J. Phys. Ther. Sci.
27: 2649–2653, 2015

Review

Locality sensitivity discriminant analysis-based feature ranking of human emotion actions recognition

Nurnadia M. Khair[1)*], M. Hariharan[1)], S. Yaacob[2)], Shafriza Nisha Basah[1)]

[1)] *School of Mechatronic Engineering, Universiti Malaysia Perlis (UniMAP): 02600, Kampus Pauh Putra, Perlis, Malaysia*
[2)] *Universiti Kuala Lumpur Malaysian Spanish Institute, Kulim Hi-Tech Park, Malaysia*

Abstract. [Purpose] Computational intelligence similar to pattern recognition is frequently confronted with high-dimensional data. Therefore, the reduction of the dimensionality is critical to make the manifold features amenable. Procedures that are analytically or computationally manageable in smaller amounts of data and low-dimensional space can become important to produce a better classification performance. [Methods] Thus, we proposed two stage reduction techniques. Feature selection-based ranking using information gain (IG) and Chi-square (Chisq) are used to identify the best ranking of the features selected for emotion classification in different actions including knocking, throwing, and lifting. Then, feature reduction-based locality sensitivity discriminant analysis (LSDA) and principal component analysis (PCA) are used to transform the selected feature to low-dimensional space. Two-stage feature selection-reduction methods such as IG-PCA, IG-LSDA, Chisq-PCA, and Chisq-LSDA are proposed. [Results] The result confirms that applying feature ranking combined with a dimensional-reduction method increases the performance of the classifiers. [Conclusion] The dimension reduction was performed using LSDA by denoting the features of the highest importance determined using IG and Chisq to not only improve the effectiveness but also reduce the computational time.
Key words: Emotion, Actions, Dimensional feature reduction

(This article was submitted Apr. 17, 2015, and was accepted May 18, 2015)

INTRODUCTION

The recent increase of dimensionality of data poses a severe challenge to many existing data mining, pattern recognition, machine learning, artificial intelligence, feature selection, and dimensionality-reduction methods with respect to efficiency and effectiveness. The problem with a high-dimensional dataset is that many of the features are irrelevant and redundant. This increases the search space size and results in a greater challenge to further process the data. This issue of dimensionality is a major obstacle in machine learning and data mining as it can lead to inferior performance of the classifier. The growing importance of knowledge discovery and data mining methods in practical applications has made feature-selection and dimensionality-reduction problems a controversial issue, especially with mining knowledge from databases with enormous numbers of records and columns. In the research field of human emotion recognition, feature selection and reduction, employed as a preprocessing step to data mining, image processing, conceptual learning, and machine learning, have been effective in reducing dimensionality, removing irrelevant and redundant data, increasing learning accuracy, and improving comprehensibility. Based on these merits, it is considered an important and necessary preprocessing step before the implementation of algorithms.

Dimensionality-reduction approaches can be broadly classified into two categories: feature selection and dimensional feature reduction. Feature selection comes under three categories: wrappers, filters, and embedded methods, which are the most effective techniques to select the most relevant features from an original dataset. Wrapper-model techniques evaluate features using the learning algorithm "wrap" as a search strategy that performs through the space of the feature subsets. The majority of feature selections in wrapper methods are genetic algorithm (GA)[1)] and sequential forward selection (SFS)[2)]. Filter-based approaches virtually always rely on the class labels, most commonly assessing correlations between the features and class label. Typical filtering methods include mutual information[3)], information gain[4)], fast correlation-based filter (FCBF)[5)], correlation-based feature selection (CFS)[6)], and fisher score[7, 8)]. These methods utilize a metric for determining the relevance or importance of each feature. Another method is the embedded method, which considers different subsets of the features during the learning process. Subsets are evaluated based on their ability to support correct classification of the training examples. Examples of this method are the weights of logistic regression[9)], random forest[10, 11)], and weight vector of SVM[12)]. The advantages of feature selection include improving model

*Corresponding author. Nurnadia M. Khair (E-mail: nurnadia1488@gmail.com)

Table 1. Summary of feature-reduction techniques in human emotion in action recognition

Ref.	First Author (Year)	Emotional State	Movement	Feature	Feature extraction / Data reduction Method	Classification/ Model	Accuracy
[13]	Lars Omlor (2006)	Happy, angry, sad, fear, normal	Muscle activity and Walking	Joint angle	PCA, fast ICA, Bayesian ICA, New algorithm	-	100%
[14]	Michelle Karg (2010)	Sad, angry, happy, neutral	Walking	Position, joint angle, joint center	PCA,KPCA, PCA-FT-PCA, KPCA-FT-PCA, PCA-FT-KPCA, KPCA-FT-KPCA	SVM and 1NN, Naive Bayes	PCA-FT-PCA in Naïve Bayes: 72%
[15]	Michelle Karg (2009)	Sad, happy, angry, neutral, PAD model (Displeased, content, bored, excited, obedient, dominant)	Walking	Velocity, stride length, cadence, joint angle	PCA, KPCA, LDA, GDA	SVM and 1NN, Naive Bayes	95%
[16]	Liyu Gong (2010)	Happy, anger, neutral, sad	Knocking	Distance, speed, acceleration, jerk	SOG descriptor	SVM	76.42%
[16]	Xin Zhao (2013)	-	Box, gestures, jog, throw-catch, walk	Joint position	SDG, SELF, PCA,LPP, LDA, SDA	kNN	95.8%

interpretability, reducing training times, and enhancing generalization by reducing over fitting.

Dimensional feature reduction is a technique employed to transform an existing high-dimensional feature space to a lower-dimensional feature space, which is then suitable for discriminative analysis. Dimension-reduction techniques can be categorized into supervised or unsupervised and linear or nonlinear. These techniques are widely used in human recognition. Table 1 summarizes several techniques of applying human emotion and action recognition.

Despite numerous approaches in the literature, feature reduction remains an ongoing research topic. Researchers continue to search for new techniques to select distinctive features such that the classification accuracy can be improved and the processing time reduced. The contributions of this paper are as follow: 1) propose a feature-selection method-based ranking to identify informative features; 2) propose a new feature-reduction method known as locality discriminant analysis (LSDA) and compare the result with the traditional principal component analysis (PCA) method; 3) propose two-stage feature-reduction techniques. The remainder of this paper is organized as follows. In subject subsection describes the database used in this experiment and feature extraction. In method subsection reviews the description of feature-based ranking approaches, the LSDA method and classification method. In result subsection presents the experimental result and final subsection, presents the discussion.

SUBJECTS AND METHODS

Subjects

In this study, a motion-captured database recorded at the Psychology Department, University of Glasgow[17] was used. The dataset contained 30 nonprofessional actors (15 males and 15 females) where each performed five actions (walk, knock, lift, throw, and a combinations of these four actions) in four emotional contexts (angry, happy, sad, and neutral). The stated results are based only on the knocking, lifting, and throwing motions from the database. Fifteen captured body joints were considered where each represented a point in 3-dimensional (3D) space, $p = [x, y, z]$, where x, y, z are the Cartesian coordinates of the marker's position. The five possible statistical features presented in Table 2 extracted from the magnitude position, speed, acceleration, jerk, angle pitch, and angle yaw were considered as features.

Table 2. Summary of feature set representations

Dynamic features	Statistical features
Position	Mean, max, min, stdev, median
Velocity	Mean, max, min, stdev, median
Acceleration	Mean, max, min, stdev, median
Jerk	Mean, max, min, stdev, median
Angle at pitch	Mean, max, min, stdev, median
Angle at yaw	Mean, max, min, stdev, median

Methods

Feature ranking-based

As discussed in the previous section, there are many techniques for the selection of distinctive features in emotion recognition. In this study, two featured ranking-based techniques, information gain (IG) and Chi-square (Chisq), are proposed because these techniques have been proven effective[4].

1. Information Gain

IG is identified as a measure of dependence between the feature and the class label. It is one of the most popular feature-selection techniques because it is easy to compute and simple to interpreting measures the amount of information

the presence or absence of a term contributes to determining the correct classification decision on a class. IG attains its maximum value if a term is an ideal indicator for class association; that is, if the term is present in a document if and only if the document belongs to the respective class. The IG of a feature X and the class labels Y are calculated as:

$$IG(X,Y)=H(X)-H(X|Y)$$

Entropy (H) is a measure of uncertainty associated with a random variable. H (X) and H (X|Y) are the entropy of X and the entropy of X after observing Y, respectively.

$$H(X)=-\Sigma_i P(x_i)log_2(P((x_i))$$

$$H(X|Y)=-\sum_j P(y_j)\sum_i P(x_i \mid y_j)log_2(P((x_i \mid y_j))$$

The maximum value of IG is one. This indicates that a feature with a high information gain is relevant. IG is evaluated independently for each feature and the features with the top-k values are selected as the relevant features.

2. Chi-square

One of the most popular feature-selection approaches is Chisq. It is used to assess two types of comparison: test of goodness of fit and test of independence. In selecting a feature, Chisq is used as a test of independence to assess whether the class label is independent of a particular feature. The Chisq score for a feature with *t* different values and *C* classes is computed using:

$$Chisq(t,C)=\sum_{i=1}^{t}\sum_{j=1}^{C}\frac{\left(N_{t,C}-E_{t,C}\right)^2}{E_{t,C}}$$

Where $N_{t,C}$ is the number of samples with the i^{th} feature value and:

$$E_{t,C}=\frac{N_{*j}N_{i*}}{N}$$

where N_{i*} is the number of samples with the i^{th} value for the particular feature, N_{*j} is the number of samples in class *j*, and *N* is the number of samples.

Feature reduction

Upon completion of the preprocessing step, the terms of high importance in the documents are acquired through the Chisq method. Although the number of features is reduced, the main problem, the high dimensionality of the feature space, remains. Therefore, to reduce the feature space dimension and the computational complexity of the machine learning algorithms used in the emotion recognition and to increase the performance, the proposed method based on LSDA is applied. The aim of these methods is to minimize information loss while maximizing the reduction in dimensionality.

1. Optimal feature reduction through LSDA

LSDA is an improvement from linear discriminant analysis (LDA), a supervised feature-selection problem described by Cai D et al.[18], which respects both discriminant and geometrical structure in the data manifold by building a nearest neighbor graph. For example, LSDA is widely used in image processing recognition. To improve the discriminative ability of the low-dimensional features, the class label information is incorporated into the feature extraction process. Assume a set of labeled points $X[x_1,x_2,\ldots,x_N]\epsilon R^{D\times N}$ denoting data points in a *D* dimensional space where the data points belong to *C* class (each class contains n_c,c=1,2, …,*C* samples, $\sum_{c=1}^{C} n_c = N$)[23].

The algorithmic procedure is formally stated below:

(i) Construct a nearest neighboring graph *G* by placing an edge between each sample and its *k* nearest neighbors. Let $N(x_i)=\{x_i^1,\ldots,x_i^k\}$ be the set of *k* nearest neighbors of x_i Then, the weight matrix of *G* in LSDA is given by:

$$W_{i,j}=\begin{cases}1, & if x_i \in N(x_j) or\ x_j \in N(x_i)\\ 0, & otherwise\end{cases}$$

(ii) The nearest neighboring graph is partitioned into two parts: a within-class graph (G_w) and a between-class graph (G_b) For each sample x_i (*i*=1, …, *N*), the set of its k nearest neighbors can be divided into two subsets $N_w(x_i)$ and $N_b(x_i)$. $N_w(x_i)$ is the set of neighbors sharing the same label with x_i. The definitions of $N_w(x_i)$ and $N_b(x_i)$ are as follows:

$$N_w(x_i)=\{x_i^j \mid l(x_i^j)=l(x_i), 1\le j\le k\}$$

$$N_b(x_i)=\{x_i^j \mid l(x_i^j)\ne l(x_i), 1\le j\le k\}$$

where $l(x_i)$ is the class label of x_i Clearly, $N_w(x_i)\cap N_b(x_i)=\varnothing$ and $N_b(x_i)\cup N_b(x_i)=N(x_i)$.

(iii) The adjacent weight matrices of G_b and G_w are defined as W_b and W_w in:

$$W_{b,ij}=\begin{cases}1, & if x_i \in N_b(x_j) or x_j \in N_b(x_i)\\ 0, & otherwise\end{cases}$$

$$W_{w,ij}=\begin{cases}1, & if x_i \in N_w(x_j) or x_j \in N_w(x_i)\\ 0, & otherwise\end{cases}$$

Then, it is found that $W=W_b+W_w$, which means that the nearest neighboring graph G is a combination of between-class graph (G_b) and within-class graph (G_w).

The objective of LSDA is to obtain a low-dimensional feature space where the nearby points with the same label are close to each other, whereas the nearby points with different labels are further apart. Thus, a reasonable criterion is to optimize the following two functions:

$$\min\frac{1}{2}\sum_{ij}(y_i-y_j)^2W_{w,ij}$$

$$\max\frac{1}{2}\sum_{ij}(y_i-y_j)^2W_{b,ij}$$

Where $y_i \epsilon R^{d\times 1}$ ($d \ll D$) is the feature extraction result of x_i

Assume that the low-dimensional features of the input data can be obtained by a transformation matrix A, that is $y_i = A^T x_i$, 1,2, …, N. After a series of deductions, the objective functions of LSDA can be derived as:

$$\arg\max_A A^T X(\beta L_b + (1-\beta)W_w)X^T A$$

$$A^T X D_w X^T A = 1$$

where D_w and D_b are diagonal matrices whose entries are the column (or row, as W_w and W_b are symmetrical) sum of W_w and W_b . That is $D_{w,ii}=\sum_j W_{w,ij}$ and $D_{b,ii}=\sum_j W_{b,ij}$. The Laplacian matrix of G_b is $L_b=D_b-W_b$ and β is a regulative parameter with $0 \leq \beta \leq 1$. The final transformation, matrix A, is obtained by maximizing the generalized eigenvalues problem:

$$A^T X(\beta L_b+(1-\beta)W_w)X^T A=\lambda X D_w X^T A$$

2. PCA

PCA is a common feature-reduction method in human action recognition. We compare the proposed algorithm with this traditional method. The methods were separately applied to the classification of datasets where the dimension acquired at the end of the PCA and LSDA application was reduced.

Classification

In this study, kNN classifier is used owing to its simplicity and accuracy for emotion recognition. The reason for using a classifier is to compare the performances of the methods in emotion recognition. Among the 30 subjects, knocking provided 1,200 trials, lifting added 1,140 trials, and throwing provided 1,190 trials when they were processed for each emotion. The ability of the statistical feature set was identified by a maximum accuracy from averaging ten times, where was noted as equal to one to ten over a tenfold cross-validation.

RESULTS

An exhaustive study was performed to compare feature-selection methods, dimensionality-reduction methods, classification accuracy, and processing time. The result was performed with maximum average accuracy and geometrical mean (G-mean). Table 3 presents the accuracy rates of the two feature-selection methods, IG and Chisq. In Table 4, two different feature-reduction methods, PCA and LSDA, are compared. In Table 5, the results of the two-stage feature reductions, IG-PCA, IG-LSDA, Chisq-PCA, and Chisq-LSDA are presented. The entire evaluation was performed using kNN and a 10-fold cross validation.

DISCUSSION

In this study, feature ranking-based, feature reduction, and two-stage feature selection-reduction were proposed to reduce the high dimensionality of a feature space composed of a large number of terms and remove redundant and irrelevant features from the feature space and thereby improve the performance of emotion recognition in different actions. In feature selection, action features were ranked based on their importance and the best features were selected using the IG and Chisq methods.

Table 3. Accuracy rate of feature-selection method using kNN classifier

	Knocking		Lifting		Throwing	
	Average	G-mean	Average	G-mean	Average	G-mean
Original	84.83	84.65	84.04	83.69	83.57	86.32
IG	85.00	84.73	84.30	83.93	86.55	86.44
Chi-square	85.00	84.76	85.09	83.95	86.64	87.00

Table 4. Accuracy rate of feature-reduction method using kNN classifier

	Knocking		Lifting		Throwing	
	Average	G-mean	Average	G-mean	Average	G-mean
PCA	70.50	69.84	61.93	60.71	62.86	61.85
LSDA	98.75	98.75	97.46	97.42	97.19	97.17

Table 5. Two-stage performance of combination of feature selection and reduction using kNN classifier

	Knocking		Lifting		Throwing	
	Average	G-mean	Average	G-mean	Average	G-mean
IG-PCA	70.42	69.85	61.67	60.17	61.68	60.87
IG-LSDA	98.83	98.83	97.72	97.68	97.23	97.20
Chi-square -PCA	70.08	69.61	62.37	61.23	61.85	61.06
Chi-square -LSDA	98.75	98.75	97.72	97.68	96.81	96.77

Feature ranking-based evaluation

Straightforward feature-selection procedures based on an evaluation of the predictive power of the individual features and ranking based on choosing the best features were evaluated in this study. From the results in Table 3, we can observe that the best feature ranking method in all actions was achieved using Chisq. Owing to its simplicity, scalability, good empirical success, and attractiveness, the method was successful in improving accuracy by 0.2% for knocking, 1.23% for lifting, and 3.54% for throwing when compared with the original features. However, the number of features selected was excessively large leading to a high computational time in the classification. Thus, feature reduction was proposed.

Feature reduction by feature transformation

In this section, the details of the performance of the feature-reduction approaches are discussed. Table 4 summarizes the results of PCA and LSDA. We determined that PCA and LSDA transformed the feature vectors with reduced numbers to a lower dimensional space. Furthermore, the recognition performance that occurred when the 450 original

Table 6. Computational time (s) of feature-reduction methods in kNN classifier.

	Original	IG	Chisq	PCA	LSDA	IG-PCA	IG-LSDA	Chisq-PCA	Chisq-LSDA
Knocking	1.55	1.64	1.66	1.16	1.14	1.01	0.69	1.23	0.67
Lifting	1.77	1.56	1.73	0.53	1.15	1.62	0.83	1.10	0.63
Throwing	2.06	2.77	1.78	0.49	0.92	0.66	0.90	0.51	0.68

features transformed to a 30 dimensional space was superior with the LSDA method. From the table, we observe that the LSDA feature-reduction techniques are were to outperform accuracy of other techniques and improved the original accuracy immediately. For example, LSDA improved the accuracy 14.1% for knocking, lifting, and throwing when compared with the original features and 14%, 13%, and 11% when compared with the Chisq method.

Two-stage feature-reduction ranking-based performance

In this section, the efficiency of the proposed two-stage feature-reduction-based methods, IG-PCA, IG-LSDA, Chisq-PCA, and Chisq-LSDA were evaluated. In Table 5, the comparison between IG-LSDA and Chisq-LSDA indicate that both approaches provided good identification rates, being marginally better than LSDA. Further, IG-LSDA achieved an identification rate of 98.83% for knocking, 97.72% for lifting, and 97.23% for throwing with ranking based of 30 features; IG-PCA and Chisq-PCA did not provide competitive results for this number of features.

Computational Time

The algorithmic complexities of all the feature-selection, feature-reduction methods considered in this study were computed to be the same. Therefore, the processing time of all the techniques was investigated and compared. The measurements were based on a computer equipped with an Intel Core i5 1.6 GHz processor and 6 GB of RAM. The results of the timing analysis, which are presented in Table 6, indicate that the two-stage IG-LSDA and Chisq-LSDA were the fastest methods. This also confirms that the contribution of the feature ranking method was successful in improving computational time in the LSDA method.

These methods were successful and indicated marginal improvement compared to the original features using ranking basis. However, the number of features selected remained excessive and the methods required high computational time for the classification. Owing to this problem, feature reduction was proposed. For example, PCA and LSDA feature reduction was applied separately to transform the original features to a lower dimensional-reduction space. The aim of these methods was to minimize information loss while maximizing the reduction in dimensionality. The results confirmed that a feature-reduction method using LSDA was superior to the original feature and feature-selection methods. This approach was also successful for improving the computational time for the classification. Finally, a two-stage feature selection-reduction method was applied to feature ranking based on a dimensional-reduction method. This was an effective contribution confirmed by the increase of accuracy and improved performance of the classifier. This means that the dimension reduction performed an LSDA by denoting the features of the highest importance determined via IG and Chisq not only improved the effectiveness but also reduced the computational time.

REFERENCES

1) Park CH, Sim KB: The novel feature selection method based on emotion recognition system. System, 2006, 4115: 731–740.
2) Rong J, Li G, Chen YP: Acoustic feature selection for automatic emotion recognition from speech. Inf Process Manage, 2009, 45: 315–328. [CrossRef]
3) Koller D, Building G: Toward Optimal Feature Selection. International Conference on Machine Learning, 1996, 284–292.
4) Bang SW, Kim J, Lee JH: An approach of genetic programming for music emotion classification. International Journal of Control. Autom Syst, 2013, 11: 1290–1299. [CrossRef]
5) Gharavian D, Sheikhan M: Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Comput Appl, 2012, 21: 2115–2126.
6) Vogt T, Andre E: Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. Proc. of Multimedia and Expo (ICME05), 2005.
7) Rani P, Liu C, Sarkar N, et al.: An empirical study of machine learning techniques for affect recognition in human-robot interaction. Pattern Anal Appl, 2006, 9: 58–69. [CrossRef]
8) Song M, Li N, Bu J, et al.: Feature Selection for Fast Speech Emotion Recognition. Proceedings of the 17th ACM International Conference on Multimedia (MM '09), 2009, 753–756.
9) Ma S, Huang J: Regularized ROC method for disease classification and biomarker selection with microarray data. Bioinformatics, 2005, 21: 4356–4362. [Medline] [CrossRef]
10) Fern M: Supervised methods with genomic data: a review and cautionary view. In: Azuaje F, Dopazo J (eds.), Wiley, 2005, pp 193–214.
11) Jiang H, Deng Y, Chen HS, et al.: Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics, 2004, 5: 81. [Medline] [CrossRef]
12) Guyon I, Weston J, Barnhill S, et al.: Gene selection for cancer classification using support vector machines. Mach Learn, 2002, 46: 389–422. [CrossRef]
13) Omlor L, Giese MA: Unsupervised learning of spatio-temporal primitives of emotional gait. Perception and interactive technologies. Proceedings of SIGGRAPH/Eurographics Workshop Graph Hardw, 2006, 4021: 188–192.
14) Karg M, Kühnlenz K, Buss M: Recognition of affect based on gait patterns. IEEE Trans Syst Cybern Biol Cybern, 2010, 40: 1050–1061. [CrossRef]
15) Karg M, Jenke R, Seiberl W, et al.: A Comparison of PCA, KPCA and LDA for Feature Extraction to Recognize Affect in Gait Kinematics. In: Third International IEEE Conference on Affective Computing and Intelligent Interaction and Workshop, 2009, 1–6.
16) Gong L, Wang T, Wang C, et al.: Recognizing Affect from Non-stylized Body Motion Using Shape of Gaussian Descriptors. In: Proceedings of the 2010 ACM Symposium on Applied Computing, 2010, 1203–1206.
17) Ma Y, Paterson HM, Pollick FE: A motion capture library for the study of identity, gender, and emotion perception from biological motion. Behav Res Methods, 2006, 38: 134–141. [Medline] [CrossRef]
18) Cai D, He X, Zhou K, et al.: Locality Sensitive Discriminant Analysis. In: Proc. of 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), 2007.