Real time vision/sensor based features processing for efficient HCI employing canonical correlation analysis

Ehab H. El-Shazly¹ · Moataz M. Abdelwahab¹ · Atsushi Shimada² · Rin-ichiro Taniguchi²

Received: 1 June 2016 / Accepted: 28 September 2016 / Published online: 18 October 2016
© Springer International Publishing Switzerland 2016

Abstract In this paper, a global algorithm for facial and gesture recognition is presented. The algorithm basically consists of three modules: features sensing and processing, dominant features selection and finally features matching. Depending on the type of data used (vision or sensor based), the proposed algorithm exploits multiple features employing 2DPCA that efficiently compact features' descriptors maintain the spatial and temporal alignment of features' components. Canonical Correlation Analysis (CCA) is employed to fuse different features from different descriptors or different performers. CCA also transforms training and testing features sets into new space where similar pairs become highly correlated pairs. Different experiments were conducted using well known data sets in addition to our newly collected data sets to verify the efficiency of the proposed algorithm. Excellent recognition accuracy, and fast performance are factors that promotes the proposed algorithm for real time implementation.

Keywords Feature extraction · Face recognition · Gesture recognition · Interactive system · CCA · 2DPCA

1 Introduction

Over the last decade, face and gesture recognition problem for Human Computer Interaction (HCI) applications had attracted many researchers. Generally HCI systems such as human identification for security, automatic access control and gaming needs robust pattern recognition algorithm. The robustness of the algorithm could be measured in terms of the achieved accuracy, processing complexity and run time. Generally any recognition algorithm consists of two phases, training phase where training data is processed to obtain discriminating features. Obtained features are being processed to remove any redundant and grouped in different classes. In testing phase the testing data is processed in the same way as training data and different classifiers could be used to match the testing set to one of the training sets. Single feature or multiple features could be fused together to improve the recognition accuracy of the proposed algorithm.

Canonical Correlation Analysis (CCA) and its extensions had been widely investigated for fusing different features to improve the recognition process. The basic function of CCA [1] is to measure the relationship between two sets of multidimensional variables and find two new basis vectors such that the correlation between the projections of the input variables using these basis vectors is maximized. Unlike Principal Component Analysis (PCA) [2], Fishers Linear Discriminant Analysis (FLDA) [3] and ICA [4], CCA take multiple modalities data into consideration so that it improves the classification performance. In [5], the author proposed a method called Multiset Canonical Corre-

✉ Ehab H. El-Shazly
ehab.elshazly@ejust.edu.eg

Moataz M. Abdelwahab
moataz.abdelwahab@ejust.edu.eg

Atsushi Shimada
atsushi@limu.ait.kyushu-u.ac.jp

Rin-ichiro Taniguchi
rin@ait.kyushu-u.ac.jp

¹ School of Electronics, Communications, and Computing,
Egypt-Japan University of Science and Technology,
New Borg El Arab City, Alexandria, Egypt

² Department of Advanced Information Technology, Kyushu
University, 744, Motooka, Nishi-ku, Fukuoka 8190395, Japan

lations using Globality-preserving Projections (MCC-GPs) for pattern recognition. Their method achieved dimensionality reduction for high-dimensional data and maintained the global geometric structure of data in a transformed low-dimensional space, so that better classification ability was achieved.

Many modifications and extensions to traditional CCA were presented recently to address better feature fusion [6–8]. In [6], the authors presented an extension of classical CCA called Tensor Canonical Correlation Analysis (TCCA) which is suitable for multidimensional data arrays (or tensors). The reported frame work achieved good results for action classification from video sequences. Local Discrimination CCA (LDCCA) was presented in [7] where the author not only considered the correlations between sample pairs but also the correlations between samples and their local neighborhoods. Discriminative classes were obtained by maximizing the local within-class correlations and minimizing local between-class correlations. In [8] the authors proposed a method called sparse representation based discriminative CCA (SPDCCA) by employing sparse representation and discriminative information with classical CCA.

In the proposed algorithm, CCA was used to fuse different features from either different repetitions or different descriptors or both of them. In addition to features fusion, CCA was used for classification as it transforms training/testing pairs into new space where similar pairs become highly correlated. Depending on using one features descriptor or multiple features descriptors, CCA yields either unique value or vector of values for canonical correlation, respectively. Classification decision is made depending on the value of canonical correlation. In case of using multiple features descriptors (the case for face recognition), weighted sum algorithm is used to find the final value of canonical correlation.

Developed algorithms in literature usually dedicated for certain application (face or gesture recognition) or certain data type, either vision based algorithms [9] or sensor based ones [10]. Development of global system for both face and gesture using either vision or sensor based features is very challenging, due to different characteristics of different features and different challenges for each application. For example face recognition systems should overcome common face images challenges such as variations in illumination conditions, poses, facial expressions, aging, and disguises such as mustaches, beards, glasses or cosmetics.

Wearable devices of all kinds are becoming popular [11,12]. In [11], the authors presented an accelerometer-based digital pen that uses trajectory recognition algorithm to convert time-series acceleration signals into features vectors. The reported algorithm achieved good recognition rates for ten digits recognition, but it requires large training data and many pre-processings filtering. On the other hand, elec-

tromyography (EMG) signals, which detects the activities of related muscles during a gesture performance are more suitable for capturing fine motions such as wrist and finger movements. In [12] the authors developed a Japanese text input method for mobile phones using surface electromyogram (sEMG). The proposed algorithm employed some signal processing techniques, generic dictionary and learning dictionary. sEMG was chosen as the proposed application depends on the fine movement of the hand fingers.

In this paper, we present an efficient and global algorithm for both face and gesture recognition based on vision and sensor based features, respectively. The proposed algorithm can handle different types of data addressing different applications including face recognition, interactive gaming and air writing. Either raw or transform domain features had been extracted and processed using Two Dimensional Principal Component Analysis (2DPCA) to chose only dominant features in the Eigen space. 2DPCA is preferred as it maintains spatial and temporal relation between adjacent pixels or feature components. CCA is employed to find a space where the projection of similar training/testing pairs become highly correlated. The training set that gives highest canonical correlation is recognized as the target set. The framework of the proposed algorithm for different applications is the same with minor adaptations as shown in Fig. 1. This paper is arranged as follows: Sect. 2 introduces briefly CCA. In Sect. 3 the proposed algorithm for face and facial expression recognition is introduced. The adaption of the proposed algorithm for sensor based applications is presented in Sect. 4. The conclusions are presented in Sect. 5.

2 Canonical Correlation Analysis

CCA is a way of measuring the linear relationship between two multidimensional variables. It finds two basis, one for each input variable, that are optimal with respect to correlations and, at the same time, it finds the corresponding correlations. In other words, it finds the two basis in which the correlation matrix between the variables is diagonal and the correlations on the diagonal are maximized. After projection to the new space where the input variables become highly correlated, dimensionality reduction achieved due to projection process (size of basis vectors). The dimensionality of these new basis is equal to or less than the smallest dimensionality of the two input variables.

Given two random sets $f \in R^{m_1 \times n_1}$ and $g \in R^{m_2 \times n_2}$, a pair of transformations u and v called canonical transformations are found so that the correlation between $f' = u^t f$ and $g' = v^t g$ is maximized. The function to be maximized is defined in Eq. (1)

Fig. 1 Block diagram of the general framework proposed for both face and gesture recognition

$$r = \frac{E[f', g']}{\sqrt{E[f'^2]E[g'^2]}} = \frac{E[u^t f, v^t g]}{\sqrt{E[u^t f f^t u]E[v^t g g^t v]}} \\ = \frac{u^t C_{fg} v}{\sqrt{u^t C_{ff} u v^t C_{gg} v}} \quad (1)$$

The maximum of r with respect to u and v is the maximum canonical correlation.

$$r = \max_{u, v} \frac{u^t C_{fg} v}{\sqrt{u^t C_{ff} u v^t C_{gg} v}} \quad (2)$$

r is the canonical correlation. Where $E[h]$ is the expectation of h and C_{ff} , C_{fg} and C_{gg} are covariance matrices.

3 Proposed algorithm for face and facial expression recognition

According to Fig. 1, the general frame work of the proposed recognition algorithm could be divided into two main phases, training phase and testing phase.

3.1 Training phase

Details of the proposed algorithm are summarized in the following steps:

1. For every face image in the training set, multiple transform domain features are obtained employing up to three transforms, e.g., (FFT, DCT and DWT). Daubechies filters were employed in case of using DWT. Employing multiple transforms allow us to obtain more features that

improve the inter class discrimination, hence improve the overall recognition accuracy.

2. To enhance the efficiency of the proposed algorithm, a dimensionality reduction technique is needed. 2DPCA that maintains the spatial relation between adjacent pixels is employed separately for each transform.
3. To group and stack all features descriptors together each projected feature descriptor will be reshaped into one vector.
4. For any data set containing N classes and, in case of using M images for each class in the training phase, we will have $L \times N \times M$ feature vectors each of size $(X \times y)$, where L is the number of used transforms. For each class X is the number of rows in original face image and y is the number of used dominant Eigen vectors. Group all obtained feature vectors to construct discriminating feature sets by stacking the $L \times M$ feature vector of each class together. At this point, N feature set equal to the number of classes each of size $(Y \times y)$ by $(M \times L)$ representing training features sets or classes had been constructed.

3.2 Testing phase

1. As in training phase, for any testing image, L transforms are to be calculated and each transformed image is projected using the corresponding projection matrix V obtained in the training phase to yield L projected feature descriptors of size $(Y \times y)$ corresponding to the L transforms.
2. Each of the obtained L projected feature descriptors obtained in the previous step are to be reshaped into L feature vectors each of size $(Y \times y \times 1)$, and these L fea-

ture vectors are to be stacked side by side constructing the testing set of size $(Y * y * L)$.

3. Canonical Correlation (CC) is to be calculated between the testing set and each of the N training sets. N vectors for CC each of size $1 * L$, containing L values arranged in descending way are obtained.
4. Weighted sum algorithm was used to calculate the final value of CC to be used for classification where the training set that gives highest CC is recognized as the target set.

Basically CCA searches for basis vectors to be used to project two input sets variables (testing set and each of the training sets) into new space where the projection of these inputs become highly correlated. The size of the outputs sets variables definitely differs from the size of input sets variables because of the projection process. Input sets must have the same number of variables (rows) but can have different numbers of observations (columns). The canonical correlation (CC) between the columns of the output sets are calculated and listed in a vector in a descending way. The J th value of CC represent the correlation between the J th columns of the output sets after projection.

In the proposed algorithm, we have N training sets each of size $(Y * y * M * L)$ and the testing set is of size $(Y * y * L)$. N vectors for canonical correlation are obtained, each containing L values where L is the number of transforms employed in the algorithm (three in our case), Y is the number of columns of the original image, y is the number of the dominant Eigen vectors and M is the number of training images per class. Weighted sum algorithm is used for each CC vector's components converting each of N canonical correlation vectors into N canonical correlation values, the training set corresponding to highest value is recognized as the target set. Validation step was performed using sample of testing data to chose best weights for the weighted sum algorithm.

3.3 Experimental results

ORL, UMIST and Japanese Female Facial Expression (JAFPE) data sets have been used to prove the effectiveness of the proposed algorithm for both face and facial expression recognition.

3.3.1 ORL data set

For the ORL data set, it contains ten different images of 40 distinct persons with pixel resolution of $92 * 112$. Different conditions such as varying lighting slightly, facial expressions (open/closed eyes, smiling/non-smiling) and facial details (with glasses/without glasses) were considered. The first five images per each class are used for training and the remaining five images are used for testing.

Table 1 ORL data set train with five images per person and test with the other five images

Technique	Accuracy (%)
The proposed DWT-CCA	93
The proposed DCT-CCA	94.5
The proposed FFT-CCA	98
The proposed (DWT, DCT, FFT)-CCA	98.5
KLDCCA [7]	97.25
SPDCCA [8]	95
LSRC [13]	97

Table 1 shows the individual results of employing either only one transform or multiple transforms as feature descriptors and CCA as metric for classification. Transform domain was employed to overcome the effect of user misalignment in spatial domain as we use the only magnitude in transform domain and any translation in spatial domain results in only phase shift in transform domain while magnitude is not affected. Fusion process could be carried out at different levels ranging from pixel fusion, feature fusion till decision fusion. In the proposed algorithm, feature fusion from different face images and different transforms is performed using CCA while decision fusion is performed using weighted sum algorithm.

Final training sets were constructed by stacking all feature vectors representing the five training images in the three transform domains, so we get training sets containing 15 columns (five images and three transforms). Each training set which is constructed of 15 columns, and the testing set that has three columns corresponding to the three transforms, are to be projected to new space where similar pairs become highly correlated. CCA process its inputs as a whole, so it fuse all features from all five face images that had been used for each class, in addition to fusing features obtained from the three transforms. The classification step is based on the CC values to be obtained after CCA. Ten CC vectors were obtained and weighted sum algorithm was employed to convert these ten vectors to ten values to be compared together. validation step was repeated many time to find best weights and final decision is based the final value of CC.

From Table 1 it is clear that the proposed concept of using CCA for multiple features fusion and classification gives better accuracy than using each feature individually and gives the best result among all recent state of the art techniques. Increasing the number of training images to seven images per class, 100 % recognition accuracy was achieved. For further validation of the proposed algorithm, it was trained with just one image per class and it yields the highest accuracy between well-known classifiers as shown in Table 2. Following the same training/testing scenario in [17]. Image number one served as training data and remaining nine images served

Table 2 ORL data set train with one images per person and test with the other nine images

Technique	Average (%)
PCA [2]	55.8
SVD [14]	58.6
2D(PC)2A [15]	62.2
F-AFMT [16]	71.2
RES [17]	78.4
The proposed FFT-CCA	79.2

as testing data for first run. In the second run, image number two served as training data and remaining nine images served as testing data. Five runs were completed and the average accuracy of these five runs listed in Table 2 showed the superiority of the proposed algorithm.

3.3.2 UMIST data set

The UMIST data set consist of 575 images of 20 people. The images in this data set covers a wide range of poses from side to frontal views. The number of different images per individual varies from 19 to 48. For this data set, FFT, DCT or both of them were used as feature descriptors. Adding other transforms (DWT) did not improve the results for this data set so we only fused both FFT and DCT descriptors. Four experiments were conducted using either Leave One Out (LOO) scenario or training with four, five and six images per person and using the remaining images as testing set. Obtained results are listed in Table 3 where the highest accuracy were achieved compared to recent techniques. In MKDA [18] the authors need pre-processing, down sampling step to achieve the mentioned results while the proposed algorithm did not need any pre-processing steps. LOO protocol achieved 100 % recognition accuracy. The experiments were conducted using Matlab 2014a running on machine with the following specs: core i7-4700MQ CPU @ 2.4 GHz, RAM 6 GB. Average run time of 0.015 s/image was achieved which promotes the proposed algorithm to real time applications. All used face data sets used in this paper has almost the same resolution (image size) so time analysis for one data set is enough as all other

Table 4 Recognition accuracy on JAFFE data set for facial expression recognition

Technique	Accuracy (%)
The proposed DWT-CCA	94.52
The proposed DCT-CCA	93.15
The proposed DWT-FFT	95.89
The proposed (DWT, DCT, FFT)-CCA	97.26
ALBP + TSALLIS + NLDAL [20]	94.59
GP [21]	93.43
CURVELET + LBP [22]	93.69

data sets will have close time analysis to that of UMIST data set.

3.3.3 JAFFE data set

We used the JAFFE data set to verify the efficiency of the proposed algorithm for both face and facial expression recognition. The data set consists of ten persons, each of them has seven types of facial expressions: angry, disgust, fear, happy, neutral, sadness, and surprise. Each expression was repeated two–four times. In our first experiment, two images for each expression for each user were used as training set and the remaining 73 images were used as testing set. This experiment is conducted to recognize the facial expression regardless of who perform that expression (person independent), the obtained results are listed in Table 4.

It is worth to mention that we used only 65 % of the data set images for training and achieved the reported result while [21] use LOO scenario which tests one image each time and trains the algorithm on the remaining data and [22] use 80 % of the data set for training. The second experiment conducted for JAFFE data set was carried out to verify the efficiency of the proposed algorithm for face recognition where we train the proposed algorithm with one image per person and test with the remaining images. In other words we used ten images as training set and 203 image as testing set, the average accuracy achieved over ten runs was 92.9 %. Using one image of each expression for each person (70 images) as

Table 3 Recognition accuracy for UMIST data set

Technique	TRAIN 4 (%)	TRAIN 5 (%)	TRAIN 6 (%)	LOO (%)
The proposed FFT-CCA algorithm	92.3	93.3	98.0	100
The proposed DCT-CCA algorithm	90.7	95.4	98.2	100
The proposed Fused (FFT-DCT)-CCA	93.9	95.6	98.9	100
MKDA [18]	93.0	95.8	97.4	–
ASIFT [19]	–	–	–	94.8

Fig. 2 Sample of the obtained features (acceleration, angular momentum and electromyography components) for Hadouken movement

training set and the remaining images as testing sets yields 100 % recognition accuracy.

4 Sensor based applications

As listed in Tables 1, 2, 3 and 4, the proposed algorithm was verified for face recognition and achieved excellent recognition rates. In this section, the proposed algorithm is adapted for early gesture recognition using a completely different features rather than that used for face recognition. The MYO sensor developed by Thalmic Labs [23], fitted around the upper part of both arms detecting muscle movements was used to capture different gestures. Each sensor has eight EMGs sensors, x , y , z accelerometer and x , y , z angular velocity, resulting in totally 28 feature components acquired for both arms ($(8 + 3 + 3 = 14) \times 2$) transmitted to computer using Bluetooth. Sample data captured by MYO sensor is shown in Fig. 2. The proposed algorithm uses either single or multiple features' components stacked side by side to represent different gestures.

4.1 Early gesture recognition for gaming

Recently, interactive systems became very common, enabling different applications. Any interactive system depends on extracting discriminating features for training the machine so that it can make true recognition decision during testing phase early before the movement ends. A newly collected data set by us at LIMU laboratory, Kyushu University containing movements from street fighting game [24] was used to verify the efficiency of the proposed algorithm. We used the same algorithm proposed for face recognition but employing features from MYO sensor instead of facial images. The obtained data from MYO sensor was used in its raw form without any preprocessing or transformation to save computational costs and achieve fast interaction. For further real time interaction, only the first X frames from each sequence were used for recognition so that the algorithm can recognize new movement during testing phase long before it ends.

Two experiments were conducted to verify the efficiency of the proposed algorithm for early recognition. The data set was recorded at frame rate of 50 Hz employing two sensor for the two arms. The experiments were conducted using Matlab 2015b running on machine with the following specification: Intel(R) Xeon(R) CPU E5-1620 V2@ 3.70 GHz, with installed RAM 15.9 GB.

4.1.1 Experiment 1

For the first experiment, five movements (Hadouken, Jump, Punch, Shoryuken and Squat) each recorded 60 times using three different orientations of the sensor around the armband were collected. Half of the collected samples were used for training and the other half were used for testing. The number of frames per sequences ranges from 67 to 162 at average frame rate of 113 frame per sequence. We tested the proposed algorithm for fixed window size using the first 45 frames which represent less than half the average number of frames per sequence. The used 45 frames window will be picked up from five different positions along the sequence beginning with a step of two frames. For example gesture $G1$ will give five sub windows $G1_1$ (frames 11:55), $G1_2$ (frames 13:57), ...and $G1_5$ (frames 19:63). The first ten frames were ignored in our experiment as it usually contains noise and no significant data. As some movements ends at only 67 frames, we pick training windows from only five positions, this will result in overall number of training windows of 750 (five gestures $\times$ five sub windows $\times$ 30 repetition).

Different features obtained from MYO sensor were used either individually or stacked side by side to train the proposed algorithm. Since acceleration signal is measured from vibrations and gravity, it is more suitable to represent noticeable, large-scale gestures like that for gaming and listed results in Table 5 confirms this fact. Stacking both acceleration and angular momentum components improves the performance of the system as it allows more discriminating features to be fused using CCA during testing phase. Achieved accuracy is calculated by counting how many test-

Table 5 Achieved accuracy for different features for gaming experiment 1

Used features	ACC. (%)	ANG. (%)	EMG (%)	ACC. + ANG. (%)
Proposed technique	94.1	89.7	52.1	94.7
PCA + CCA	57.5	41.6	36	76.3
2DPCA + NN	84.9	82.4	21.2	84.4
2DPCA + correlation	90.9	85.5	37.7	86.8

Table 6 Run time and used Eigen vectors for gaming experiment 1

Used features	ACC.	ANG.	EMG	ACC. + ANG.
Run time (ms)	9.8	7.2	25.6	7.6
Used Eigen vectors	19	7	45	5

Table 7 Achieved accuracy for different features for gaming experiment 2

Used features	ACC. + ANG. (%)
Proposed technique	98.2
PCA + CCA	36.1
2DPCA + NN	96.3
2DPCA + correlation	96.3

ing windows truly classified over the total number of testing windows.

For each feature, only the dominant Eigen vectors representing more than 99 % of the energy in the Eigen space were chosen. Depending on the number of chosen Eigen vectors, the size of the final features' descriptor size varies and recognition time increases with increasing the descriptor's size as listed in Table 6.

4.1.2 Experiment 2

This experiment was conducted to examine the efficiency of the proposed algorithm to discriminate between movements from the same class but performed with different velocities and strength. The data set was increased to nine movements including Gaurd, Hadouken, Jump, Strong Punch, Weak Punch, Power Up, Strong Shoryken, Weak Shoryken and Squat. Each movement was recorded 37 times with the average number of frames per movement equal to 120 frames. For each movement, 20 sequences were used as training set and the remaining 17 sequences were used as testing set. The same window segmentation as experiment one was employed resulting in total number of training windows of 900. According to results obtained in experiment 1, we used the stacked acceleration and angular momentum components as features descriptor, as it gives the best performance. Obtained accuracy is listed in Table 7.

As listed in Table 7, the proposed algorithm was able to discriminate between challenging movements of the same

trajectory with different strength. The run time increased to 12.2 ms due to the increase in both number of classes and the used Eigen vectors to almost twice the number, but it is still promising for real time implementation. To verify the efficiency of the proposed algorithm, we compared our results to other known techniques (PCA and minimum distance classifiers). From Tables 5, 7 it is clear that 2DPCA outperforms the traditional PCA, that is, because it maintains the spatial and temporal alignment for features' components and consecutive frames while using PCA will disorder either features components or frames order. In addition, CCA outperforms traditional correlation or minimum distance classifier as it first transform testing set and all training sets to a new space where similar pairs become highly correlated. The achieved excellent accuracy, fast run time and low computational complexity promotes our proposed algorithm for real time implementation. Both experiments 1 and 2 were conducted using offline sequences, for online performance suitable event detection technique could be jointed with the proposed recognition algorithm.

4.2 Air writing application

Air writing for human computer interaction was chosen to test the proposed algorithm and a new large data set containing ten gestures representing digits from zero to nine was collected. The collected sequences have varying length, so a simple threshold was used to detect the active segment position. CCA can match only feature descriptor having the same number of frames so, all obtained active segments were to be resized to have equal length sequences. Different experiments and different training/testing splitting scenarios were conducted to verify the efficiency of the proposed algorithm.

4.2.1 Experiment 1

In this experiment, ten gestures representing digits from zero to nine as shown in Fig. 3 were recorded by three different performers, each gesture was repeated 20 times by the same

Fig. 3 Ten gestures air writing data set

Table 8 Achieved accuracy for air writing application, experiment 1

Used features	ACC. (%)
40 % training, 60 % testing scenario	98.8
Leave one person out scenario	90

Table 9 Run time and used Eigen vectors for experiment 1

Used features	40 % training, 60 % testing scenario	Leave one person out scenario
Used Eigen vectors	40	125
Run time (ms)	3.5	11.5

person resulting in a total number of sequences equal to 600. All used ten gestures belongs to large scale gestures which was performed by moving the whole arm, so acceleration signal is the best choice according to reported results. In this experiment two training/testing splitting scenarios were employed. In the first scenario, 40 % of the sequences served as training set and 60 % served as testing set, eight repetitions of each person were randomly picked out to train the system and remaining 12 repetitions were used for testing. Ten random runs for training sequences selection were used and the average results are listed in Table 8. In the second scenario, Leave One Person Out scenario was implemented where all the sequences of one person were used as testing set and the sequences of the other two persons serves as training set. The obtained results listed in Table 8 is the average results for the three performers. Features energy is calculated in the Eigen space, and number of Eigen vectors that contain most of the energy were used to represent the features. The number of used Eigen vectors changes according to energy distribution in each experiment resulting in changing the size

Table 10 Achieved accuracy for air writing application, experiment 2

Used features	ACC. (%)
40 % training, 60 % testing scenario	98.7
Leave one person out scenario	86.3

of final descriptor and consequently the recognition time as shown in Table 9. From Table 9, recognition time in ranges of milliseconds and excellent accuracy was achieved for both scenarios, promoting the proposed algorithm for real time implementation.

4.2.2 Experiment 2

In this experiment, a new edited data set is created from the recorded one. It contains 20 gestures representing numbers consisting of two and three digits (0–9, 10, 21, 32, 43, 54, 65, 76, 87, 98, 100) as shown in Fig. 4. The same training/testing splitting scenarios as in experiment one were conducted for the new edited data set. The obtained results listed in Table 10 prove the efficiency of the proposed algorithm.

4.2.3 Experiment 3

In this experiment we tested the proposed algorithm for the data set collected in [25] for comparison. It was collected employing a prototype developed by the authors and consist of ten digits performed by 20 performers (13 male and 7 females) each gesture was repeated ten times resulting in overall 2000 sequences. In [25], the authors employed threshold classifier, Bayes linear classifier and an improved dynamic time-warping algorithm for classification. Various features and segmentation steps for both acceleration and

Fig. 4 Two digit gestures construction

sEMG signals were needed and a score-based classifier was used to find the final decision. The experiment in [25] was conducted using extended data set containing 19 gestures. The complete data set is not available, we have only data for the ten digits. For comparison, we used published accuracy for the ten digits used in our experiment (0–9) and it was 94.95 % with standard deviation of 9.35 which is comparable to our results which is 89.5 % with standard deviation of 0.5 using 40 % training, 60 % testing scenario. Our proposed algorithm uses only the raw acceleration signal as feature descriptor.

5 Conclusion

In this paper, we proposed an efficient and global algorithm for both vision and sensor based features. The proposed algorithm achieved the highest reported accuracy for face and facial expression recognition. Employing the same algorithm, excellent results for gesture recognition were achieved using features obtained from MYO sensor. Either single raw features or multiple features were employed either in spatial domain or in transform domains. 2DPCA was employed to extract the dominate features maintaining the spatial and temporal alignment of features' components or adjacent pixels in case vision based features. After reshaping and grouping dominant features, CCA was employed to fuse different features and find a new space where similar train/test pairs become highly correlated. Obtained results verified the efficiency of the proposed algorithm for both vision and sensor based features processing for HCI. Excellent accuracies, simplicity of implementation and small run time are all factors that promotes the proposed algorithm for real time implementation.

References

- Hotelling H (1936) Relations between two sets of variates. *Biometrika* 28(3/4):321–377
- Turk MA, Pentland AP (1991) Face recognition using eigen faces. In: *IEEE Computer Society Conference on computer vision and pattern recognition*, Proceedings CVPR'91. IEEE, pp 586–591
- Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugen* 7(2):179–188
- Comon P (1994) Independent component analysis, a new concept? *Signal Process* 36(3):287–314
- Yuan Y-H, Sun Q-S (2014) Multiset canonical correlations using globality preserving projections with applications to feature extraction and recognition. *IEEE Trans Neural Netw Learn Syst* 25(6):1131–1146
- Kim T-K, Wong S-F, Cipolla R (2007) Tensor canonical correlation analysis for action classification. In: *IEEE conference on computer vision and pattern recognition*, CVPR'07. IEEE, pp 1–8
- Peng Y, Zhang D, Zhang J (2010) A new canonical correlation analysis algorithm with local discrimination. *Neural Process Lett* 31(1):1–15
- Guan N et al (2012) Sparse representation based discriminative canonical correlation analysis for face recognition. In: *11th international conference on machine learning and applications (ICMLA)*, vol 1. IEEE, pp 51–56
- Mitra S, Acharya T (2007) Gesture recognition: a survey. *IEEE Trans Syst Man Cybern Part C Appl Rev* 37(3):311–324
- Zhu C, Sheng W (2011) Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living. *IEEE Trans Syst Man Cybern A Syst Hum* 41(3):569–573
- Wang J, Chuang F (2012) An accelerometer-based digital pen with a trajectory recognition algorithm for handwritten digit and gesture recognition. *IEEE Trans Ind Electron* 59(7):2998–3007
- Hatano A, Araki K, Matsuhara M (2009) A Japanese input method for mobile terminals using surface EMG signals. In: *Proceedings of 22nd Annu. Conf. Jpn.-Soc. Artif. Intell.*, pp 5–14
- Li C-G, Guo J, Zhang H-G (2010) Local sparse representation based classification. In: *20th international conference on pattern recognition (ICPR)*, pp 649–652
- Zhang D, Chen S, Zhou Z-H (2005) A new face recognition method based on svd perturbation for single example image per person. *Appl Math Comput* 163(2):895–907
- Li J, Pan J-S (2008) A novel pose and illumination robust face recognition with a single training image per person algorithm. *Chin Opt Lett* 6(4):255–257
- Chen YM, Chiang J-H (2010) Face recognition using combined multiple feature extraction based on Fourier-Mellin approach for single example image per person. *Pattern Recognit Lett* 31(13):1833–1841
- Kuo C-H, Lee J-D (2012) Face recognition based on a two-view projective transformation using one sample per subject. *IET Comput Vis* 6(5):489–498
- Araabi BN, Gharibshah Z (2012) Face recognition with manifold-based kernel discriminant analysis. In: *The 2012 international joint conference on neural networks (IJCNN)*, pp 1–8
- Morel J-M, Yu G (2009) Asift: a new framework for fully affine invariant image comparison. *SIAM J Imaging Sci* 2(2):438–469
- Liao S, Fan W, Chung ACS, Yeung D-Y (2006) Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features. In: *IEEE international conference on image processing*, pp 665–668
- Cheng F, Jiangsheng Y, Xiong H (2010) Facial expression recognition in jaffe data set based on gaussian process classification. *IEEE Trans Neural Netw* 21(10):1685–1690
- Saha A, Wu QMJ (2010) Facial expression recognition using curvelet based local binary patterns. In: *IEEE international conference on acoustics speech and signal processing (ICASSP)*, pp 2470–2473
- <http://www.thalmic.com/myo/techspecs>. Accessed Aug 2015
- <http://www.streetfighter.com/>. Accessed Aug 2015
- Lu Z, Chen X, Li Q, Zhang X, Zhou P (2014) A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices. *IEEE Trans Hum Mach Syst* 44(2):293–299