algorithms

MDPI

Article

Facial Expression Recognition Based on Discrete Separable Shearlet Transform and Feature Selection

Yang Lu, Shigang Wang * and Wenting Zhao

College of Communication Engineering, Jilin University, Changchun 130022, China; yanglu16@mails.jlu.edu.cn (Y.L.); zhaowt0910@gmail.com (W.Z.)

* Correspondence: wangsg@jlu.edu.cn

Received: 11 December 2018; Accepted: 25 December 2018; Published: 31 December 2018

Abstract: In this paper, a novel approach to facial expression recognition based on the discrete separable shearlet transform (DSST) and normalized mutual information feature selection is proposed. The approach can be divided into five steps. First, all test and training images are preprocessed. Second, DSST is applied to the preprocessed facial expression images, and all the transformation coefficients are obtained as the original feature set. Third, an improved normalized mutual information feature selection is proposed to find the optimal feature subset of the original feature set, thus we can retain the key classification information of the original data. Fourth, the feature extraction and selection of the feature space is reduced by employing linear discriminant analysis. Finally, a support vector machine is used to recognize the expressions. In this study, experimental verification was carried out on four open facial expression databases. The results show that this method can not only improve the recognition rate of facial expressions, but also significantly reduce the computational complexity and improve the system efficiency.

Keywords: facial expression recognition; discrete separable shearlet transform; feature selection; linear discriminant analysis; support vector machine

1. Introduction

Facial expression recognition (FER) is attracting a lot of research attention because of its prospective applications in human–computer interactions and intelligent transportation systems [1–4]. Researchers have found that facial expressions can be analyzed to identify emotions, behavioral information, and psychological activities. In the field of multimedia, if an effective FER system can recognize users' facial expressions in real time, it can feedback the changes of customers' facial expressions to the multimedia system, and the multimedia system can provide different multimedia contents for the customers [5–8]. Therefore, facial expressions not only reflect the inner thoughts of human beings but are also an indispensable part of interpersonal communication [9].

FER usually involves three steps: preprocessing the image, extracting the facial expression features, and training and recognizing the expression feature model. Facial expression feature extraction is the most important part of an FER system. An effective expression feature extraction greatly improves the recognition performance. Many algorithms have been developed for this purpose. The study addressed in [10] employed biorthogonal wavelet entropy to extract multiscale features, and used a fuzzy multiclass support vector machine to be the classifier; experimental results demonstrate the effectiveness of the proposed algorithm. The study in [11] proposed a novel facial emotion recognition method based on discrete wavelet transform, principal component analysis, and cat swarm optimization. They used discrete wavelet transform to extract wavelet coefficients and principal component analysis was utilized to reduce the features. Finally, a single-hidden-layer neural network was used as the classifier. Experimental results demonstrate the feasibility of the

proposed algorithm. Gabor is widely used in the pattern recognition fields, such as image processing and feature extraction [12]. Gabor wavelet transform is an important feature extraction tool because of its multiresolution analysis. After the original image is decomposed by Gabor wavelets at different decomposition scales, we can obtain the approximated and detailed information of the target image at different levels. Unfortunately, for FER, the Gabor wavelet transform still has defects. Because the two-dimensional (2D) separable wavelet formed by the one-dimensional (1D) wavelet transform has limited directions, it cannot best represent the high-dimensional features with line or surface singularities. In fact, the features with line or surface singularities are very common in facial expression images, such as the outline of the mouth and the eyes. These are remarkable and precise features of human facial expressions. To overcome this limitation, multiscale geometric analysis tools have been developed, such as curvelet transform [13] and shearlet transform [14]. Although curvelet basis functions are advantageous for approximating linear singularities, the edge of a facial expression image is usually a curve rather than a straight-line, so curvelet transform has limitations in expression recognition. The discrete separable shearlet transform (DSST) is a multiscale geometric framework for image analysis [15]. It can better detect the edge and detail information because it has characteristics like multidirectional, multiscale, localization, and anisotropy. However, the coefficients of the image after DSST are large, so it is not desirable to use all the transformation coefficients as the expression feature set. Therefore, it is necessary to find the optimal feature subset of the original feature set so as to retain the key classification information of the original data and improve the identification efficiency of the system. In this paper, a novel approach to facial expression recognition based on DSST and normalized mutual information feature selection is proposed.

The remainder of this paper is organized as follows: Section 2 describes the theoretical methods adopted in this study, Section 3 describes the methodology, Section 4 presents the experimental results and discussions, and finally, Section 5 presents the conclusions and future work.

2. Theoretical Analysis

DSST is a multiscale geometric framework for image analysis that is designed to represent information—not only across several scales but also across several orientations—so that it can efficiently represent geometric features, like edges and other landmarks in images. Compared with shearlet transform [14], the construction of DSST is simpler [15]. It has many properties, such as multidirectional, multiscale, localization, and anisotropy. These properties enable DSST to detect edges and other elongated geometric features very effectively. These features occupy the dominant position in facial expression images.

The traditional shearlet transform (ST) was introduced in [14]. Compared with ST, an advantage of DSST is that the separable scaling function, $\phi \in L^2(R)$, and separable shearlet generators, $\psi^{(0)}$, $\psi^{(1)} \in L^2(R^2)$, can be selected. First, we define the following:

$$\phi(x_1, x_1) = \phi_1(x_1)\phi_1(x_2) \tag{1}$$

$$\psi^{(0)}(x_1, x_2) = \psi_1(x_1)\phi_1(x_2) \tag{2}$$

$$\psi^{(1)}(x_1, x_2) = \psi^{(0)}(x_2, x_1). \tag{3}$$

We construct the separable shearlet generators, $\psi \in L^2(R^2)$, and the related scaling function, $\phi \in L^2(R)$, on the horizontal cone. The vertical cone is computed similarly. We assume that $\phi \in L^2(R)$ is a compactly supported scaling function, and

$$\phi_1(x_1) = \sum_{n_1 \in Z} h^0(n_1)\sqrt{2}\phi_1(2x_1 - n_1). \tag{4}$$

Additionally, we define the compactly supported wavelet function as:

$$\psi_1(x_1) = \sum_{n_1 \in Z} g(n_1)\sqrt{2}\phi_1(2x_1 - n_1). \tag{5}$$

Thus, the shearlet generator function can be expressed as:

$$\psi(x_1, x_2) = \psi_1(x_1)\phi_1(x_2). \tag{6}$$

Moreover, the scaling function can be expressed as:

$$\phi(x_1, x_2) = \phi_1(x_1)\phi_1(x_2), \tag{7}$$

where h^0 and g are conjugate mirror filters for a fixed $J > 0$. Thus, we can write:

$$f(x) = \sum_{n \in Z^2} f_J(n) 2^J \phi\Big(2^J x_1 - n_1, 2^J x_2 - n_2\Big), \tag{8}$$

where $f_J(n) = f\big(2^{-J}n\big)$. From the above discussion, we conclude that DSST $\Big\langle f, \psi_{j,k,m} \Big\rangle$ $(j = 0, 1, \cdots, J-1)$ can be calculated as:

$$\Big\langle f, \psi_{j,k,m} \Big\rangle = w_{J-j,J-j/2}\Bigg(\Big(\Big(\widetilde{f_J}(S_k) * \Phi_k \Big) *_1 \overline{h^0_{j/2}} \Big)_{\downarrow 2^{j/2}} \Bigg)(m), \tag{9}$$

where $*_1$ is the 1D convolution, $\downarrow 2^j$ is down-sampled by 2^j along the horizontal axis, S_k is a shearing sampling matrix, and $\Phi_k(n)$ comprises the filter coefficients,

$$\Phi_k(n) = \langle \varphi(S_k(\cdot)), \phi(\cdot - n) \rangle, n \in Z^2 \tag{10}$$

$$\overline{h}_{j/2}(n_1) = h_{j/2}(-n_1) \tag{11}$$

$w_{j_1,j_2}(c), j_1, j_2 > 0, c \in l\big(Z^2\big)$ is the separable wavelet transform.

3. The Proposed Algorithm

3.1. Framework of the Algorithm

Figure 1 presents the flowchart of our proposed method. It is summarized as follows:

1. To reduce the computational complexity and satisfy the requirement of DSST for the input image size, the original image is first preprocessed.
2. After preprocessing, the test and training images are discrete separable shearlet transformed, and all DSST coefficients are extracted as the original expression feature set.
3. The improved normalized mutual information feature selection method proposed in this paper is used to find the optimal feature subset of the original feature set. The feature subset retains the key classification information of the original set.
4. After the feature extraction and selection, the feature space is reduced by employing linear discriminant analysis (LDA).
5. The support vector machine (SVM) is used to recognize the expressions.

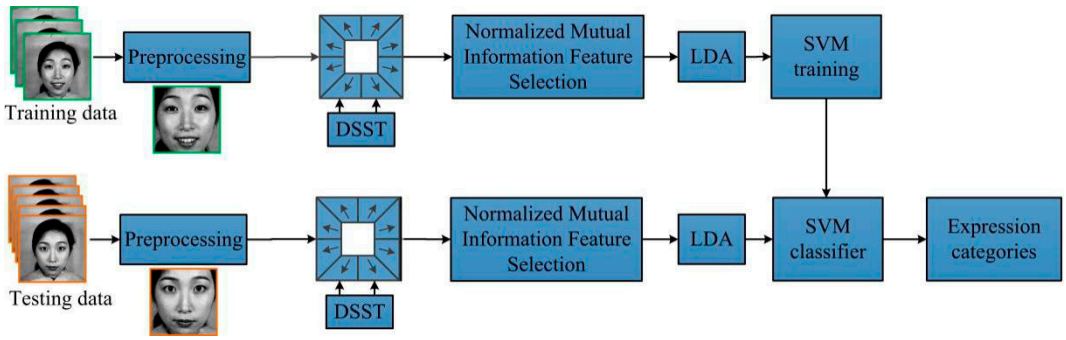


Figure 1. The framework of the proposed method.

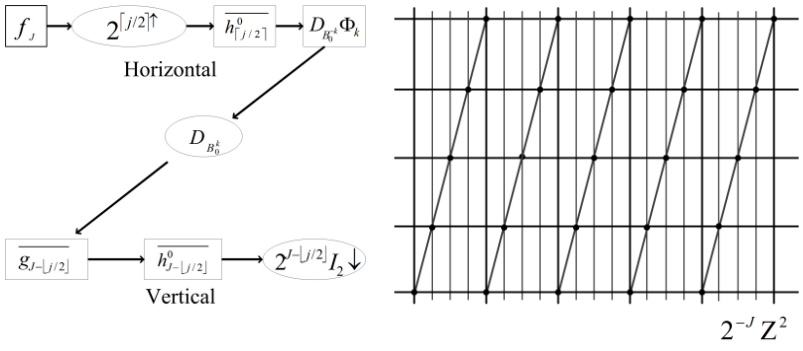
3.2. Preprocessing

The original images in a dataset usually have issues with inconsistent size and contain too much redundant information. Moreover, the expression is mainly reflected by the eyes, nose, and mouth area, whereas the surrounding area is basically useless. Therefore, it is unnecessary to extract features from the whole image, and the processing of redundant information will only increase the workload of the system. Thus, image preprocessing is necessary. Normalization and equalization were performed on the original images. We followed the preprocessing method given by [16]. The facial images were detected and all images were normalized to a gray-level image of size 64 × 64 pixels.

3.3. Discrete Separable Shearlet Transform

Given an $N \times N$ image, the DSST procedure described in Section 2 as a fixed decomposition scale, j, can be summarized as follows, and the procedure is shown in Figure 2.

1. f_J is up-sampled by $2^{j/2}$ to obtain $f_{J\uparrow 2^{j/2}}$.
2. Compute the 1D convolution of $f_{J\uparrow 2^{j/2}}$ and $h^0_{j/2}$, where $h^0_{j/2}$ is a 1D low-pass filter to obtain $\widetilde{f}_J$.
3. $\widetilde{f}_J$ is up-sampled by the shear matrix S_k to obtain $\widetilde{f}_J(S_k(n))$.
4. $\overline{h^0_{j/2}}$ is down-sampled by $2^{j/2}$ to obtain $\overline{h^0_{\lfloor j/2 \rfloor}}$. The 1D convolution of $\overline{h^0_{\lfloor j/2 \rfloor}}$ and $\widetilde{f}_J(S_k(n))$ is then computed.
5. Use the separable wavelet transform call and proceed through all the scales, $j, j = 0, 1, \cdots, J-1$.



(a) **(b)**

Figure 2. (**a**) Two computation steps for discrete separable shearlet transform (DSST) coefficients: refinement along the horizontal axis (top) and resampling associated with shear matrix (middle), followed by the separable wavelet transform along the vertical axis (bottom); (**b**) refinement along the horizontal axis to obtain the coefficients when $j = 4$ and $k = 1$.

DSST has no limitation on the number of directions. If we define the direction number as L, the feature vectors will still have $\left[\left(N/2^j\right)^2 + N^2/2^j\right] \times L$ elements, where N is the size of the input image

(e.g., $N = 64$, $j = 2$, $L = 6$). The number of feature elements is $\left(\left(64/2^2\right)^2 + 64^2/2^2\right) \times 6 = 7680$, which is still a high dimension for subsequent processing and recognition. Therefore, in this paper, we propose an improved method for normalized mutual information feature selection to find the optimal feature subset of the original feature set. Moreover, the feature subset not only retains the key classification information of the original set, but also reduces the amount of data.

3.4. Feature Selection

Battiti [17] defined the feature selection problem as the process of selecting the most relevant $k, k < M$ features subset $\mathbf{W} = \{t_1, t_2, \cdots t_w, \cdots, t_k\}$ from an initial set of $\mathbf{T} = \{t_i | i = 1, 2, \cdots, M\}$, and proposed a greedy selection method to solve it. Ideally, the problem can be solved by maximizing $\mathrm{M}\, I(\mathbf{C}; \mathbf{W})$, the joint mutual information between the class variable $\mathbf{C}$ and the subset of selected features $\mathbf{W}$. Pablo et al. [18] used an incremental search pattern to solve the problem of feature selection. They redefined a criterion function G as:

$$G = \mathrm{NM}\, I(c_j; t_i) - \frac{1}{|\mathbf{W}|} \sum_{t_s \in W} NM\, I(t_i; t_s), \tag{12}$$

where $\mathrm{NM}\, I(c_j; t_i)$ represents the normalized mutual information of c_i and t_i, $\mathrm{NM}\, I(t_i; t_w)$ represents the normalized mutual information of t_i and t_w, $\mathrm{NM}\, I(t_i; t_w)$ can be calculated as:

$$\mathrm{NM}\, I(t_i; t_w) = \frac{\mathrm{M}\, I(t_i; t_w)}{\min[H(t_i), H(t_w)]}, \tag{13}$$

where $H(t_i)$ is the entropy of information for t_i and $H(t_w)$ is the entropy of information for t_w. However, the above method has the drawback of unequal weight in normalization, since the weight $\min[H(t_i), H(t_w)]$ depends on the variables t_i and t_w. To eliminate this drawback, an improved algorithm is proposed in this paper.

Here, every feature t_i is quantized by employing the same number of levels (N) that has been decided to achieve the expected quantization error. Algorithm 1 illustrates the quantization algorithm, where t_i represents the original feature; U_{pper} and L_{ower} represent the maximum and minimum values of the original feature respectively; S_{tep} stands for quantization step; P_{artition} is the segmentation vector, which represents the level of quantization range segmentation; y_i represents the quantized value of the original feature; Q_{uantiz} is a quantization function defined in MATLAB; and C_{odebook} represents the set of quantized values.

The number of quantization levels progressively increases until the quantization error is smaller than a predefined constant ξ, which is the expected quantization error. Here, we used $\xi = 0.05$.

Clearly, $|\Omega_{\mathbf{T}}| = N$, where $\Omega_{\mathbf{T}}$ is the alphabet of the variable $\mathbf{T}$, and the entropy function $H(\mathbf{T})$ of $\mathbf{T}$ satisfies Jensen's inequality [19]:

$$H(\mathbf{T}) \leq \log_2 \left[\sum_{t_i \in \Omega_T} p(t_i) \frac{1}{p(t_i)} \right] \tag{14}$$

$$H(\mathbf{T}) \leq \log_2(|\Omega_{\mathbf{T}}|) \tag{15}$$

Here, $p(t_i)$ is the probability distribution of t_i, so, $H(\mathbf{T}) \leq \log_2(N)$. The joint mutual information of $\mathbf{T}$ and $\mathbf{Y}$ is computed as:

$$\mathrm{M}\, I(\mathbf{T}; \mathbf{Y}) \leq \min[H(\mathbf{T}), H(\mathbf{Y})]. \tag{16}$$

Hence, it is clear from (15) and (16) that:

$$\mathrm{M}\, I(\mathbf{T}; \mathbf{Y}) \leq \log_2(N), \tag{17}$$

where $\log_2(N)$ is the upper bound of the mutual information $\mathrm{M}\, I(\mathbf{T}; \mathbf{Y})$ and does not depend either on $\mathbf{T}$ or $\mathbf{Y}$.

Algorithm 1 Quantization algorithm

Input: M-Total number of features, $t_i(i = 1, \cdots, i, \cdots, M)$-Feature data, ξ-The expected quantization error
Output: N-Number of quantization levels, $y_i(i = 1, 2, \cdots, i, \cdots, M)$-Quantized data
1. **begin**
2. $N = 2$
3. **while** 1 **do**
4. $E_{\max} = -1e + 16$ **for** $i = 1$ to M **do**
5. $U_{\text{pper}} = \max(t_i)$ $L_{\text{ower}} = \min(t_i)$, $S_{\text{tep}} = (U_{\text{pper}} - L_{\text{ower}})/N$, $P_{\text{artition}} = [L_{\text{ower}} : S_{\text{tep}} : U_{\text{pper}}]$, $C_{\text{odebook}} = [U_{\text{pper}} - L_{\text{ower}}, L_{\text{ower}} : S_{\text{tep}} : U_{\text{pper}}]$, $[y_i, E_{\text{Qerror}}] = Q_{\text{uantiz}}[t_i, P_{\text{artition}}, C_{\text{odebook}}]$
6. **if** $E_{\text{Qerror}} > E_{\max}$ **then**
7. $E_{\text{Qerror}} = E_{\max}$
8. **end**
9. **end**
10. **If** $E_{\max} < \xi$ **then**
11. Break
12. **end**
13. $N = N + 1$
14. **end**
15. **end**

To eliminate the drawback of unequal normalization weights, we propose to use (17) to normalize the mutual information instead of (12) in [18].

Hence, the improved criterion function G of the feature selection problem based on normalized mutual information is:

$$G = \mathrm{NM}\, I(\mathbf{C}; t_i) - \frac{1}{|\mathbf{W}|} \sum_{t_w \in W} NM\, I(t_i; t_w) = \frac{\mathrm{M}\, I(\mathbf{C}; t_i)}{\log_2(|\Omega_{\mathbf{C}}|)} - \frac{1}{|\mathbf{W}|} \sum_{t_w \in W} \frac{\mathrm{M}\, I(t_i; t_w)}{\log_2(N)}. \tag{18}$$

The improved feature selection algorithm based on normalized mutual information can be summarized as follows:

1. Initialization: Assume $\mathbf{T} = \{t_i | i = 1, 2, \cdots, M\}$ as the original set of features; initialize $\mathbf{W}$ as an empty set.
2. Calculate the joint mutual information of each feature and class: $\mathrm{M}\, I(\mathbf{C}; t_i)$.
3. Find the first selected feature: find the feature t_i that maximizes $\mathrm{M}\, I(\mathbf{C}; t_i)$; delete t_i from set $\mathbf{T}$; and then add t_i to set $\mathbf{W}$, i.e., $\mathbf{T} = \mathbf{T} \backslash \{t_i\}$, $\mathbf{W} = \mathbf{W} \cup \{t_i\}$.
4. Repeat the following procedure until $|\mathbf{W}| = k$: (a) Computed the feature-feature mutual information $\mathrm{M}\, I(t_i; t_w)$; (b) find the next selected feature: $t_i \in \mathbf{T}$ that maximizes the criterion function G shown in (18); (c) delete t_i from set $\mathbf{T}$, then add t_i to set $\mathbf{W}$, i.e., $\mathbf{T} = \mathbf{T} \backslash \{t_i\}$, $\mathbf{W} = \mathbf{W} \cup \{t_i\}$; (d) the output set $\mathbf{W}$, which contains k selected features, is the most relevant features subset from an initial set of $\mathbf{T} = \{t_i | i = 1, 2, \cdots, M\}$.

3.5. Dimension Reduction Based on Linear Discriminant Analysis

Dimension reduction can not only reduce data dimensions, but also extract effective information and discard useless information. Some well-known methods used for dimension reduction of a feature space are the principal component analysis (PCA) [20] and the linear discriminant analysis (LDA) [21]. In this study, LDA is used for data dimension reduction.

The within-class scatter matrix $\mathbf{U}_{\mathrm{A}}$ and between-class scatter matrix $\mathbf{U}_{\mathrm{B}}$ are defined as follows:

$$\mathbf{U}_{\mathrm{A}} = \sum_{j=1}^{C} V_j \left(\overline{m}_j - \overline{\overline{m}}\right) \left(\overline{m}_j - \overline{\overline{m}}\right)^T, \tag{19}$$

$$\mathbf{U}_{\mathrm{B}} = \sum_{j=1}^{C} \sum_{m_k \in c_j} \left(m_k - \overline{m}_j\right)\left(m_k - \overline{m}_j\right)^T, \tag{20}$$

where $V_j(j = 1, 2, \cdots, C)$ is the number of vectors in the jth class c_j, C is the number of classes (here, C represents 6 facial expressions), $\overline{\overline{m}}$ is the mean of all the vectors, $\overline{m}_j$ is the mean of the class c_j, and m_k is the vector of a specific class. Therefore, the optimal discrimination projection matrix of LDA can be written as:

$$\mathbf{D}_{\mathrm{opt}} = \arg\max_{\mathbf{D}} \frac{\left|\mathbf{D}^T \mathbf{U}_B \mathbf{D}\right|}{\left|\mathbf{D}^T \mathbf{U}_A \mathbf{D}\right|} = [d_1, d_2, \cdots, d_o]. \tag{21}$$

The size of $\mathbf{D}_{\mathrm{opt}}$ is $o \times r, o \leq C - 1$, and r is the number of elements in a vector. When the within-class scatter matrix $\mathbf{U}_{\mathrm{A}}$ is nonsingular, according to Lagrange multiplier method, the column vector of the optimal projection matrix $\mathbf{D}_{\mathrm{opt}}$ satisfies the characteristic equation $\mathbf{U}_B d_i = \lambda \mathbf{U}_A d_i, i = 1, 2, \cdots, o$; put the characteristic equation into Equation (21) and we get $\mathbf{D}_{\mathrm{opt}} = \arg\max|\lambda|$. Therefore, we only need to reserve the eigenvectors that correspond to o eigenvalues with larger absolute values, and the low-dimensional space composed of o eigenvectors is the low-dimensional space we are looking for. Because the rank of each $\left(m_k - \overline{m}_j\right)$ in $\mathbf{U}_B$ is 1, the maximum rank of $\mathbf{U}_B$ is C. However, if we know the first $C - 1$ $\overline{m}_j$, the last $\overline{m}_C$ can be linearly represented by the first $C - 1$ $\overline{m}_j$. Thus, the maximum rank of $\mathbf{U}_B$ is $C - 1$, so there's at most $C - 1$ eigenvectors, thus $o \leq C - 1$. Thus, LDA maximizes the total scattering of the data while minimizing the within scattering of the classes.

3.6. Facial Expression Recognition and Classification Using Support Vector Machine

In machine learning, SVM [22] uses a kernel function to map the data in an input space to a high-dimensional feature space in which we can process a problem in linear form. The steps of facial expression recognition based on SVM are as follows:

1. Selection of the kernel function: in this method, we choose radial basis function (RBF) as the kernel function, as shown in (22). The reasons for choosing RBF are that RBF can realize nonlinear mapping and solve nonlinear separable problems, and, compared with other kernel functions, RBF has only one parameter σ, so its model complexity is lower than that of others.

$$K(x, x_{\mathrm{c}}) = \exp\left(-\frac{\|x - x_{\mathrm{c}}\|^2}{\sigma^2}\right) \tag{22}$$

2. Selection of RBF parameter σ and penalty coefficient E: when using the RBF kernel function, the values of two parameters, σ and E, are considered. However, there is no theoretical guidance on how to select the two parameters' values, therefore, we use Dr. Lin's tool grid.py in LibSVM to select the optimal values of σ and E, $E = 128$ and $\sigma = 0.0078125$.
3. Construction of the multiclass SVM classifier: we adopt the one-against-one voting strategy of SVM. In the training stage, we use 6 categories of samples to construct $6 \times (6 - 1)/2 = 15$ SVM binary classifiers. We save the results of each SVM binary classifier into an array of structural cells, and hence save all the information needed for multiclass SVM classification into the array of structural cells. In the multiclass SVM classification stage, the training samples are successively passed through the 15 SVM binary classifiers, and the category of data is determined by one-against-one voting strategy.

4. Results and Discussion

To verify the effectiveness of the proposed algorithm, we performed experiments on the Japanese Female Facial Expression (JAFFE) [23], extended Cohn-Kanade (CK+) [24], MMI Facial Expression Database (MMI) [25], and Psychological Image Collection at Stirling (PICS) [26] facial expression datasets. These are the four most comprehensive datasets currently available for facial expression

research. We applied a 10-fold cross-validation scheme, i.e., out of 10 subjects, data from a single subject were reserved as the validation data for testing the algorithm proposed in this paper, whereas the data for the remaining nine subjects were used as the training data.

4.1. Experimental Database

JAFFE dataset: the JAFFE dataset contains 213 images of seven facial expressions (six basic facial expressions + one neutral) posed by 10 Japanese female models. Each image was rated on six emotion adjectives by 60 Japanese subjects. The database was planned and assembled by Michael Lyons, Miyuki Kamachi, and Jiro Gyoba. The photos were taken at the psychology department in Kyushu University. The images in the JAFFE dataset are all positive faces, and the original images are adjusted and pruned to make the position of eyes and size of the faces roughly the same. The illumination is all positive light sources but the intensity of illumination is different. Because the expression database is open access and the expression calibration is very standard, it is used for training and testing in most articles on expression recognition nowadays. All 213 images of the JAFFE database were used for six-class expression recognition in this study. For each subject, we randomly chose 14 images for training and used the rest for testing. Each time, data from a single subject were reserved as the validation data for testing the algorithm proposed in this paper, whereas the data for the remaining nine subjects were used as the training data.

CK+ dataset: the Cohn-Kanade AU-Coded Facial Expression Database is for research in automatic facial image analysis and synthesis and for perceptual studies. Cohn-Kanade is available in two versions and a third is in preparation, and CK+ dataset is the second version. It includes both posed and non-posed (spontaneous) expressions and additional types of metadata, and it consists of 123 university students aged 18–30 years, of which 65% are female, 15% are African-American, and 3% are Asian or Latino. This dataset is much larger than JAFFE and is available free of charge. The database contains 593 image sequences, of which 327 have emotion labels. This database is a popular database in facial expression recognition. Many articles use this database for training and testing. We selected 320 image sequences from the CK+ dataset. The only selection criterion was that a sequence could be labeled as one of the six basic emotions. The sequences come from 96 subjects, with one to six emotions per subject. Because we studied facial expression recognition based on static images in this paper, for each sequence, the neutral face and three peak expression frames were used for prototypic expression recognition, resulting in 960 images (i.e., 108 anger, 120 disgust, 99 fear, 282 joy, 126 sadness, and 225 surprise). The peak expression frame was chosen because it reflects the best state of expression. Therefore, we can extract the features that best reflect this expression. More precisely, we distributed the images randomly into 10 groups with roughly equal numbers of subjects. Nine groups were used as the training data to train classifiers, and the remaining group was used as the test data.

MMI dataset: the MMI facial expression database is an ongoing project that aims to deliver large volumes of visual data of facial expressions to the facial expression analysis community. A major issue hindering new developments in the area of automatic human behavior analysis in general, and affect recognition in particular, is the lack of databases with displays of behavior and affect. To address this problem, the MMI facial expression dataset was conceived in 2002 as a resource for building and evaluating facial expression recognition algorithms. The database addresses a number of key omissions in other databases of facial expressions. In particular, it contains recordings of the full temporal pattern of facial expressions ranging from neutral, through a series of onset, apex, and offset phases, and back again to a neutral face. The database consists of over 2900 videos and high-resolution still images of 75 subjects. It is fully annotated for the presence of action units (AUs) in videos and partially coded on frame-level, indicating for each frame whether an AU is in either the neutral, onset, apex, or offset phase. A small part was annotated for audio-visual laughter. The database is freely available to the scientific community. Then, 96 image sequences were selected from the MMI database; the sequences were obtained from 20 subjects with one to six emotions per subject. The neutral face and

three peak frames of each sequence (384 images in total) were used for six-class expression recognition. We randomly chose all images from 15 people for training and used the rest for testing.

PICS dataset: the PICS was planned and assembled by Psychology, School of Natural Sciences University of Stirling. It has many subset databases, such as the Stirling/ESCR 3D face dataset, 2D face sets, 3D face sets, and the other image sets; this paper chose the pain expression subset. It contains 599 images of 13 females and 10 males, including the same seven expression categories as in JAFFE. The resolution of each image is 181 × 241. The database is also freely available to the scientific community. From the PICS dataset, we chose 180 images of 10 people, including six expression categories and three images per person. In this study, we randomly chose nine subjects for training and used the rest for testing.

4.2. Recognition Rates of the Proposed Method

The recognition rates of the proposed method were evaluated for each dataset separately under the settings, as mentioned above. The 3D feature plots of the proposed method for the six expressions after applying LDA on four datasets are shown in Figure 3a–d, and the recognition rates are shown in Table 1.

In the legend in Figure 3, H represents happiness, Su represents surprise, Sa represents sadness, D represents disgust, A represents anger, and F represents fear. It is observed that the algorithm clearly classifies the features of six kinds of expressions, which provides a powerful tool for subsequent classification and recognition. It is clear from Table 1 that the proposed FER system achieved a high recognition rate on the four datasets. Moreover, the recognition rate decreased significantly without the feature selection method. The experimental results show that the improved feature selection method proposed in this paper plays an important role in the high recognition of the FER system.

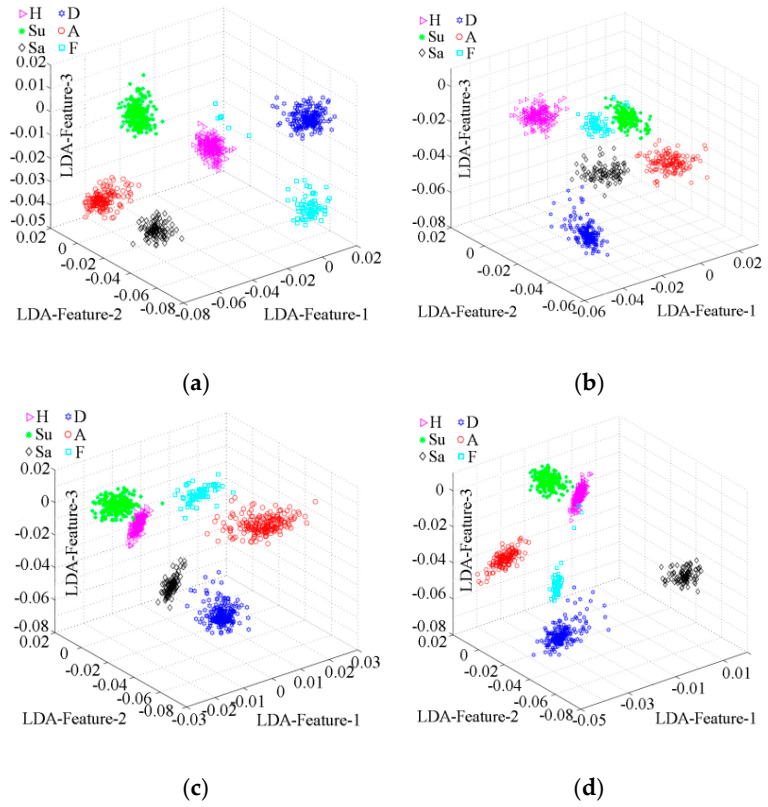


Figure 3. 3D feature plots of the proposed method for recognizing the expressions on the four datasets: (**a**) Japanese Female Facial Expression (JAFFE) dataset, (**b**) extended Cohn-Kanade (CK+) dataset, (**c**) MMI Facial Expression (MMI) dataset, and (**d**) Psychological Image Collection at Stirling (PICS) dataset.

Table 1. Comparison of average accuracy rate of the proposed method with and without feature selection on the four data sets.

Method	Data Set	Average Accuracy Rate (%)
With feature selection	JAFFE	98.00
	CK+	95.17
	MMI	96.02
	PICS	97.33
Without feature selection	JAFFE	94.06
	CK+	91.50
	MMI	92.23
	PICS	92.50

To further verify that this algorithm is independent of the dataset and has strong robustness, we employed 4-fold cross-validation on the dataset. Out of the four datasets, one was utilized as the training data and the others were used as the testing data. This process was repeated four times. The weighted average recognition rates of the proposed method are shown in Table 2.

Table 2. Average accuracy rate of the proposed method trained on one dataset and tested on the others.

Training Dataset.	Testing Datasets	Average Accuracy Rate (%)
JAFFE	CK+, MMI, PICS	87.89
CK+	JAFFE, MMI, PICS	84.97
MMI	JAFFE, CK+, PICS	85.62
PICS	JAFFE, CK+, MMI	86.38

As seen from Table 2, the proposed algorithm not only has a high recognition rate for a single database, but also has a high recognition rate when training one database and testing the other three databases. This proves that the algorithm has strong robustness.

4.3. Contrast Experiment

Table 3 compares the proposed FER method with state-of-the-art methods, which were selected because they use frequency domain features, a similar testing strategy, and the same dataset. For a fair comparison, we implemented some of these methods, and for those which we did not, we quoted their published results. A 10-fold cross-validation scheme was used on each dataset. For the four datasets, the weighted average recognition rates of the state-of-the-art methods and the proposed method are shown in Table 3. Table 3 clearly indicates that the proposed method outperforms the state-of-the-art methods for the four datasets. The weighted average recognition rate of the proposed method was 7.62%, 1.46%, 1.98%, 4.27%, 10.71%, and 6.25% higher than the rates of the methods proposed in [27–32], respectively.

Table 3. Comparison of the proposed method with state-of-the-art methods in terms of average accuracy rate.

Method	Average Accuracy Rate/%
Ref. [27] method	89.01
Ref. [28] method	95.17
Ref. [29] method	94.65
Ref. [30] method	92.36
Ref. [31] method	85.92
Ref. [32] method	90.38
Proposed method	96.63

Moreover, the quantization algorithm has a complexity of $O(M)$, and M is the number of features. The experiments were performed in MATLAB R2014a, [Intel-(R) Core-(TM) (3.60 GHz) with a RAM capacity of 56 GB]. The proposed framework has a complexity of $O(TM)$, and T is the number of input expression images.

In order to analyze the computational cost of the proposed framework, we selected the most efficient method (that is, [27] from the above experiments of Table 3). The framework of [27] took 1493 ms, 1998 ms, 1215 ms, and 2031 ms to recognize an expression frame from JAFFE, CK+, MMI, and PICS datasets of facial expressions, respectively. On the other hand, our framework took 1185 ms, 1826 ms, 1002 ms, and 1725 ms to recognize an expression frame from the same datasets. Thus, our framework not only achieved high recognition rate, but it is also less expensive in terms of computational.

5. Conclusions

At present, facial expression recognition algorithms are mainly composed of three modules: image preprocessing, feature extraction, and recognition. However, feature selection is also very important for facial expression recognition and is a topic worthy of further study. In this paper, a novel approach to facial expression recognition based on DSST and normalized mutual information feature selection is proposed. This method was tested on four different datasets. After preprocessing, the test and training images were discrete separable shearlet transformed, and all DSST coefficients were extracted as the original expression feature set. The improved feature selection method proposed in this paper was used to find the optimal feature subset of the original feature set. Moreover, the feature subset retained the key classification information of the original set. For dimension reduction, we used LDA. SVM was used as the recognizer. The experimental results show that the weighted average recognition rate of the proposed algorithm was 96.63%, which is significantly higher than the recognition rate of existing facial expression recognition systems. Unfortunately, our framework is not yet ready for use in real-time scenarios because there exist several factors in a real-time scenario that might decrease the performance of the framework, such as complicated background, image rotation, and blur. Therefore, further study is needed to tackle these issues and maintain the same high recognition rate in a real time scenario.

Author Contributions: The experimental measurements and data collection were carried out by Y.L. and S.W. The manuscript was written by Y.L. with the assistance of S.W. and W.Z. All authors reviewed the manuscript.

Funding: This research was funded by National Natural Science Foundation of China: 61631009; National Key Research and Development Plan of 13th Five-year Plan: 2017YFB0404800; Fundamental Research Funds for the Central Universities: 2017TD-19.

Acknowledgments: Thanks to my family for their support. Thanks to my teachers and classmates for giving me guidance on my studies. Thanks to my school Jilin University.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, W.; Zhang, Y.; Ma, L.; Guan, J. Multimodal learning for facial expression recognition. *Pattern Recognit.* **2015**, *48*, 3192–3202. [CrossRef]
2. Lekdioui, K.; Messoussi, R.; Ruichek, Y. Facial decomposition for expression recognition using texture/shape descriptors and SVM classifier. *Signal Process. Image Commun.* **2017**, *58*, 300–312. [CrossRef]
3. Allaert, B.; Mennesson, J.; Bilasco, I.M. Impact of the face registration techniques on facial expressions recognition. *Signal Process. Image Commun.* **2018**, *61*, 44–53. [CrossRef]
4. Liu, Y. Facial Expression Recognition with Fusion Features Extracted from Salient Facial Areas. *Sensors* **2017**, *17*, 712. [CrossRef] [PubMed]
5. Stai, E.; Kafetzoglou, S.; Tsiropoulou, E.E.; Papavassiliou, S. A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content. *Multimed. Tools Appl.* **2018**, *77*, 283–326.

6. Pouli, V.; Kafetzoglou, S.; Tsiropoulou, E.E.; Dimitriou, A.; Papavassiliou, S. Personalized multimedia content retrieval through relevance feedback techniques for enhanced user experience. In Proceedings of the IEEE 2015 13th International Conference on Telecommunications (ConTEL), London, UK, 13–15 July 2015; pp. 1–8.
7. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. Recommender systems survey. *Knowl.-Based Syst.* **2013**, *46*, 109–132. [CrossRef]
8. Balabanovic, M.; Shoham, Y. Fab: Content-based, collaborative recommendation. *Commun. ACM* **1997**, *40*, 66–72. [CrossRef]
9. Mohammadi, M.R.; Fatemizadeh, E.; Mahoor, M.H. PCA-based dictionary building for accurate facial expression recognition via sparse representation. *J. Vis. Commun. Image Represent.* **2014**, *25*, 1082–1092. [CrossRef]
10. Zhang, Y.D.; Yang, Z.J.; Lu, H.M. Facial Emotion Recognition Based on Biorthogonal Wavelet Entropy, Fuzzy Support Vector Machine, and Stratified Cross Validation. *IEEE Access.* **2017**, *4*, 8375–8385. [CrossRef]
11. Wang, S.H.; Yang, W.; Dong, Z. Facial Emotion Recognition via Discrete Wavelet Transform, Principal Component Analysis, and Cat Swarm Optimization. In Proceedings of the 7th International Conference on Intelligence Science and Big Data Engineering (IScIDE 2017), Dalian, China, 22–23 September 2017; Springer: Berlin, Germany, 2017; pp. 203–214.
12. Selesnick, I.W. Wavelets, a modern tool for signal processing. *Phys. Today* **2007**, *60*, 78–79. [CrossRef]
13. Tang, M.; Chen, F. Facial expression recognition and its application based on curvelet transform and PSO-SVM. *Optik-Int. J. Light Electron Opt.* **2013**, *123*, 5401–5406. [CrossRef]
14. Hou, B.; Zhang, X.; Bu, X. SAR Image Despeckling based on Nonsubsampled Shearlet Transform. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 809–823. [CrossRef]
15. Lim, W.Q. The discrete shearlet transform: A new directional transform and compactly supported shearlet frames. *IEEE Trans. Image Process.* **2010**, *19*, 1166–1180.
16. Sun, W.Y. *Facial Expression Recognition Arithmetic Research*; Beijing Jiaotong University: Beijing, China, 2006; pp. 37–38. (In Chinese)
17. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550. [CrossRef] [PubMed]
18. Pablo, A.E.; Tesmer, M.; Perez, C.A. Normalized Mutual Information Feature Selection. *IEEE Trans. Neural Netw.* **2009**, *20*, 189–201.
19. Wu, S.H. Generalization and application of Jensen inequality reinforcement. *J. Sichuan Univ. (Nat. Sci. Ed.)* **2005**, *3*, 437–443.
20. Hong, Y.Y.; Wu, C.P. Day-Ahead Electricity Price Forecasting Using a Hybrid Principal Component Analysis Network. *Energies* **2012**, *5*, 4711–4725. [CrossRef]
21. Wei, Y.; Yue, Y. Research on Fault Diagnosis of a Marine Fuel System Based on the SaDE-ELM Algorithm. *Algorithms* **2018**, *11*, 82. [CrossRef]
22. Du, J.L.; Liu, Y.Y.; Yu, Y.N. A Prediction of Precipitation Data Based on Support Vector Machine and Particle Swarm Optimization (PSO-SVM) Algorithms. *Algorithms* **2017**, *10*, 57. [CrossRef]
23. Lyons, M.; Akamatsu, S.; Kamachi, M. Coding Facial Expressions with Gabor Wavelets. In Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.
24. Lucey, P. The extended Cohn-Kanade dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. In Proceedings of the IEEE 3rd International Workshop on CVPR for Human Communicative Behavior Analysis, San Francisco, CA, USA, 18 June 2010; pp. 94–101.
25. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Amsterdam, The Netherlands, 6 July 2005; p. 5.
26. PICS Database. Available online: http://pics.psych.stir.ac.uk (accessed on 11 November 2018).
27. Lu, Y.; Wang, S.G.; Zhao, W.T.; Zhao, Y. A novel approach of facial expression recognition based on shearlet transform. In Proceedings of the IEEE 5th Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, 14–16 November 2017; pp. 398–402.
28. Ghimire, D.; Lee, J. Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. *Sensors* **2013**, *13*, 7714–7734. [CrossRef]

29. Uçar, A.; Demir, Y.; Güzeliş, C. A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering. *Neural Comput. Appl.* **2016**, *27*, 131–142. [CrossRef]
30. Rivera, A.R.; Castillo, J.R.; Chae, O. Local directional number pattern for face analysis: Face and expression recognition. *IEEE Trans. Image Process.* **2013**, *22*, 1740–1752. [CrossRef] [PubMed]
31. Lu, G.M.; Li, X.N.; Li, H.B. Research on Recognition for Facial Expression of Pain in Neonates. *Acta Opt. Sin.* **2008**, *11*, 664–667.
32. Li, Y.Q.; Li, Y.J.; Li, H.B. Fusion of Global and Local Various Feature for Facial Expression Recognition. *Acta Opt. Sin.* **2014**, *34*, 172–178.