Pattern Recognition Letters 000 (2017) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Using data dimensionality reduction for recognition of incomplete dynamic gestures

Miguel Simão[a], Pedro Neto[a,*], Olivier Gibaru[b]

[a] University of Coimbra, Department of Mechanical Engineering - PÓLO II, Coimbra, 3030-788, Portugal
[b] École National Superieure d'Arts et Métiers - ParisTech, 8, Bd Louis XIV, 59046 Lille Cedex, France

ARTICLE INFO

Article history:
Available online xxx

MSC:
41A05
41A10
65D05
65D17

Keywords:
Gesture recognition
Data dimensionality reduction
Human-robot interaction

ABSTRACT

Continuous gesture spotting is a major topic in human-robot interaction (HRI) research. Human gestures are captured by sensors that provide large amounts of data that can be redundant or incomplete, correlated or uncorrelated. Data dimensionality reduction (DDR) techniques allow to represent such data in a low-dimensional space, making the classification process more efficient. This study demonstrates that DDR can improve the classification accuracy and allows the classification of gesture patterns with incomplete data, i.e., with the initial 25%, 50% or 75% of data representing a given dynamic gesture (DG) - time series of positional and hand shape data. Re-sampling raw data with bicubic interpolation and principal component analysis (PCA) were used as DDR methods. The performance of different classifiers is compared in the classification 95 different signs of the UCI Australian Sign Language (High Quality) Dataset. Experimental tests indicate that the use of PCA-based features result in a classification accuracy that is higher with 25% of gesture data (93% accuracy) than with 100% of gesture data (82% accuracy). These results were obtained from a non-trained data set and the recognized gestures are used to control a robot in an collaborative process.

1. Introduction

The ability to interact with a robot in a natural and intuitive way, for example using speech and gestures, has brought important advances to the way our societies look to robots. The paradigm for robot usage has changed in the last few years, from a concept in which robots work with complete autonomy to a scenario in which robots cognitively collaborate with human beings. This brings together the best of each partner, robot and human, by combining the coordination and cognitive capabilities of humans with the robots' accuracy and ability to perform monotonous tasks. For this end, robots and humans have to understand each other and interact in a natural way, creating a co-working partnership. This will allow a greater presence of robots in our companies, schools, hospitals, etc., with consequent positive impact on society' s life standards. The current problem is that the existing interaction modalities are neither intuitive nor reliable. Instructing and programming an industrial robot by the traditional teaching method is a tedious and time-consuming task that requires technical expertise.

The robot market is growing and the human-robot interaction (HRI) interfaces will have a main role in the acceptance of robots as co-workers. Gestures and other natural interaction modalities may decrease the need for technical expertise in robot programming, therefore decreasing the cost of owning a robot.

Multimodal HRI interfaces combining gestures, speech and tactile based-actions are expected to be in a near future the standard for a reliable and intuitive interaction process. Nonverbal communication cues in the form of gestures are considered to be an effective way to approach natural HRI. For instance, a person can point to indicate a position to a robot, use a dynamic gesture to instruct a robot to move and a static gesture to stop the robot [17,24]. In this scenario, the user has little or nothing to learn about the interface, focusing on the task and not on the interaction [23]. For all the reasons mentioned above, continuous and real-time gesture spotting (segmentation and recognition) are key factors to bridge the gap between laboratory research and real world application of novel HRI modalities [22]. A major challenge in continuous gesture recognition has to do with the fact that there are movement segments between gestures that have no meaning. Such inter-gesture transition periods are transition frames and are known as Movement Epenthesis (ME). Most studies treat ME as a classification problem [8].

Some gestures, although not all, can be defined by their spatial trajectory. This is particularly true for pantomimic gestures, which are often used to demonstrate a certain motion to be done, e.g., a

* Corresponding author.
E-mail address: pedro.neto@dem.uc.pt (P. Neto).

http://dx.doi.org/10.1016/j.patrec.2017.01.003
0167-8655/

circle. Burke and Lasenby focused with success on using PCA and Bayesian filtering to classify these time series [2]. The importance of DDR PCA to reduce the dimensionality of a dataset representing human gestures for HRI has been studied in [3,11]. In a different study, PCA is applied to a continuous stream of time series data capturing body motions with an accuracy of 91% [14]. A novel approach called joint sparse principal component analysis (JSPCA) to jointly select useful features and enhance robustness to outliers is proposed in [25]. In [12] it is proposed a unified sparse learning framework by introducing the sparsity or L1-norm learning, which further extends the locally linear embedding (LLE)-based methods to sparse cases. Four methods based on L2,1-norm for linear dimensionality reduction are proposed in [13]. These methods are robust to outliers and have more freedom to jointly select the useful features for a low-dimensional representation. Experimental results on image datasets show that such algorithms obtain competitive performance compared with other DDR methods. A DDR approach using sparsified singular value decomposition (SSVD) technique to identify and remove trivial features before applying feature selection is proposed in [15]. A reference study presents a sequence kernel dimension reduction approach (S-KDR) in which spatial, temporal and periodic information is combined in a principled manner and an optimal manifold is learned [20]. A novel support vector number reduction method is in [6]. The support vector number is reduced by more than 99.5% without accuracy degradation.

A recent study indicates that after DDR PCA the classification accuracy is higher with incomplete gesture data than with complete gesture data [21]. These results were obtained in a relatively small library of 10 hand/arm dynamic gestures. In [1] it is presented a gesture recognition solution from scale independent and partial input data (25%, 50% and 75% of the total gesture length) with an error rate of 3%. For complete data the accuracy is 100%, however, these results were obtained from a relatively small library of 16 classes of simple gestures in an x-y plane. Reasoning with incomplete data can be associated in some way to the concept of anticipation. For example we may identify a gesture before it is finished. An important study in the field represents each possible future using an anticipatory temporal conditional random field (ATCRF) that models the rich spatial-temporal relations through object affordances [10]. Human activity prediction has been studied for task recognition. Tasks can be modelled with a Dynamic Bayesian Network (DBN) in order to estimate the current task, predicting the most probable future pairs of action-object and correcting possible misclassification [16]. The combination of data from different sources greatly influences the ability to predict gesture and voice patterns [19]. When successful, such prediction ability allows to increase safety and reliability in HRI process. In addition, prediction attenuates the negative effect of machine communication delays and algorithm processing time in the HRI loop.

Gesture classification has been studied over the years. However, there remains the problem with reliability and intuitiveness, which are key factors for a system's acceptance by end-users. Other challenges are related with the ability to perform the recognition in real-time and continuously, recognize gesture patterns with incomplete data, and performing DDR keeping or increasing the recognition accuracy. DDR also allows to reduce training data in supervised classification methods and consequently reduce the training time. The number of existing studies approaching pattern classification with incomplete data for prediction/anticipation purposes is still very limited.

1.1. Overview and proposed approach

In a real-world system we usually have a device setup that captures features of what we are analyzing. In the case of dynamic gestures, time dependencies may also be important. As opposed to static gestures, of which we only need snapshots of the data at certain instants, dynamic gestures require that data are recorded over time, generating a multivariate time series. Oftentimes, the data are sampled at high rates, quickly generating high dimension data sets. Gesture patterns contain spatial variation between sequences and also temporal variation, which is not necessarily linear. Time series recognition is a currently active research topic.

In this work we introduce an approach to perform the recognition of time series segments, targeted at natural language processing (NLP) of hand gestures. This approach has the advantage of allowing accurate classification to be performed with partial (incomplete) gesture data Fig. 1.

To overcome the problem of classification of time sequences, i.e., DGs, we present two approaches to their feature extraction and DDR: one based on up- or down-sampling of gesture frames using interpolation, and another based on PCA. The PCA approach has the advantage of being capable of yielding a good classification even before the gesture is finished – with incomplete/partial gesture data. Complex DGs are defined by a large set of features, with a variable number of frames, and with both trajectory and finger movement. The performance of different classifiers is compared in the process to classify the 95 different signs of the UCI Australian Sign Language (High Quality) Dataset. Independently of the type of gesture and feature extraction approach, the predictor for a certain sample is represented by the vector $\mathbf{z} \in \mathbb{R}^d$:

$$\mathbf{z}^{(i)} = [f_1 \ f_2 \ \dots \ f_d] \tag{1}$$

where f_i is the ith element of $\mathbf{z}$.

2. Data dimensionality reduction and classification

2.1. Overview

Lets assume we have a data set $\mathcal{S}$ which contains a number of samples and corresponding labels. A specific sample $\mathcal{S}(i)$ of the data set is represented by the matrix $\mathbf{X}^{(i)}$ and its label is $\mathbf{t}^{(i)}$. A function f is then used to extract the feature-vector $\mathbf{z}$ from each sample:

$$\begin{array}{lll} f: & \mathbb{R}^{d\times n} & \rightarrow \mathbb{R}^b \\ & \mathbf{X} & \rightarrow \mathbf{z} \end{array} \tag{2}$$

Above, b is the dimensionality of the feature vector. This transformation can have several steps and DDR may be one of them. The vector $\mathbf{z}$ is the input for the classifiers and $\mathbf{t} \in \{0,1\}^{n_{classes}}$ is the target value of the classifier for that sample (supervised learning). If the target class has the number o, the target vector $\mathbf{t}^{(o)}$ is defined by:

$$\mathbf{t}_j^{(o)} = \delta_{oj}, \quad j = 1, \dots, n_{classes} \tag{3}$$

where δ is the Kronecker delta and $\mathbf{t}_j$ is the jth element of $\mathbf{t}$.

The transformation f may still yield a vector with very high dimensionality b, which may be detrimental for the classification. Therefore, intermediate processing techniques, such as PCA and interpolation, are proposed for DDR and are defined as in (4). Nevertheless, DDR can be done either before or after feature extraction.

$$\begin{array}{lll} r: & \mathbb{R}^b & \rightarrow \mathbb{R}^{b'} \\ & \mathbf{z} & \rightarrow \mathbf{z}' \end{array}, \qquad b' < b \tag{4}$$

2.2. Resampling with bicubic interpolation

Resampling is done with bicubic interpolation to transform a DG sample $\mathbf{X}^{(i)}, \ i \in i^D, \ \mathbf{X} \in \mathbb{M}^{d\times n}$, which has a variable number of frames n, into a fixed-dimension sample $\mathbf{X}', \ \mathbf{X}' \in \mathbb{M}^{d\times k}$. In this

Please cite this article as: M. Simão et al., Using data dimensionality reduction for recognition of incomplete dynamic gestures, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.01.003

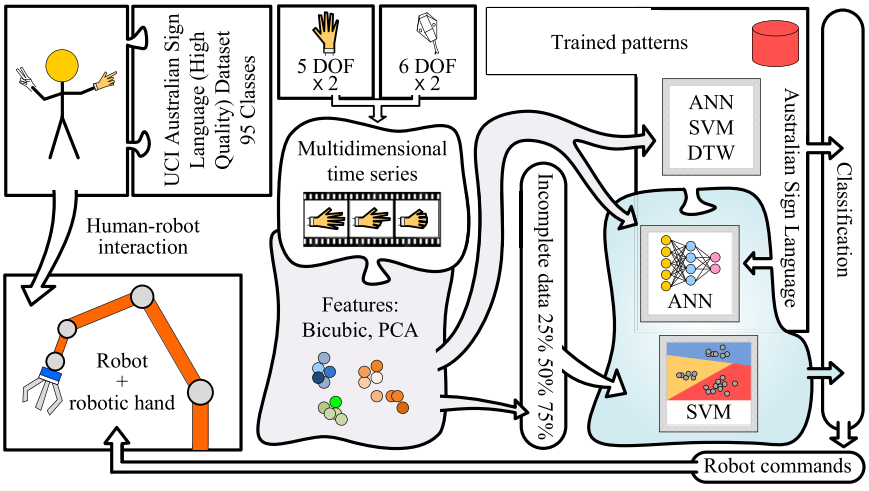

Fig. 1. Overview of the proposed gesture recognition system.

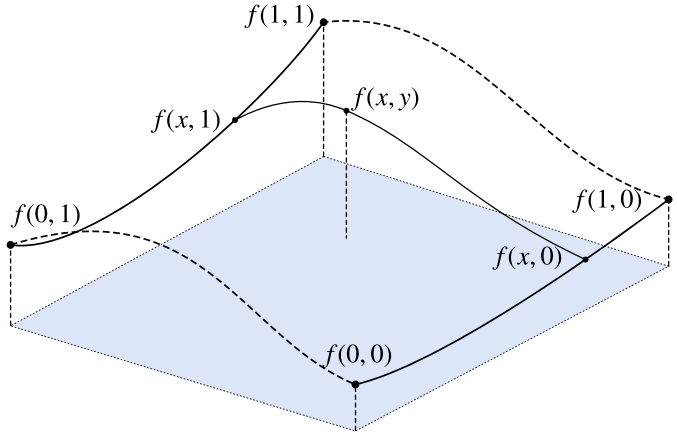

Fig. 2. Representation of the result of bicubic interpolation on a 2 × 2 grid of points $f(0, 0), f(1, 0), f(0, 1), f(1, 1)$.

work, we propose down-sampling the gesture data so that $k \leq n$, being k arbitrarily defined as the minimum n in all of the samples i so that $i \in i^D$. In this way, the proposed transformation is effectively down-sampling the gesture data, thus reducing the dimensionality of the feature vector.

$$\begin{aligned} interp: \quad & \mathbb{R}^{d\times n} \rightarrow \mathbb{R}^{d\times k} \\ & \mathbf{X} \rightarrow \mathbf{X}' \end{aligned} \tag{5}$$

The bicubic interpolation method [9] yields a surface p described by 3rd order polynomials in both dimensions of space. Given a patch of dimension 2 × 2, there are 4 data points in which we know the values f and derivatives f_x, f_y and f_{xy}. The derivatives are not known at the boundaries, but they can be estimated using finite differences. The interpolated values inside the uniformized 2 × 2 sector are given by:

$$p(x, y) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}x^i y^j \tag{6}$$

A representation of the sector is in Fig. 2.

The problem is determining the 16 coefficients a_{ij}. The function values and 3 derivatives at the 4 points provide $4 \times 4 = 16$ linear equations, which can be written as an equation system $\mathbf{A}\alpha = \mathbf{x}$ with:

$$\alpha = [a_{00}\ a_{10}\ a_{20}\ a_{30}\ a_{01}\ \ldots\ a_{33}]^T \tag{7}$$

$$\mathbf{x} = [f(0,0)\ f(1,0)\ \ldots\ f_x(0,0)\ \ldots\ f_y(0,0)\ \ldots\ f_{xy}(0,0)\ \ldots\ f_{xy}(1,1)]^T \tag{8}$$

The matrix $\mathbf{A}$ is nonsingular, so the equation system can be rewritten as $\alpha = \mathbf{A}^{-1}\mathbf{x}$. This process is used for all patches in the bi-dimensional grid. The derivatives at the boundaries of a patch are maintained across neighbouring patches. In order to apply the method to the whole data grid efficiently, techniques such as Lagrange polynomials, cubic splines or cubic convolution algorithms are used. The resulting interpolated data points are smother and have less artifacts than those using other interpolation methods, such as bilinear interpolation.

2.3. Principal component analysis

PCA is a mathematical tool that performs an orthogonal linear transformation of a set of n p-dimensional observations, $\mathbf{X} \in \mathbb{R}^{n\times p}$, into a space defined by the principal components (PC). The PC have necessarily a size less than or equal to the number of original dimensions, p. The first component has the largest possible variance observed in the observations. Each of the following PC is orthogonal to the preceding component and has the highest variance possible under this orthogonality constraint. The PC are the eigenvectors of the covariance matrix and its eigenvalues are a measure of the variance in each of the PC. Therefore, PCA can be used for reducing the dimensionality of data by projecting that data into the PC space and truncating the lowest-ranked dimensions. These dimensions have the lowest eigenvalues, so truncating them retains most of the variance present in the data.

The first step in PCA is centering the data, because PCA is sensitive relative to the scaling of the original dimensional space. This is done by subtracting each of a dimension's values by its overall average.

The PC transformation is very often determined by another matrix factorization method, the SVD of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{9}$$

where $\mathbf{X} \in \mathbb{M}^{n\times p}$ is the original data matrix. $\mathbf{\Sigma} \in \mathbb{M}^{n\times p}$ is a diagonal matrix with the singular values of $\mathbf{X}$, $\mathbf{U}$ is a $n \times n$ matrix whose

columns are orthogonal unit vectors that are the left singular vectors of $\mathbf{X}$, and $\mathbf{V} \in \mathbb{M}^{p\times p}$ is a matrix whose columns are unit vectors, the right singular vectors. Both $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, so that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$. The singular values $\sigma_1, \sigma_2, \ldots$ in the diagonal of $\mathbf{\Sigma}$ are the positive square roots, $\sigma_i = \sqrt{\lambda_i} > 0$, of the nonzero eigenvalues of the Gram matrix $\mathbf{K} = \mathbf{X}^T\mathbf{X}$, thus being always positive.

The implementation used for this purpose was Matlab's *pca* function. Since the input data matrix $\mathbf{X}$ is most often rectangular, the function uses the aforementioned SVD method (9) for the matrix decomposition. The singular values, i.e., the variance in each of the PC, are the eigenvalues of the covariance matrix of $\mathbf{X}$. The covariance matrix of p sets variates $\{\mathbf{x}_1\}, \ldots, \{\mathbf{x}_p\}$, $\mathbf{x}_i = \mathbf{X}_{\bullet i}$ is defined by $\mathbf{W} \in \mathbb{M}^{p\times p}$:

$$\mathbf{W}_{ij} = \mathrm{cov}(\mathbf{x}_i, \mathbf{x}_j) \equiv \left\langle (\mathbf{x}_i - \mu_i)\left(\mathbf{x}_j - \mu_j\right)\right\rangle, i, j = 1, \ldots, p \tag{10}$$

where μ and $\langle\,\rangle$ denote mean value, being $\mu_i = \langle \mathbf{x}_i \rangle$. $\mathbf{W}$ can also be written as $\mathbf{W}_{ij} \equiv 1/n - 1\mathbf{X}\mathbf{X}^T$. The product $\mathbf{X}\mathbf{X}^T$ has as eigenvectors the columns of $\mathbf{U}$.

Although PCA is most often performed to reduce the dimensionality of the observations, in this work we preferred to use the PCs as features. The first PC or singular vector $\mathbf{U}_{\bullet 1}$ determines the direction in the PC-space in which there is the most variance during a DG. The variance is measured by the respective singular value, $\mathbf{\Sigma}_{11}$. Therefore, we expect these values to produce good features for the DG classification, even if the gesture is incomplete. We also used PCA to represent gesture features in lower two- and three-dimensional spaces, for easier visualization.

2.4. Classifiers

In this study, three distinct classification methods where applied: dynamic time warping (DTW), support vector machine (SVM) and artificial neural networks (ANN). These methods are commonly used for the classification of time series data. The ANN will be parameterized in different ways to best adapt to the classification problem in study.

3. Experimental results

3.1. UCI Auslan dataset

The proposed approach was tested with the Australian Sign Language (Auslan) signs (High Quality) dataset, from the UCI machine learning repository [7]. Each sample of this set is a multidimensional time series. The whole data set was used, with 95 distinct classes and 27 examples for each one of the classes. The samples were obtained from one native Auslan signer over a period of 9 weeks.

The data acquisition setup consisted of two Fifth Dimension Technologies (5DT) gloves and two Ascension Flock-of-Birds magnetic position trackers. Each of the subjects two hands was equipped with a glove and a tracker. The data gloves measured finger flexion for each of the 5 fingers and the trackers recorded positional and orientation of the hands – 6 degrees of freedom (DOF). These 22 degrees of freedom were measured over time at a rate of close to 100 frames per second. The average length of each sign is about 60 frames. Each frame is represented as a 15 dimensional feature vector consisting of hand position (X,Y,Z), roll, yaw, pitch, and bend measurements of different fingers.

In this work, a sample in the data set is represented by $\mathcal{S}$:

$$\mathcal{S} = \left\{\mathbf{X}^{(i)},\ t^{(i)}\right\},\ \mathbf{X}^{(i)} \in \mathbb{R}^{d\times n},\ t^{(i)} \in \{1, \ldots, L\} \tag{11}$$

where d is the number of DOF of the system, n is the number of data frames in the sample, t is the target class number and L is the number of classes of the set.

Table 1
Accuracy on the validation set of the classifiers trained only with full gesture data.

	DTW	FE1-ANN	FE2-ANN
Accuracy (%)	77.02	86.67	85.72

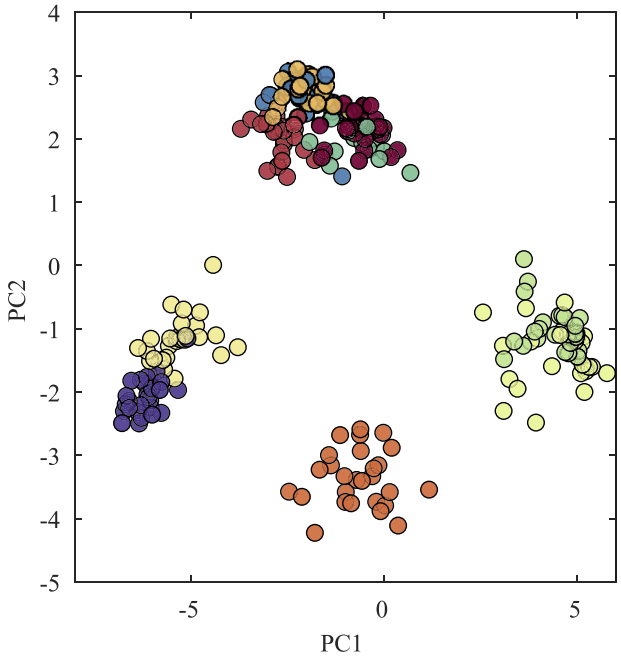

Fig. 3. Distribution of features (FE2 with 100% of data) in a reduced principal component space. Only the first 10 classes are represented and colours discriminate the classes.

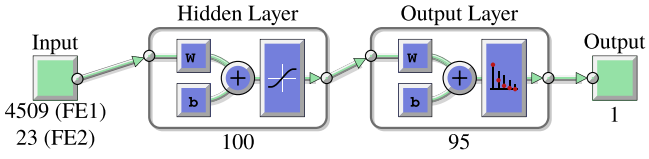

Fig. 4. ANN architecture used for classification for test cases FE1 and FE2.

3.2. Feature extraction

Two distinct feature extraction methods are proposed:

1. Re-sampling the samples with bicubic interpolation (FE1);
2. Extracting principal vectors and values using PCA (FE2).

In the first case, FE1, given a sample $\mathbf{X}^{(i)}$: $i \in i^D$ with n frames ($\mathbf{X}^{(i)} \in \mathbb{M}^{22\times n}$), the goal is to resample it to a fixed size p. The number p can be chosen arbitrarily, but in order to reduce the number of features, p should be below a lower bound such that $p \leq n,\ \forall n|\mathbf{X}^{(\mathbf{i})} \in \mathbb{M}^{22\times n}$. For this data set, this lower bound is 41 and we chose a $p = 20$, which is about half of the minimum original gesture length. Applying the bicubic interpolation algorithm results in a matrix $\mathbf{X}' \in \mathbb{R}^{22\times p}$. The following step is to transform $\mathbf{X}'$ into a vector $\mathbf{z} \in \mathbb{R}^{22p\times 1}$, which is done by concatenating every frame vertically:

$$\mathbf{z}^{(i)} = \begin{pmatrix} \mathbf{Z}^{(i)}_{\bullet 1} \\ \vdots \\ \mathbf{Z}^{(i)}_{\bullet (28p)} \end{pmatrix} \tag{12}$$

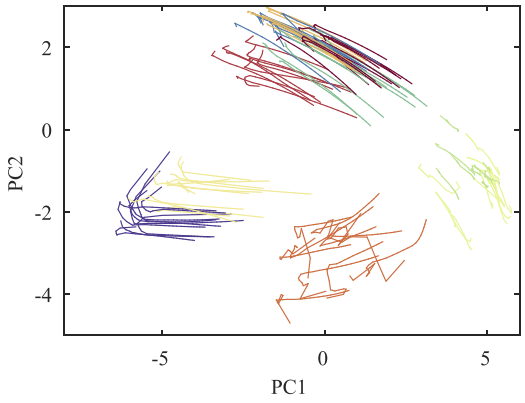

Fig. 5. Representation of the evolution of FE2 features over gesture completion rate. Each line represents one sample and its colour refers to the target class.

Table 2
Classification accuracy on the validation data set for test case FE2.

	Accuracy over time (%)				
Classifier	25%	50%	75%	100%	Mean
ANN	93.06	90.88	88.31	81.92	88.08
SVM	83.79	81.29	78.18	73.89	80.01

In a second case, FE2, we use PCA to extract features. The advantage is that it allows us to obtain features from incomplete gestures and still obtain coherent features. From each sample $\mathbf{X}^{(i)} : i \in i^D$, $\mathbf{X} \in \mathbb{M}^{22\times n}$ we can extract b feature vectors $\mathbf{z}_k^{(i)} : k \in]0, 1]$, where k defines the fraction of the number of frames that were used:

$$\mathbf{z}_k^{(i)} = \mathbf{U}_{\bullet 1} \tag{13}$$

where $\mathbf{U}_{\bullet 1}$ is the first singular vector. The singular vector has the same dimensionality as the data source. It is calculated using the partial sample $\mathbf{X}_{\bullet \mathbf{m}}^{(i)}$, so that:

$$\text{pca}(\mathbf{X}_{\bullet \mathbf{m}}^{(i)}), \quad \mathbf{m} = \{1, \ldots, \lceil nk \rceil\} \tag{14}$$

where n is the number of frames of the sample and $\lceil nk \rceil$ represents the ceiling function, since $\lceil nk \rceil \in \mathbb{N}$. Therefore, $\lceil nk \rceil$ represents the cutoff frame, that is, the last frame within the sample that is used for feature extraction.

The last feature processing step is feature scaling. Feature scaling is essential for achieving smaller training times and often better classification performance with less training. It harmonizes the values of different features so that all of them fall within the same range. This is especially important when some features have distinct orders of magnitude. The scaling function chosen was linear rescaling, l:

$$l(\mathbf{x}) = \frac{2\mathbf{x} - \widehat{\mathbf{X}}^T}{\widehat{\mathbf{X}}^T} \tag{15}$$

where $\widehat{\ }$ is the max+min operator defined in (16). $\mathbf{X}^T = \left(\cup \mathbf{z}^{(i)} : i \in i^T\right)$ is the set of unscaled features of the training set. This operator is valid both for static and dynamic gestures but the sample subsets used should be exclusive.

$$\widehat{\mathbf{X}}_i = \max \mathbf{X}_{i\bullet} + \min \mathbf{X}_{i\bullet}, \quad i = 1, \ldots, d \tag{16}$$

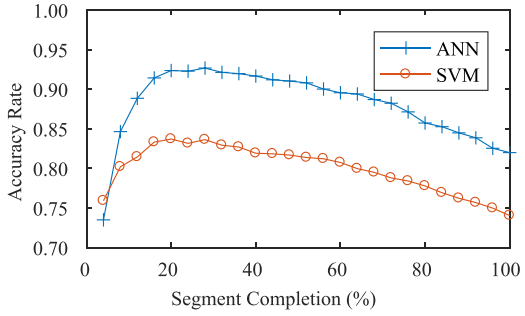

Fig. 7. Evolution of the accuracy with percentage of data used.

3.3. Results and discussion

The available samples $\mathcal{S}(i) : i \in \mathbf{i}^D$ were divided in two sets of approximately the same size: a training set ($i \in \mathbf{i}^{DT}$) and a validation set ($i \in \mathbf{i}^{DV}$). Each set has about the same number of samples per class. In this case, we use all the available samples in the data set, 27 samples per each of the 95 classes (2565 samples total). All the accuracy results presented in this study correspond to those obtained from the validation set.

We present the results for the DTW approach, FE1 and FE2 (full gesture) in Table 1. The DTW classifier achieved a 77.02% accuracy rate, while the proposed feature sets FE1 and FE2 achieved significantly better results (86.67% and 85.72%, respectively) when discriminated with an ANN. A representation of the features for the first 10 classes for FE2 is shown in Fig. 3. In this $\mathbb{R}^2$ space, the classes show good separability, with low intra-class dispersion.

The ANN architecture is composed by one hidden layer with 100 nodes and 95 output neurons (classes) in both approaches, Fig. 4. The sole difference between FE1 and FE2 is the size of the input feature vector, 440 and 23 for FE1 (12) and FE2 (13), respectively. In both cases the transfer function is the hyperbolic tangent in the first layer and the *softmax* function in the second layer.

While for the FE1 case the gesture frames are interpolated and the features are extracted after the gesture is finished, for FE2 the

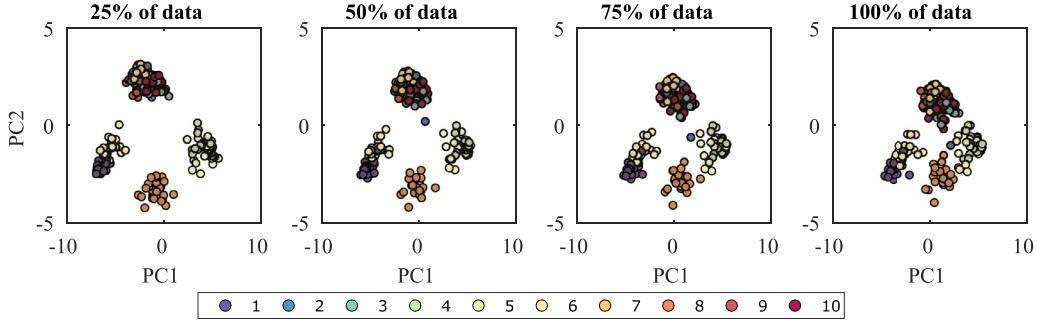

Fig. 6. Plots of the features obtained from the validation data set (including the sets with incomplete data – 25%, 50%, 75% and 100% of the data) in a reduced principal component space. The features were centered and scaled. Each color represents a different class.

Fig. 8. Visualization of different stages of robot teleoperation process: (a) starting point, (b) virtual joystick guidance to a goal, (c) forceful stop command, (d) rotation of the end-effector, (e) gesture-command to open the gripper, (f) grabbing a bottle and puling it up, (g) rotation of the end-effector, (h) safe collaboration with the robot. NOTE: The virtual joystick mode moves the end-effector in a direction defined by the vector that joins a center position in which the hand is closed and the position of the hand when it is moved.

features can be extracted at any time during a gesture. This allows for recognition even before the gesture is even finished. If we plot the evolution of the features of a gesture sample over time, we obtain what is shown in Fig. 5. Over time, the features still form defined clusters, which are precursors to good classifiers. Given that, the ANN for FE2 was trained and validated with 25 sets of features originated from $\mathbf{i}^{DT}$, $\mathbf{z}_k^{(i)} : k \in \{1/25, 2/25, 3/25, \ldots, 25/25\}$, see (13).

For FE2, the accuracy results are displayed in Table 2. In this table we are displaying the accuracy using 25, 50, 75 and 100% of the gesture data. The best accuracy, reaching 93.06%, was obtained when only about 25% of the initial gesture data was used. When more data was used, the accuracy decreased to 81.92%, Fig. 7. It is also possible to see the evolution of the FE2 features obtained from 25, 50, 75 and 100% of the data in Fig. 6. Even at 25% there is already good separation of the classes and the clusters are maintained over time. There are very few studies concerning the classification with partial data, so it is difficult to compare our results. Most of the community assumes that more data should report better accuracy. In this study we showed that it is possible to obtain state of the art accuracy in a real dataset with as little as 25% of the initial gesture data. Using the same proposed PCA features, the accuracy decreases over time. This can be justified by the fact that the initial part of the gesture has better linear correlation between variables, so the PCA coefficients are more stable. As the number of frames increases, the gesture becomes more complex and it is not as easily described by just the first principal vector.

The UCI Auslan dataset has been applied in a number of studies related with pattern classification. In a recent study the authors randomly selected four subsets of the whole data set with each subset containing 20 categories [18]. It is reported that the best classification results were obtained with the proposed order preserving sparse coding method (MTO-SC), with an accuracy of about 94% in the classification of subsets with 20 categories. In [4] it is reported an accuracy of about 94% using SVM and logistic regression models. A Dual Square-Root Function (DSRF) descriptor obtained by calculating gradient-based shape features of normalized rigid body motion trajectories was applied with an accuracy of 88%, [5].

3.4. Interacting with a real robot

The proposed classification system was tested in a real robot. For the HRI process, the robot is controlled using gestures in a collaborative task: preparing a breakfast meal. This is a combination of single robot tasks such as pick, place, hold, and carry actions. In such task the robot grabs a cereal box (pouring the contents into a bowl) and grabs a yogurt bottle (also pouring its contents into the same bowl), Fig. 8. Our setup is composed by a robot with 6 DOF, a data glove and a magnetic tracker. From the 95 gestures in the UCI Auslan dataset we are only using/recognizing 10 gestures from one hand/arm. The classified gestures are used as input to teleoperate the robot, i.e., they are directly associated to the robot commands: stop motion, move along X, Y or Z in Cartesian space, rotate the robot end-effector in turn of X, Y or Z, and open/close the gripper. In practice, the human âguidesâ the robot to close the target objects using specific gestures, and then open or close the gripper using other gestures. A specific gesture is associated to robot STOP command, i.e., the human can stop the robot at any instant. This teleoperation mode allows to use the robot as a tool, in which the human can save robot target points so that the robot can replicate the task if objects are in same poses or if the robot has perception abilities to adjust to new positions of objects.

4. Conclusions

This paper demonstrated that dynamic gesture data can be subject to DDR making the classification process more efficient: increase accuracy, reduce training time and classify gesture patterns with incomplete data. We concluded that the classification accuracy is higher with 25% of initial gesture data (93% accuracy) than with 50% (91% accuracy), 75% (88% accuracy) or 100% (83% accuracy) of gesture data. These results, obtained from the classification of 95 different patterns, can be explained by the noise we have in classifiers input data. Recognized gestures proved to be a natural and intuitive human-robot interface.

Future work will be dedicated to explore the ability to classify gesture patterns with initial 25% in the context of anticipation activities in the process of human-robot interaction. In addition, this approach will be tested with other type of input data (obtained from other sensing technology).

Acknowledgements

This work was supported in part by the Portuguese Foundation for Science and Technology (FCT), SFRH/BD/105252/2014, and Portugal 2020 project POCI-01-0145-FEDER-016418 by UE/FEDER through the program COMPETE2020.

References

[1] C. Appert, O. Bau, Scale detection for a priori gesture recognition, in: International Conference on Human Factors in Computing Systems, 2010, pp. 879–882. Atlanta, United States. doi: 10.1145/1753326.1753456. https://hal.inria.fr/inria-00538339.

[2] M. Burke, J. Lasenby, Pantomimic gestures for human–robot interaction, IEEE Trans. Rob. 31 (5) (2015) 1225–1237, doi:10.1109/TRO.2015.2475956.
[3] S. Calinon, A. Billard, Incremental learning of gestures by imitation in a humanoid robot, in: Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on, 2007, pp. 255–262, doi:10.1145/1228716.1228751.
[4] F. Dalvi, H. Cate, Z. Hussain, Sign Language Recognition using Temporal Classification, Stanford University (2015).
[5] Y. Guo, Y. Li, Z. Shao, Dsrf: a flexible descriptor for effective rigid body motion trajectory recognition, in: 2016 IEEE International Conference on Mechatronics and Automation, 2016, pp. 1673–1678, doi:10.1109/ICMA.2016.7558815.
[6] H.G. Jung, Support vector number reduction by extending iterative preimage addition using genetic algorithm-based preimage estimation, Pattern Recognit. Lett. 84 (2016) 43–48. http://dx.doi.org/10.1016/j.patrec.2016.08.003. URL: http://www.sciencedirect.com/science/article/pii/S0167865516301970
[7] M. Kadous, M.W. Kadous, S.C. Sammut, Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series, School of Computer Science and Engineering, PhD Thesis, University of New South Wales, 2002.
[8] D. Kelly, J. Mc Donald, C. Markham, Weakly supervised training of a sign language recognition system using multiple instance learning density matrices, IEEE Trans. Syst. Man Cybern. B Cybern. 41 (2) (2011). 526–41. doi: 10.1109/TSMCB.2010.2065802.
[9] R. Keys, Cubic convolution interpolation for digital image processing, IEEE Trans. Acoust. 29 (6) (1981) 1153–1160, doi:10.1109/TASSP.1981.1163711.
[10] H.S. Koppula, A. Saxena, Anticipating human activities using object affordances for reactive robotic response, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 14–29, doi:10.1109/TPAMI.2015.2430335.
[11] D. Kulic, W. Takano, Y. Nakamura, Online segmentation and clustering from continuous observation of whole body motions, IEEE Trans. Rob. 25 (5) (2009) 1158–1166, doi:10.1109/TRO.2009.2026508.
[12] Z. Lai, W.K. Wong, Y. Xu, J. Yang, D. Zhang, Approximate orthogonal sparse embedding for dimensionality reduction, IEEE Trans. Neural Netw. Learn Syst. 27 (4) (2016) 723–735, doi:10.1109/TNNLS.2015.2422994.
[13] Z. Lai, Y. Xu, J. Yang, L. Shen, D. Zhang, Rotational invariant dimensionality reduction algorithms, IEEE Trans Cybern PP (99) (2016) 1–14, doi:10.1109/TCYB.2016.2578642.
[14] J.F.S. Lin, V. Joukov, D. Kulic, Full-body multi-primitive segmentation using classifiers, in: 2014 IEEE-RAS International Conference on Humanoid Robots, 2014, pp. 874–880, doi:10.1109/HUMANOIDS.2014.7041467.
[15] P. Lin, J. Zhang, R. An, Data dimensionality reduction approach to improve feature selection performance using sparsified svd, in: 2014 International Joint Conference on Neural Networks (IJCNN), 2014, pp. 1393–1400, doi:10.1109/IJCNN.2014.6889366.
[16] V. Magnanimo, M. Saveriano, S. Rossi, D. Lee, A bayesian approach for task recognition and future human activity prediction, in: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014, pp. 726–731, doi:10.1109/ROMAN.2014.6926339.
[17] P. Neto, D. Pereira, J.N. Pires, a.P. Moreira, Real-time and continuous hand gesture spotting: an approach based on artificial neural networks, in: 2013 IEEE International Conference on Robotics and Automation, 2013, pp. 178–183, doi:10.1109/ICRA.2013.6630573.
[18] B. Ni, P. Moulin, S. Yan, Order preserving sparse coding, IEEE Trans Pattern Anal Mach Intell 37 (8) (2015) 1615–1628, doi:10.1109/TPAMI.2014.2362935.
[19] S. Rossi, E. Leone, M. Fiore, A. Finzi, F. Cutugno, An extensible architecture for robust multimodal human-robot communication, in: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013, pp. 2208–2213, doi:10.1109/IROS.2013.6696665.
[20] A. Shyr, R. Urtasun, M.I. Jordan, Sufficient dimension reduction for visual sequence classification, in: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010, 2010, pp. 3610–3617, doi:10.1109/CVPR.2010.5539922.
[21] M. Simão, P. Neto, O. Gibaru, Taking advantage of data dimensionality reduction for dynamic gesture recognition from incomplete data, Workshop on Behavior Adaptation, Interaction and Learning for Assistive Robotics, IEEE RO–MAN 2016, NYC, USA, 2016.
[22] M.A. Simão, P. Neto, O. Gibaru, Unsupervised gesture segmentation by motion detection of a real-time data stream, IEEE Trans. Ind. Inf. in press (99) (2016). 1–1. doi: 10.1109/TII.2016.2613683.
[23] Y. Song, D. Demirdjian, R. Davis, Continuous body and hand gesture recognition for natural human-computer interaction, ACM Trans. Interact. Intell. Syst. 2 (1) (2012) 1–28, doi:10.1145/2133366.2133371.
[24] M.T. Wolf, C. Assad, M.T. Vernacchia, J. Fromm, H.L. Jethani, Gesture-based robot control with variable autonomy from the JPL BioSleeve, in: 2013 IEEE International Conference on Robotics and Automation, IEEE, 2013, pp. 1160–1165, doi:10.1109/ICRA.2013.6630718.
[25] S. Yi, Z. Lai, Z. He, Y. ming Cheung, Y. Liu, Joint sparse principal component analysis, Pattern Recognit. 61 (2017) 524–536. http://dx.doi.org/10.1016/j.patcog.2016.08.025. URL: http://www.sciencedirect.com/science/article/pii/S0031320316302370

Please cite this article as: M. Simão et al., Using data dimensionality reduction for recognition of incomplete dynamic gestures, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.01.003