecewise representation of the data is appropriate. A variety of algorithms to obtain a proper linear representation of sequential data have been presented (see, for instance, [14], [15] and [16]). However, when working with subjective data, special considerations must be taken into account when selecting the cut points where a linear model will be fitted over the data. In this case, we propose a linear segmentation algorithm based on the time-honored idea of looking for feature points where extreme changes on the data trend occurs. We call this method Trend Segmentation Algorithm (TSA). On the first stage of the algorithm, the feature points of the VAS evaluation where the trend of the data presents a deviation from a straight course must be located. For that purpose the curvature of the data at each point needs to be calculated. Let $f(t)$ be a VAS evaluation at time $t = \{1, \ldots, T\}$. Following [17], the n-order tangent at time t is calculated as:

$$f^n(t) = wf(t) - \sum_{i=-n,\ i\neq 0}^{n} w_i f(t+i)\,, \tag{1}$$

where $w_i = 1/(2|i|)$, and w is the sum of all the weights w_i.

That is, we compute the tangent at t as a weighted average of the VAS evaluation in the n consecutive points surrounding t. The weight w_i is inversely proportional to the distance from the closest point to the point t. The curvature at each point t is computed as the absolute value of the difference between the tangent at that point and the tangent at point $t-1$:

$$C(t) = |f^n(t) - f^n(t-1)|\,. \tag{2}$$

A point t is a feature point if it satisfies one of the following conditions:

1. $t = 1$ or $t = T$ (initial and final point)
2. $C(t) > max(C(t+1)\,, C(t-1))$ (local maximums)

Algorithm 1. Trend Segmentation Algorithm (TSA)

Input: VAS evaluation f, R^2_{min}
Output: $\{CP\}$ **(set of Cut Points)**
1. Obtain the feature points of curve $f \rightarrow \{FP\} = \{fp_1, fp_2, \ldots, fp_N\}$
2. Repeat for each pair of consecutive feature points fp_i**, and** fp_{i+1}
Fit a regression line $(\hat{Y})$ **in the current segment** $[fp_i\,, fp_{i+1}]$
if $(R^2(\hat{Y}) \geq R^2_{min})$ **then**
Store the initial and final points of the current segment as cut points $(fp_i \in CP\,, fp_{i+1} \in CP)$
else
Subdivide the segment to reduce the error and go to 2
end if
3. Joint identical regression lines between the selected cut points.

That is, the feature points are points with relevant changes in the curvature of the original VAS evaluation f. Notice that, since we work with a discretization of the curvature, there will be a smoothing effect depending of the n value. An example of the selection of feature points in a VAS evaluation using the 5-order curvature is shown in Fig. 5. In this example, a total of 38 points where the trend of the data suffered a relevant change were selected as feature points. Given a set of N feature points $\{FP\}$, the second stage of TSA consists on the selection of points where a piecewise linear model can be properly fitted ($Cut\ Points$=$\{CP\}$). As other segmentation algorithms, TSA needs some method to evaluate the quality of fit for a proposed segment. A measure commonly used in conjunction with linear regression is the coefficient of determination. Further, in order to ensure a linear fitting in each section, an Anderson-Darling Normality test is applied to the residuals of each fitted line (see, for instance, [18]). The pseudo code of the TSA is presented in (Algorithm 1). The input of the algorithm is the VAS evaluation f. The output of the algorithm will be a set of $Cut\ Points$ among which a linear model can be fitted with an error lower or equal than the allowed

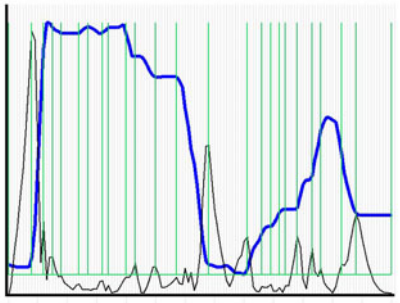


Fig. 5. Example of the selection of feature points (green marks) based on the curvature (black line) of a subjective sequential serie (blue line)

by the parameter R^2_{min}. Given a VAS evaluation f, all its feature points are obtained and stored in the set $\{FP\}$. Next, a linear model is tested in each segment defined by each pair of consecutive points in $\{FP\}$. If the regression error is low (R^2 is higher than R^2_{min}) then the feature points that define the segment are stored as *Cut Points*. Otherwise, the linear model is not proper for the observations in the segment and, as a consequence, the segment is divided. In this work we propose to store all the points in the segment as *Cut Points*. Finally, in order to reduce the number of generated sections, the consecutive segments with identical regression lines are joined together. Following the example shown in Fig. 5, the linear representation of the VAS evaluation after the application of TSA is shown in Fig. 6. In this example, 22 feature points were selected as *Cut Points*. That is, 21 linear sections were enough to represent the VAS evaluation with an R^2_{min} of 0.75. Notice that, the main advantage of the TSA algorithm is its special care when selecting *Cut Points* to fit a linear model. This righteousness becomes very important when working with subjective sequential data because the trend of the data is kept. In order to choose the optimal linear representation of a specific dataset, a trade-off between the global error and the complexity of the representation (number of generated segments) is considered by the minimization of:

$$C = \alpha(1 - R^2) + (1 - \alpha)\frac{\text{number of segments}}{T - 1}, \tag{3}$$

where the parameter α is set to 0.5 to grant similar relevance for both terms. In this case, R^2 is a global average over the R^2 of the linear models fitted in each of the segments obtained from TSA. In order to make comparable two linearized curves it is necessary to align the linear segments of each curve. To achieve this aim, an OR operation between the *Cut Points* selected from each curve is done. Figure 7 shows an example of the OR operation between the set of *Cut Points* selected from a two VAS evaluations. The outcome aligned segmentation for both VAS evaluations is shown in Fig. 7(c).

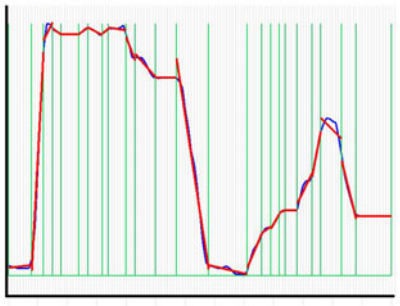
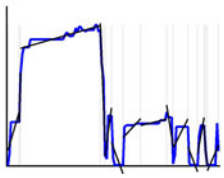
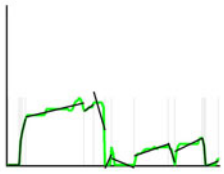


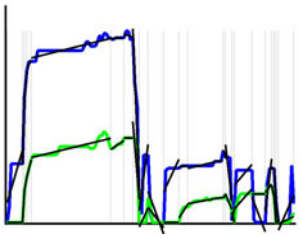
Fig. 6. Example of linear sections generated between the selected feature points



(a) First evaluation



(b) Second evaluation



(c) Combination

Fig. 7. Example of the TSA applied to two series acquired from the same traffic safety expert in different knowledge acquisition experiments

4 Similarity Definitions

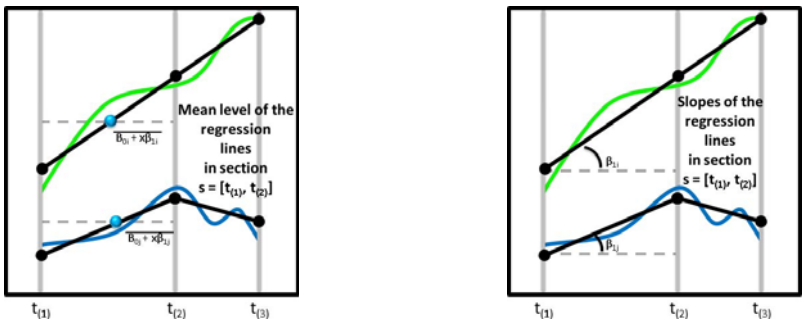
One of the main tasks of the present work is to define a proper similarity measure between subjective sequential data. The similarity between two VAS evaluations can be measured in many ways. Given a pair of aligned linearized curves, it is possible to define a set of similarity measures taking advantage of the characteristics of the linear representation proposed in Section 3. In this work, we propose two similarity measures based on the difference of levels and the difference of angles between the linear regression lines obtained from the TSA representation of the curves.

4.1 Mean Level Based Similarity

Let $k = [t_{(1)}, t_{(2)}]$ be a common section defined for the curves f_i and f_j. Let $\hat{Y}_i = \beta_{0i} + x\beta_{1i}$ and $\hat{Y}_j = \beta_{0j} + x\beta_{1j}$ be the regression lines fitted in the section k of the curves f_i and f_j, respectively. The mean level similarity is based on the mean levels of the regression lines $\hat{Y}_i$ and $\hat{Y}_j$ over the section k (see Fig. 8(a)). The mean level similarity calculated in the section k, denoted by $s_0(k)$, is obtained as one minus the ratio between the Euclidean distance (d) of the mean levels of the regression lines $\hat{Y}_i$ and $\hat{Y}_j$ and the worst possible distances between them:

$$s_0(k) = 1 - \frac{d}{\breve{d}}, \tag{4}$$

where $s_0(k)$ is in $[0, 1]$. The worst distance $\breve{d}$ is calculated from the maximum possible distance that the mean level of the curves could have in all the dataset. In this case, for a set of VAS evaluations ranging in $[0, 100]$, the maximum possible



(a) Mean Level Similarity (b) Angle Level Similarity

Fig. 8. Similarities between two sections of two segmented sequential series

distance $\breve{d}$ is 100. Finally, the overall mean level section-wise similarity for the curves f_i and f_j is calculated as the weighted sum of all the sectional similarities as follows:

$$S_0(f_i, f_j) = \frac{\sum_{k=1}^{K} w(k)\ s_0(k)}{\sum_{k=1}^{K} w(k)}\,, \tag{5}$$

where $w(k)$ is the width of the section $k = 1, \ldots, K$.

4.2 Angle Based Similarity

The angle based section-wise similarity considers the angle formed by the regression lines defined in the sections $k = 1, \ldots, K$. Let β_{1i} and β_{1j} be the slopes of the regression lines $\hat{Y}_i$ and $\hat{Y}_j$, respectively (see Figure 8(b)). The angle between the regression lines is calculated as:

$$\theta = atan(|\beta_{1i} - \beta_{1j}|). \tag{6}$$

The angle based similarity calculated in the section k, denoted by $s_1(k)$, is obtained as the relation between the angle θ and the worst possible angle $\breve{\theta}$ of the section k as follows:

$$s_1(k) = 1 - \frac{\theta}{\breve{\theta}_k}\,, \tag{7}$$

where $s_1(k)$ is in $[0, 1]$. The worst angle $\breve{\theta}$ is established as the maximum possible change in an analyzed section. The maximum possible angle between two regression lines at the section k can be calculated as:

$$\breve{\theta}_k = atan(|\frac{2\breve{d}}{w(k)}|)\,, \tag{8}$$

where $w(k)$ is the width of the section $k = 1, \ldots, K$.

Finally, the overall angle based section-wise similarity for the curves f_i and f_j is calculated as the weighted sum of all the sectional similarities as follows:

$$S_1(f_i, f_j) = \frac{\sum_{k=1}^{K} w(k)\, s_1(k)}{\sum_{k=1}^{K} w(k)}\,, \tag{9}$$

where $w(k)$ is the width of the section $k = 1, \ldots, K$.

5 Experiments

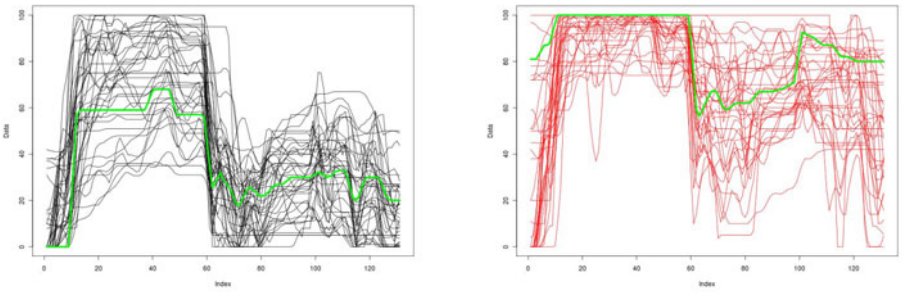
Given the set of curves from the CABINTEC dataset, it is possible to generate similarity matrices with the definitions presented in the Section 4. For our purposes, an unique similarity matrix mas built:

$$S_{0,1} = \frac{S_0 + S_1}{2}\,. \tag{10}$$

In this case, as the proposed similarities matrices (S_0, and S_1) are obtained as the weighted mean of a set of similarities obtained from individual sections, a deviation from the Euclidianess may occur. Following [19], the deviation from Euclidianess of each similarity matrix of the CABINTEC dataset was calculated as the ratio of the smallest negative eigenvalue to the largest positive eigenvalue of the similarity matrices (r_{mm}). When the negative eigenvalues are relatively small in magnitude, those negative eigenvalues can be interpreted as a noise contribution. However, if the negative eigenvalues are relatively large, possibly important information could be rejected by neglecting them (see [20] for a complete description). The deviation of each matrix is presented in Table 1. Several techniques have been proposed to solve this problem ([20]). In this work, Multidimensional Scaling was applied to represent the data set in a Euclidean space. As mentioned before, the CABINTEC database consist of 76 VAS evaluations of a group of 38 traffic safety experts. That is, the same simulated driving session was evaluated twice for each expert. We will illustrate the performance of the similarity measures defined in Section 4 based on two kind of experiments on the CABINTEC dataset. The first one is a cluster experiment, whose main objective is to know if there are meaningful classes of experts that can be grouped together. In addition, it is possible to detect wrong evaluations when the two evaluations of the same experts are grouped in different clusters. The second experiment is based on a new measure for outlier detection. If there is a high difference between the two evaluations of the same expert, then the expert is considered an outlier and should be studied carefully.

Table 1. Euclidianess deviation of the similarity matrices of the dataset CABINTEC

Similarity	**Lowest Eigenvalue**	**Highest Eigenvalue**	**Deviation from Euclidianess (r_{mm})**
S_0	-0.1602	58.6125	0.0027
S_1	-0.0165	50.3231	0.0003



(a) Cluster 1 (b) Cluster 2

Fig. 9. Clusters of the CABINTEC database with the mean level and angle similarities

5.1 Clustering

Clustering is an initial and fundamental step in data analysis. It is an unsupervised technique whose goal is to reveal a natural partition of data into a number of meaningful subclasses or clusters. Accurate clustering requires a precise definition of the nearness between a pair of objects, in terms of either the pairwised similarity or distance. Clustering of sequential data differs from clustering of static feature data mainly in how to compute the similarity between two data objects. In general, depending upon the specific characteristics of the sequential data, different clustering studies provide different methods to compute this similarity. Once the similarity of sequential data is defined, many general-purpose clustering algorithms can be used to partition the data. In this work, we focus on clustering sequential data in which each sequential object is represented as a set of regression lines defined from a linearization algorithm. In our work, we test the capability of the similarity measure presented in (10) in order to achieve accurate clustering of the CABINTEC experts' evaluations. We will use this similarity to perform a partitioning clustering of the experts' evaluations into clusters around k representative objects or medoids among the sequential experts' evaluations of the dataset (see [21] for a complete description of the PAM algorithm). The clusters generated for the CABINTEC database are shown in Fig. 9. To apply PAM method, we will work with the dissimilarity defined as $1 - S_{0,1}$. For each cluster, the medoid (a representative VAS evaluation of the cluster) is remarked with a green line. In this case, two clearly identifiable patterns were found. In the simulated driving exercise, the driver received a phone call from second 5 to second 60. In the second cluster (Fig. 9(b)), the traffic safety experts considered the phone call as the maximum fault giving a 100 in their VAS evaluations. However, in the first cluster (Fig. 9(a)), the traffic safety experts did not consider to answer a phone call as the maximum risk in which a driver could fall. In this way, as the experts' evaluations bunched in cluster 2 were giving the maximum possible risk level during all the phone call, they were unable to penalize the action of driving with no hands on the steering wheel given from second 38 to second 50. On the other side, this action (no hands

Table 2. Clustering error of the CABINTEC database with several similarity measures

Method		Bad clustered experts	Error (%)
TSA	S_0	5	13.2
	S_1	12	31.6
	$\frac{S_0+S_1}{2}$	**1**	**2.6**
DTW		3	7.9
Euclidean distance		6	15.8
Hausdorff distance		12	31.6
Kendall correlation		16	42.1
Pearson correlation		17	44.7

on wheel) was detected and penalized by most of the experts bunched on the cluster 1. At the second half of the risk evaluation (from second 60) the experts of the first cluster punish in a moderated way a group of risky situations of the driver leaving a margin on their VAS range to punish riskier situations that could come. In the same way, the experts of the second cluster, detected most of the risky situations given at the second half of the evaluation. However, these experts continued with high VAS values until the end of the risk evaluation. The clusters generated with the similarities proposed in this work are very helpful to select the kind of data that will be considered in the future stages of the research. On the one hand, we have identified experts (cluster 2) whose major concern is the distraction of the driver while doing a secondary task (like a phone call). On the other hand, we have identified experts (cluster 1) that are more concerned about the driving efficiency regardless of the number of tasks of the driver. It is well known that is very difficult to conduct a systematic study comparing the impact of similarity metrics or distances on cluster quality, because objectively evaluating cluster quality is difficult in itself. In practice, manually assigned category labels are usually used as a baseline criteria for evaluating clusters. Nevertheless, in this case we know a relevant information: the expert that generated each evaluation. We estimate our clustering error as the number of experts such that the two evaluations of the same expert are grouped in different clusters. This error measure was used to compare the methodology proposed in this work with the clustering based on other well-known distance measures (DTW: Dynamic time warping, Euclidean and Hausdorff distance, and correlation of Kendall and Pearson). The results are shown in Table 2. The best result is achieved by the combination of the mean level and angle similarities. These results show the complementarity of both similarity measures achieving an error reduction from 13.2% and 31.6% to 2.6%.

Other Data. Additionally to the CABINTEC dataset, several well-known databases, out of the driving risk problem, were analyzed: the ECG200, Gun Point, Coffee, and Growth databases. A summary of these databases is shown in Table 3 (see [22] and [23] for a complete description). The Trend Segmentation Algorithm and the Similarity measure proposed in Section 4 were applied to all the series. Results are presented in Figure 10. For the ECG200 [22], two patterns

were found among the 200 curves of the dataset. On the first one (see Fig. 10(b)), the main valley is generated faster with a steeper slope. After that, a decreasing behavior is observed until the end. On the second one (see Fig. 10(c)), the main valley is reached later with a moderated slope and an increasing behavior is observed until the end. In this case, it is clear that the angle and mean level similarities are useful to separate these patterns. In the same way, for the Gun Point database (see Fig. 10(e) and 10(f)) [22], the patterns found among the 200 curves are evidently dependent of the width of the main peak. For this clusters, the similarities presented in this paper shows relevance when separating the patterns. For the Coffee database (see Fig. 10(h) and 10(i)) [22], the two patterns found among the curves show a major relevance on the information given by the mean level similarity. In this case, the curves are mainly identified by its level since they have a similar angle behavior. Finally, for the Growth database (see Fig. 10(k) and 10(l)) [22], the patterns identified among the curves are clearly discovered by the behavior of their slope. In this case, each cluster is characterized by the tilt of each curves while they increase along their 31 registers.

5.2 Outlier Detection

One of the first tasks in any outlier detection method is to determine what an outlier is. This labor strongly depends on the problem under consideration. In this case, we are interested in the detection of experts that generated heterogeneous evaluations (or even random evaluations) during the acquisition process. An expert should be considered an outlier if a very high distance between the two risk evaluations of the experts is observed. Let f_i^1, and f_i^2 be the two evaluations obtained from expert i. Given the similarity measure presented in Section 4 two sets of evaluations are defined:

$$F(i)_{1,2} = \{f_j : S_{0,1}(f_i^1, f_j) > S_{0,1}(f_i^1, f_i^2)\}\,, \tag{11}$$

$$F(i)_{2,1} = \{f_j : S_{0,1}(f_i^2, f_j) > S_{0,1}(f_i^1, f_i^2)\}\,. \tag{12}$$

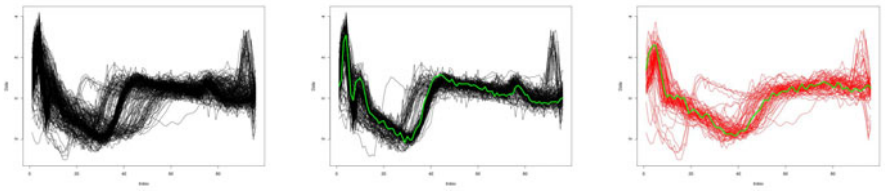
That is, $F(i)_{1,2}$ is the set of the experts' evaluations such that the similarities between them and the first evaluation of expert i are higher than the similarity between the two evaluations of expert i. Hence, the evaluations between evaluations 1 and 2 of expert i are considered.

$$\delta(i) = \#\{F(i)_{1,2} \cap F(i)_{2,1}\}\,. \tag{13}$$

Table 3. Summary of the databases considered in the clustering experiments

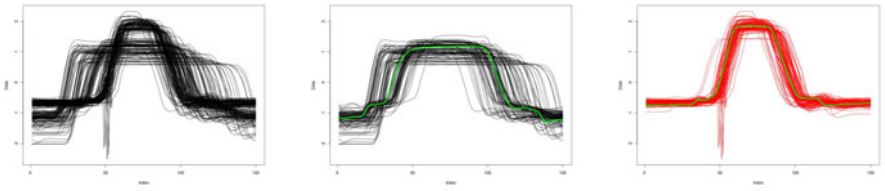
Database Name	Number of series	Time Series Length	Figure
ECG200	200	96	10(a)
Gun Point	200	150	10(d)
Coffee	56	286	10(g)
Growth	93	31	10(j)

ECG200 database



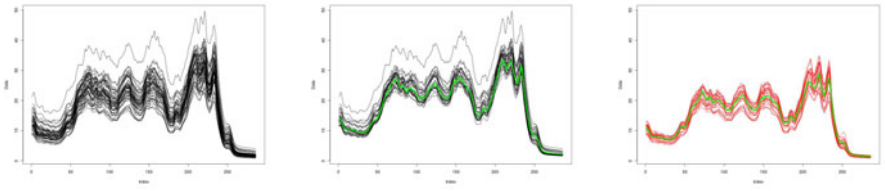
(a) Original data (b) Cluster 1 (c) Cluster 2

Gun Point database



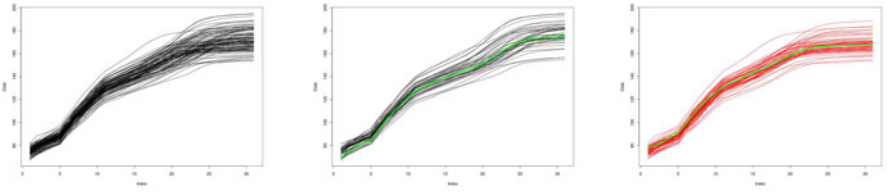
(d) Original data (e) Cluster 1 (f) Cluster 2

Coffee database



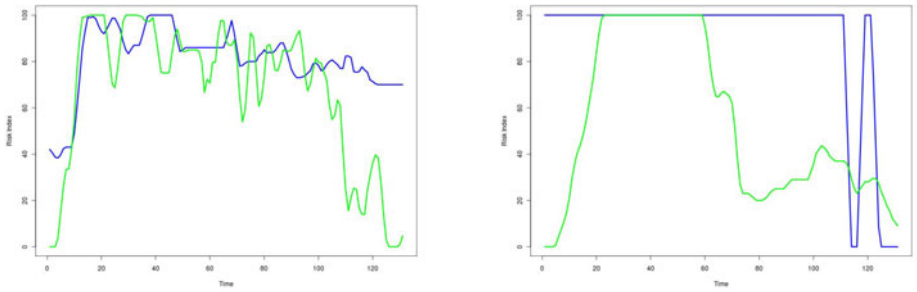
(g) Original data (h) Cluster 1 (i) Cluster 2

Growth database



(j) Original data (k) Cluster 1 (l) Cluster 2

Fig. 10. Clustering of several databases with the mean level and angle similarities



(a) Expert 22 (b) Expert 23

Fig. 11. Outliers of the CABINTEC database with the mean level and angle similarities

In the same way, $F(i)_{2,1}$ is the set of the experts' evaluations such that the similarities between them and the second evaluation of expert i are higher than the similarity between the two evaluations of expert i. Hence, the evaluations between evaluations 2 and 1 of expert i are considered. Given that, we deal with sequential data, in general $F(i)_{1,2} \neq F(i)_{2,1}$. To obtain the outliers evaluations in our experiments, we define the following dissimilarity measure: That is, given the similarity measure $S_{0,1}$, the dissimilarity measure evaluated on expert i equals the number of evaluations between his two evaluations. On the one hand, if the two evaluations of expert i are very similar, there will be very few elements in sets $F(i)_{1,2}$ and $F(i)_{2,1}$, and as a consequence, $\delta(i)$ will be very low. On the other hand, if the two evaluations of expert i are not similar, there will be very a lot of elements in sets $F(i)_{1,2}$ and $F(i)_{2,1}$, and as a consequence, $\delta(i)$ will be very high. Next, we calculate this dissimilarity measure in the CABINTEC database. Table 4 presents the values of the dissimilarity function (13) in the experts' VAS evaluations of the CABINTEC database.

Table 4. Dissimilarity measure for outliers detection in the CABINTEC database

δ **value**	**0**	**1**	**2**	**4**	**5**	**6**	**10**	**12**	**20**	**24**
Number of experts	24	5	2	1	1	1	1	1	1	1

Notice that in 24 out of 38 experts (63.2%) no other evaluations were found between the two expert's evaluations. That is, there are no neighbors in common between the first and the second evaluations of these experts. On the other hand, there were two experts with 20 and 24 neighbors in common between their two evaluations. That is, the two evaluations of the same expert are strongly different. Figure 11 shows the evaluations of these two experts that are considered as outliers. For the expert 22 (Fig. 11(a)), a group of contradictions could be observed between his two evaluations. In addition, from the second 100 , the first evaluation of the expert (green line), shows a decreasing risk value until the end of the evaluation and, on the other side, the second evaluation (blue line) shows

a high level VAS evaluation the whole time. For the expert 23 (Fig. 11(b)), his second evaluation (blue line) shows a total disinterest on the experiment.

6 Conclusions

The main contribution of this paper, is a novelty methodology for the analysis of subjective sequential data. First, a linear segmentation algorithm for the proper representation of subjective data, based on the location of feature points, has been developed. This algorithm is useful to represent sequential data in a piecewise model emphasizing the trend of the data. Next, two similarity measures have been defined from the differences between the level and the angle of the lines of the piecewise representation. This similarities were defined in order to cover the two more relevant characteristics of the trend: behavior (angle) and scale (level). The methodology proposed in this work, focused on the representation and similarity measurement of subjective data, have been used for clustering several experts' risk evaluations of a simulated driving exercises. Further, a novel dissimilarity measure for outlier detection of paired sequential data have been proposed. The results of the cluster and outlier detection experiments show that both level and angle based similarities contain complementary and relevant information about the data trend. In the future, clustering of the individual segments of a linear segmentation representation of sequential data will be performed. In this way, potential high driving risk areas will be detected and studied for its prediction.

Acknowledgments. Supported by the Minister for Science and Innovation of Spain: CABINTEC (PSE-37010-2007-2) and VULCANO (TEC2009-10639-C04-04). Thanks to CONACYT and CONCYTEY from México for supporting the project through their scholarship programs.

References

1. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: Lomet, D.B. (ed.) FODO 1993. LNCS, vol. 730. Springer, Heidelberg (1993)
2. Chan, K., Fu, W.: Efficient time series matching by wavelets. In: Proceedings of the 15th IEEE International Conference on Data Engineering (1999)
3. Perng, C., Wang, H., Zhang, S., Parker, S.: Landmarks: a new model for similarity-based pattern querying in time series databases. In: Proceedings of the 15th IEEE International Conference on Data Engineering (2000)
4. Keogh, E., Pazzani, M.: An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In: KDD, pp. 239–243 (1998)
5. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J.: Mining of concurrent text and time series. In: Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, pp. 37–44 (2000)

6. Park, S., Kim, S.W., Chu, W.W.: Segment-based approach for subsequence searches in sequence databases. In: Proceedings of the 16th ACM Symposium on Applied Computing (2001)
7. Wang, C., Wang, S.: Supporting content-based searches on time series via approximation. In: Proceedings of the 12th International Conference on Scientific and Statistical Database Management (2000)
8. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE, 257–286 (1989)
9. García-García, D., Parrado-Hernandez, E., Díaz-de-Maria, F.: Anderson-darling: A goodness of fit test for small samples assumptions. P. Recognition 44, 1014–1022
10. Panuccio, A., Bicego, M., Murino, V.: A hidden markov model-based approach to sequential data clustering. In: Caelli, T.M., Amin, A., Duin, R.P.W., Kamel, M.S., de Ridder, D. (eds.) SPR 2002 and SSPR 2002. LNCS, vol. 2396, pp. 734–742. Springer, Heidelberg (2002)
11. Brazalez, A., et al.: CABINTEC: Cabina inteligente para el transporte por carretera. In: Proc. of the Congreso Español de Sistemas Inteligentes de Transporte (2008)
12. Siordia, O.S., Martín, I., Conde, C., Reyes, G., Cabello, E.: Driving risk classification based on experts evaluation. In: Proceedings of the 2010 IEEE Intelligent Vehicles Symposium (IV 2010), San Diego, CA, pp. 1098–1103 (2010)
13. Cork, R.C., Isaac, I., Elsharydah, A., Saleemi, S., Zavisca, F., Alexander, L.: A comparison of the verbal rating scale and the visual analog scale for pain assessment. Technical Report 1, Int. Journal of Anesthesiology (2004)
14. Keogh, E., Chu, S., Hart, D., Pazzani M.: Segmenting time series: A survey and novel approach. In: Data Mining in Time Series Databases, pp. 1–22 (1993)
15. Lachaud, J., Vialard, A., de Vieilleville, F.: Analysis and comparative evaluation of discrete tangent estimators. In: Andrès, É., Damiand, G., Lienhardt, P. (eds.) DGCI 2005. LNCS, vol. 3429, pp. 240–251. Springer, Heidelberg (2005)
16. Zhu, Y., Wu, D., Li, S.: A piecewise linear representation method of time series based on feature points. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 1066–1072. Springer, Heidelberg (2007)
17. Basri, R., Costa, L., Geiger, D., Jacobs, D.: Determining the similarity of deformable shapes. Vision Research 38, 135–143 (1995)
18. Romeu, J.L.: Anderson-darling: A goodness of fit test for small samples assumptions. Selected Topics in Assurance Related Technologies 10(5), 1–6 (2003)
19. Pękalska, E., Duin, R.P.W., Günter, S., Bunke, H.: On not making dissimilarities euclidean. In: Fred, A., Caelli, T.M., Duin, R.P.W., Campilho, A.C., de Ridder, D. (eds.) SSPR&SPR 2004. LNCS, vol. 3138, pp. 1145–1154. Springer, Heidelberg (2004)
20. Pekalska, E., Paclík, P., Duin, R.P.W.: A generalized kernel approach to dissimilarity-based classification. Journal of Machine Learning Research, Special Issue on Kernel Methods 2(12), 175–211 (2001)
21. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
22. Keogh, E., Xi, X., Wei, L., Ratanamahatana, A.: The ucr time series classification/clustering (2006), http://www.cs.ucr.edu/~eamonn/time_series_data/
23. Ramsay, J., Silverman, B.: Functional Data Analysis, Secaucus, NJ, USA. Springer Series in Statistics (2005)