EAI Endorsed Transactions
on e-Learning

Research Article EAI.EU

Lip language identification via Wavelet entropy and K-nearest neighbor algorithm

Ran Wang[1], Yifan Cui[1], Xinyu Gao[1], Wei Chen[1], Mingbo Hu[1], Qian Li[1], Jiahui Wei[1], XianWei Jiang[1,*]

[1]School of Mathematics and Information Science, Nanjing Normal University of Special Education, Nanjing 210038, China

Abstract

INTRODUCTION: Image processing technology is widely used in lip recognition, which can automatically detect and analyse the unstable shape of human lips.
OBJECTIVES: In this paper, we propose a new algorithm using Wavelet entropy (WE) and K-nearest neighbor (KNN) improves the accuracy of lip recognition.
METHODS: At present, the two most commonly used technologies are wavelet transform and *K*-nearest neighbor algorithm. Wavelet transform is a set of image descriptors, and the *K*-nearest neighbor algorithm has high accuracy. After a large number of experiments, we propose a lip recognition method based on Wavelet entropy and *K*-nearest neighbor, which combines Wavelet entropy, *K*-nearest neighbor and K-fold cross validation.
RESULTS: This method reduces the calculation time and improves the training speed. The best result of the experiment improves the accuracy to 80.08%.
CONCLUSION: Therefore, our algorithm is superior to other state-of-the-art approaches of lip recognition.

Keywords: Lip language identification, Wavelet entropy, *K*-nearest neighbor, Wavelet transform, K-fold cross validation

Received on 29 June 2021, accepted on 05 August 2021, published on 11 August 2021

doi: 10.4108/eai.11-8-2021.170669

*Corresponding author. Email: jxw@njts.edu.cn

1. Introduction

1.1. What is Lip language identification

Lip speech recognition is a technology that combines machine vision and natural language processing to identify speech content directly from the image of someone speaking. Lip recognition system using machine vision technology, continuous identify faces from the image, determine which is the speaker, to extract the person mouth change characteristics of continuous, then enter the characteristics of continuous variation to the lip recognition model, identify the corresponding pronunciation speech population type, then according to identify the pronunciation, calculated that the most likely of natural language statements. In the process of lip recognition, the relationship between mouth shape and pronunciation, pronunciation and text, is not the only corresponding, there are often multiple possible alternative results, need to calculate the most possible result in real time.

1.2. Literatures

In recent years, image processing techniques have been extensively developed for human lip recognition, which can automatically detect and analyse the unstable shape of human lips and distinguish in real time whether the user is speaking or not. Examples include audiovisual speech

recognition (AVSR) [1], visual speech recognition (VSR) [2, 3], speaker recognition [4-6], intelligent human-computer interaction (IHCI) [7], vision-based voice activity detection (VVAD), etc. Research in the field of speech technology has achieved remarkable results both at home and abroad. In 1984, the University of Illinois developed a lip recognition system, which improved the accuracy of speech recognition by using visual information of functional features of the lips as an aid to speech recognition [8]; Kass used the Snake model to fit the contours of human lips [9-11]; in 2012, Liu proposed a lip tracking method for speaker recognition that considers the darkest part of the grayscale image as the corner of the mouth and the relatively large scatter between adjacent pixels as the contour line [5]; in 2015, Goldschen used principal component analysis [12, 13] and the coefficient matrix method to extract lip image features, and then used a Hidden Markov model for lip recognition [14]. Later, Chiou et al. combined these two methods for feature extraction of input data and used Hidden Markov Models to identify isolated words [15]. It is worth mentioning that in recent years, more and more online technology giants are studying lip recognition technology. In December 2017, Sogou held a media communication meeting in Beijing to publicly introduce this new lip recognition technology, at which Chen Wei, technical director of Sogou's voice interaction center, stated that with Sogou's strong technical background in natural language processing, Sogou is in the leading position in this field, and Sogou uses Mature convolutional neural network technology to model Chinese lip recognition sequences, using a large amount of real lip data to train the network model and build a vocabulary of more than 100,000 words. Sogou tested the trained model, and the network model had an accuracy rate of over 60%, higher than Google's 50%. Meanwhile, Sogou applied the technology to some vertical scenes and the recognition accuracy exceeded 90%.

1.3. Shortcomings of state-of-the-art methods

In the process of human communication, the face of the speaker contains extremely rich facial expressions and a variety of lip movement information. It is so much facial information superimposed together that makes it difficult to interpret the lip information. Because of these reasons, many recognition methods cannot well cover the details of the lip information. For example: computer-assisted lip-reading system (CALRS) proposed in 2008, it uses object-oriented image processing method to recognize the correct lip and neural network. The system uses SOMNN (self-organizing map neural network) to accurately compare the lip images of Putonghua speech and process the lip features to help the hearing-impaired to pronounce correctly. However, from the experiment results, the performance of the recognition system is only good enough for 10 phonemes, not for all 37 Mandarin phonemes [16]. Using this method can identify too few phonemes to meet the needs of our daily conversation and life, and the recognition rate is not high.

Until 2012, a calculation method that can obtain the characteristics of moving lips was proposed. It was tested by Yun Long lay et al. Using a method called principal component analysis and mouth change rate. It is a lip recognition system that uses image processing technology, neural network algorithm and database to help people or computer understand lip [17]. It mainly detects and analyses the dynamic changes of lip shape. However, because it uses neural network to gather similar image features first, and then recombines them, the recombining rate of its real-time system largely depends on the lip shape. So, from the experimental results, the researchers discovered that the major problem in automatic lip reading is to recognize the lip shape sets when possessing the word with the similar pronunciation voice. The same algorithm as this method is HMM (Hidden Markov Model). In application, HMM is mainly used for statistics. It infers the results by using the number of hidden states and the probability of each case, combined with their predictable state chain. It can greatly improve the efficiency of calculation results, but it will also increase the probability of calculation uncertainty. Because it depends on a variety of States and its corresponding observation objects, once one party makes a mistake, the result will deviate greatly from the correct answer.

Another example is GMM (Gaussian mixed model) algorithm, which combines Gaussian function with linear function. GMM algorithm not only has fast calculation speed, but also has high precision. Theoretically, it can simulate almost any type of distribution. But its shortcomings are also obvious. It needs more data than huge data, and more possibilities of results should be considered. Therefore, a large amount of data is needed to simulate the situation in order to improve the accuracy of its calculation. Once the amount of data provided is not enough or the data situation is not enough, then the variance of its calculation will increase, which cannot guarantee the accuracy of the calculation. Compared with traditional classification methods, support vector machine has better classification effect and significant advantages of elegant mathematical processability [18]. It performs well in the second class classification algorithm, but we often need to solve the multi classification problem in the calculation, and SVM does not perform well in solving the multi classification problem. In the face of large-scale samples, especially when it comes to the calculation of n-order matrix, it will consume too much system memory and running time, and the calculation efficiency will be very slow.

1.4. Our contribution

The Wavelet entropy (WE) algorithm not only effectively reduces the number of features, but also improves the performance of the recognizer [19]. In addition, Wavelet entropy values of different types are significantly different,

which proves that Wavelet entropy is very effective for the change of lip shape. The reason why we can get good performance may be that we can analyze the transient characteristics of non-stationary variation of lip shape [20]. In pattern recognition, the K-nearest neighbor (KNN) is a classification method based on the nearest training sample in feature space [21]. The KNN is a basic and simple method in pattern recognition. When the data distribution information is insufficient, it is one of the preferred classification methods [22].

Our lip recognition technology (WE+KNN) product the advantages of both methods. "WE + KNN" is superior to the most advanced lip recognition method in recognition accuracy. In addition, the technology can not only reduce the training time, but also meet the actual needs of online recognition calculation time. This shows the effectiveness of "WE + KNN" [20].

2. Dataset

We took 26 categories of lip images as a data set, that is, **Figure 1**, from these 26 categories of continuous analysis validation. Pinyin is composed of initials and finals. The different pronunciation of initials is determined by the different parts and methods of pronunciation. The area where airflow is blocked during pronunciation is called the articulation area. Initial pronunciation refers to the way and condition in which the larynx, mouth and nose control airflow during pronunciation. It can be observed from three aspects: the blocking mode, whether the vocal cords are vibrating or not, and the intensity of airflow. Therefore, our data analysis is based on these parts for preliminary analysis [23]. There are 39 vowels in Chinese Pinyin. Vowels consist mainly of vowels or vowels with nasal consonants. A vowel in which the position of the tongue, the shape of the lips, and the opening of the mouth remain unchanged are called monosyllabic. The difference of unit sounds is mainly caused by different mouth and tongue positions. The lifting and stretching of the tongue, the flattening and turning of the lips, and the opening and closing of the mouth all produce different forms of resonance, resulting in vowels of different timbre. Polyphonic phonemes are composed of polyphonic sounds, which are the vowels in which the position of the tongue and the shape of the lips change during the vowel pronunciation. We will conduct follow-up analysis based on these characteristics.

Figure 1. Lip shape in pinyin

3. Methodology

3.1. Wavelet entropy

Wavelet transform is a set of image descriptors that can analyze images at any resolution. At present, wavelet transform has become the choice method for many images analysis and classification problems. Nevertheless, the main disadvantage of wavelet transform is that it requires large storage space and expensive computation. Therefore, WE propose a new feature, WE, which aims to extract the entropy from the coefficients.

In statistics, entropy is defined as a random system to measure its randomness quantitatively. Entropy also represents our uncertainty about the source of information. Thus, entropy can be used to characterize the texture of the input image. The entropy of the image can be approximated from the histogram of the image. The histogram shows the different grayscale probabilities in the image [24].

The state of entropy refers to the degree of chaos in a system and is used to indicate how evenly any one kind of energy is distributed in space. The more evenly distributed the energy, the greater the state of entropy. When the energy of a system is completely evenly distributed, the state of the system is at its maximum. Wavelet state is a measure of the disorder degree of signal energy distribution in subspace. In fact, a very ordered signal is a narrowband signal (such as a periodic signal of a single frequency), with a relative wavelet energy of 1, while the relative wavelet energy of other frequency bands is 0, so the ordered signal has a wavelet equivalent of O or very close to 0. For a signal with a very uniform energy distribution, the relative wavelet energies of all frequency bands are approximately equal, so the wavelet state will be a relatively large value.

As for the concept of stew, there are many ways to define it, depending on the use of different situations to choose a different variety of children, among which the most commonly used is the Shannon child. Shannon's state theory points out that for an uncertain system, if a random variable X with a finite number of values is used to represent its state characteristics, the probability of its value being X is

$$Q_i = Q\{X = x_i\}, I = 1, 2, \ldots, n \ \sum_{i=1}^{n} Q_i = 1$$

Then the information entropy of X is defined as:

$$\mathrm{H(X)} = \sum_{i=1}^{n} Q_i \ln (1/Q_i) \quad (1)$$

Information desks can be used to quantitatively estimate the complexity of random signals. Spectral differences based on the concept of Shannon's differences are information differences that can be used to analyze the complexity of signals. The narrower the spectrum peak in the power spectrum, the smaller the spectrum blue is, which indicates that there is obvious oscillation rhythm in the signal and the complexity is small. On the contrary, the flatter the power spectrum, the larger the spectrum. However, the power spectrum estimation based on Fourier transform is only applicable to stationary signals, and the frequency resolution of the spectrum estimation is proportional to the signal length used. A short time window will cause serious sidelobe leakage effect and make the power spectrum distortion. Wavelet transform can locate and analyze non-stationary time-varying signals simultaneously in time domain and frequency domain.

(i) By using binary discrete wavelet transform, the signal can be decomposed into various components at different scales, and the wavelet coefficient $C_j\ (k)$ is obtained. The energy at different scales can be directly estimated by using these wavelet coefficients. The implementation process of binary discrete wavelet transform is equivalent to using a set of high-pass and low-pass filters repeatedly to decompose the time series signals step by step. After each decomposition, the sampling frequency of the signal is reduced by one time, and then the decomposition process of the low-frequency component is repeated, so as to obtain the next level of two decomposition components. Let the high frequency component coefficient of the signal at time k be $cD_{j,k}$ under the JTH decomposition scale after the above transformation. The low-frequency component coefficients are $cA_{j,k}$ and x. After single reconstruction, the signal components D_j and A_j are obtained.

The detail signal energy of j with different resolutions is $M_j = \sum_k |C_j(\mathrm{K})|^2$. Where the wavelet coefficient $C_j(k) =< x(t), Q_{j,k}(t) > C_j(k)$ and the total energy of the signal is zero.

$$\mathrm{M} = ||x(t)||^2 = \sum_j^n \sum_k^n |C_j(k)|^2 = \sum_j M_j \qquad (2)$$

(ii) Therefore, the normalized wavelet $P_j = \frac{E_j}{E}$. Clearly $\sum_j P_j = 1$. Similar to entropy Wavelet entropy defined:

$$\mathrm{WEE} = -\sum_j^n Q_j \ln(Q_j) \qquad (3)$$

Therefore, Wavelet Inversion *WEE* can reflect the chaotic degree of multi-frequency component signals and provide the dynamic characteristics of the signals.

WE is a new method to analyze the transient characteristics of complex signals. It is used to calculate the entropy of the probability density function PDF of the variable of the energy distribution of the wavelet coefficients in the wavelet domain. The WE value has the physical meaning of order/disorder degree of multi-scale time-frequency resolution signal. It combines wavelet transform and tangency entropy to estimate the disorder and order degree of a specific image with a specific spatial frequency resolution [25, 26]. Therefore, Wavelet entropy is of great significance for lip recognition.

WE can be used to represent the mean value of information and uncertainty in lip feature signals, and to represent the useful information in the dynamic process of signals. After the multi-scale wavelet transform of the signal, the wavelet coefficients of each scale are transformed into a probability distribution sequence. The entropy calculated by the sequence can reflect the difference and degree of different lip shapes. Therefore, we analyze different types of lip shapes and decompose the lip image by Wavelet entropy, as shown in **Figure 2**. Combined with the change characteristics of lip shape and tongue position, we find out the differences between different lip shapes to obtain more accurate values.

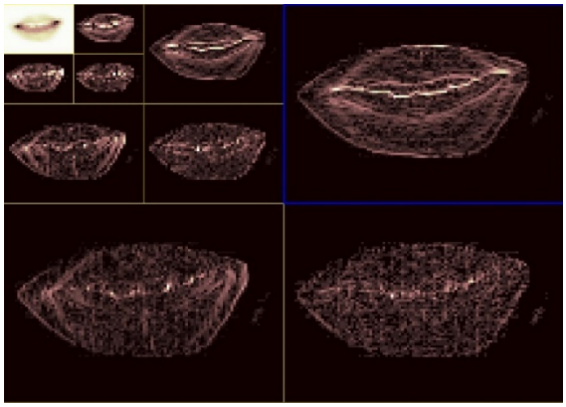

Figure 2. Wavelet entropy decomposition of lip

3.2. KNN

K-nearest neighbor (KNN) Classification Algorithm is a data mining degradation algorithm with mature theory and low complexity. The basic idea is that in the sample room, when the next proximity samples belong to a category, the samples belong to the same category. The nearest neighbor is a single or multi-dimensional feature vector used to describe the sample at the nearest point. The closest point criterion may be the Euclidean distance of the feature vector [21]. The KNN classification algorithm consists of two steps: 1) In order to find a set of KNN for a given set,

it searches the training set using a certain distance metric; 2) It classifies queries based on the majority of the classes in KNN [19].

KNN is one of the oldest and simplest methods of pattern classification. However, it often produces the result of competition, in some fields, when skillfully combined with prior knowledge, it is already ahead of the highest level of technology. KNN regular pattern classification classifies each unlabeled sample through the majority of labels in the KNN in the training set. Therefore, its performance depends largely on the distance metric used to identify the nearest neighbor. KNN approach is one of the simplest methods in exponential data mining classification technology. Because of its simple implementation and excellent performance, it has been widely used in data mining and machine learning applications.

KNN means k nearest neighbors, which means that each sample can be represented by its nearest k neighbors. The core idea of KNN algorithm is that if most of the k most adjacent samples of a sample in the feature space belong to a certain category, then the sample also belongs to this category and has the characteristics of samples on this category.

This method only determines the category of the samples to be divided according to the category of the nearest one or several samples. The KNN method is only related to a very small number of adjacent samples in category decision making. Since the KNN method mainly depends on the surrounding limited adjacent samples, rather than on the method of discriminating the class domain to determine the category. Therefore, the KNN method is more suitable than other methods for the sample set to be divided which has more crossover or overlap in the class domain. So, this paper uses KNN to process experimental data. The algorithm is divided into the following steps:

Firstly, in order to find a group of KNN for a given set, it uses a certain distance metric to search the training set [27].

Secondly, it classifies the query in accordance with the majority of a class in KNN [28]. Suppose that $\{(a_i, b_i)\}_{i=1}^{n} \in D$ is the training set, and a_i is a v-dimensional vector, b_i is the label of class.

For a query a_j in a testing set (a_j, b_j), the algorithm obtain its unknown b_j in these ways:

i. Calculate the distances between a_j and each a_i of the training set. It is useful to measure the weight of neighbors so that the weight of closer neighbors is more significant than that of distant neighbors. For example, a common weighting scheme is to assign a weight of $1/D$ to each neighbor, where D is the distance to the neighbor[29].

ii. Put all the distances in descending order.

iii. Choose k samples in the training set which are the nearest to a_j.

iv. Determine the occurrence frequency of these k sample categories.

v. The class with the highest frequency of occurrence is regarded as the class label of a_j.

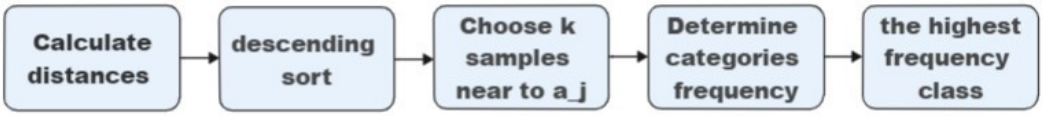

Figure 3. The process of obtain the class label of aj

The process is shown in **Figure 3**.

Simple calculation by KNN algorithm, we can quickly get a more accurate Lip language recognition results. In the absence of prior knowledge, most KNN classifications use simple Euclidean distances to measure differences between examples represented as vector inputs. However, the Euclidean distance measure does not take advantage of any statistical laws in the data that might be estimated from a large set of training samples. Ideally, the distance measurement of the KNN classification algorithm should be adapted to the specific problem being solved. For example, it is difficult to use the same distance measure in both mouth recognition and speech recognition to achieve optimal classification, even though in both tasks the distance is a fixed-size image calculated between the same tasks. In fact, many researchers have shown that KNN classification can learn distance measures from the example of labeling, and can be significantly improved. Even linear transformations with simple (global) input characteristics can yield better KNN classification. Our work on lip recognition builds on the success of these previous methods and takes it in a new direction [30].

3.3. K-fold cross validation

The data in machine learning modeling is divided into two parts, one is the test set, the other is the training set. The test set is separate from the training set and is used for the final evaluation of the model. When the training set is used in the training process, the phenomenon of overfitting often appears in the training process. Overfitting means that the data in the training set can match the model well, while other data outside the training set cannot match the model as well as the data in the training set. Since the data from the test set is used for the final evaluation of the model, we cannot use the test data to debug the model. The best way to solve the over-fitting is to separate the training data into another part to adjust the model, and this part of data is called validation data.

Validation data is part of the training data, but it does not participate in the training process to ensure the objectivity of the matching degree of the evaluation model to data outside the training set. Such assessments are usually done using K-fold cross validation (circular validation). Firstly, all the original data of the training set are divided into *K* group (*k*-fold), each subset is separately made a validation set, and the rest of the *k*-1 group of subsets are training sets. *K* models are obtained, and these *K* models are evaluated by validation sets respectively, and the cross-validation error is the sum average of the final error MSE (Means Square Error). Cross-validation (cyclic validation) is a validation method that makes efficient use of limited data, and its evaluation results can be closer to the final evaluation of the model by the test set. Such results can be used as indicators for model optimization [31].

The following is a simple example to illustrate the *k*-fold process. The raw data are shown below.

[0.2,0.4, 0.5, 0.6, 0.7, 0.8, 0.9,0.3]

If K=4,

$$\text{Fold } \alpha = [0.4,0.8] \quad (4)$$

$$\text{Fold } \beta = [0.5,0.7] \quad (5)$$

$$\text{Fold } \gamma = [0.6,0.9] \quad (6)$$

$$\text{Fold } \Omega = [0.2,0.3] \quad (7)$$

We can observe the following Table 1, four models will be used for cross-validation, respectively for training and testing.

Table 1. Four groups of cross-validation

Model	Train	Test
α	Fold α + Fold β	Fold Ω
β	Fold β + Fold Ω	Fold α
γ	Fold α + Fold Ω	Fold β
Ω	Fold γ + Fold Ω	Fold γ

4. Experiments Results and Discussions

4.1. Statistical Results

As can be seen in Ten runs of our method Table 2, our method obtained an overall accuracy 78.30± 1.30%, which is relatively effective and feasible. Additionally, the highest accuracy rate reached 80.08% and the lowest accuracy rate reached 75.78% in the results of ten runs. The gap between the results of each run is small, which reflects the stability of the test.

Table 2. Ten runs of our method

Run	Our Method
1	75.78
2	76.95
3	78.13
4	76.95
5	79.30
6	78.91
7	78.13
8	**80.08**
9	78.91
10	78.91
Average	**78.30± 1.30**

4.2. Training method comparison

Random subsampling, leave-one-out validation and K-fold cross validation are recommended as common cross validation methods. As full use of data in training and verification, K-fold cross validation was adopted in our experiment. Assuming K-fold partition, K-1 folds of data set are for training and remained folds are for validation. As a rule of thumb, value of K is frequently chosen as 5, 10, etc. In this test, 5-fold, 10-fold, 15-fold and 20-flod cross validation were compared. It is observed that 10-flod cross validation reaches 0.7% higher accuracy rate than 5-fold. Meanwhile, 15-flod is superior to 10-flod of value 0.5% and 20-fold is superior to 15-fold of value 0.3%. However, the value of K should be picked. Large value of K leads to large variance of estimator, consuming more calculation time. Contrarily, small value of K leads a large estimator deviation. Thus, 10-fold cross validation can be regarded as appropriate, which prevents overfitting and achieves out-of-sample estimation.

4.3 Comparison to State-of-the-art Approaches

Several state-of-the-art approaches were compared to our method. The comparison results were listed in Table 3. As can be observed that ANN, SVM, HMM, GMM, DNN were employed and the average accuracy was from about 50% to 80%, which indicates that the recognition rate is not high and it is not easy to improve the identification of lip language. Relatively speaking, our method has more advantages and potentials, which combines Wavelet entropy, K-nearest neighbor and K-fold cross-validation. Among them, Wavelet entropy can reduce the number of features, improve the training speed of the classifier, and reduce the need for memory. KNN is most suitable for image data with extremely low-dimensional features. Especially, KNN only retains feature vectors during training, which saves training time and only needs to be calculated during the verification phase. In fact, since each new instance will not be trained, the calculation time will be shortened. K-fold cross-validation can avoid overfitting and realize out-of-sample estimation. All of these enhanced the advantages of our method. Thus, proposed method is superior to other state-of-the-art approaches.

Table 3. Comparison with state-of-the-art approaches

Method	# Images/words	Overall Accuracy
K-Means + ANN [5]	6	64.4
Eigen lips + SVM	26	70.6
HOG+SVM [32]	26	71.3
CD-HMM [33]	300	47.48
CD-GMM [33]	300	49.19
CD-DNN [33]	300	77.19
GMM-HMM [34]	180	63.4
DNN-HMM [34]	180	64.9
WE-KNN (proposed)	520/26	78.30± 1.30

5. Conclusions

According to the description, wavelet transform technology is a series of image descriptors, no matter what kind of resolution the image needs, it can study and analyze the image. However, it also has many disadvantages. Because of its time scale representation, it is very suitable for the analysis of nonstationary signals in essence, but it needs more storage space and is very expensive in calculation. DWT allows image decomposition with different kinds of coefficients preserving the image information. The major disadvantage of this technique is that its excessive features increase computation times and storage memory [24].

K-nearest neighbor algorithm has high accuracy, is not sensitive to outliers, and does not need data input assumption, but its time complexity and space complexity are very high. The choice of k in KNN classification depends on the number of features and the number of training samples. With a low value of k, the classification result is more influenced by individual samples [35]. When the sample data is unbalanced, for example, the sample size of a class is very large, while the sample size of other classes is very small. At this time, when it inputs a sample, most of the k-nearest values of the class with large sample size may lead to its classification error. In addition, it is highly dependent on training data. Although all machine learning algorithms are highly dependent on data, this phenomenon is particularly serious for K-nearest neighbor algorithm. If we have one or two wrong data in the training data set, and it happens to be next to the value we need to classify, then the predicted data will not be accurate enough. From this, we can see that the fault tolerance of K-nearest neighbor algorithm to the training data is too poor.

The combination of Wavelet entropy, K-nearest neighbor and K-fold cross validation is an innovation. This method can effectively reduce the number of features, improve the training speed of classifier and reduce the demand for memory. There are significant differences in different types of we values, indicating that Wavelet entropy is very effective for lip shape change. In addition, KNN only retains the feature vectors during training, rather than training each vector, so it can reduce the calculation time and improve the efficiency. "We + KNN" combines the advantages of the two methods, making it superior to the most advanced lip recognition method in recognition accuracy. In the future continuous research, we will unremittingly optimize the algorithm, save time, strive to achieve better recognition accuracy than at present, and contribute to lip recognition.

Acknowledgements.
This work was supported by Natural Science Foundation of Jiangsu Higher Education Institutions of China (19KJA310002), The Philosophy and Social Science Research Foundation Project of Universities of Jiangsu Province (2017SJB0668).

References

[1] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, "A Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion," *IEEE Transactions on Multimedia,* vol. 18, no. 3, pp. 326-338, 2016.

[2] S. L. Wang, A. W. C. Liew, W. H. Lau, and S. H. Leung, "An Automatic Lipreading System for Spoken Digits

With Limited Training Data," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 18, no. 12, pp. 1760-1765, 2008.

[3] J. Shin, H. I. Kim, and R. H. Park, "New interface for musical instruments using lip reading," *Image Processing Iet,* vol. 9, no. 9, pp. 770-776, 2015.

[4] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading," *IEEE Transactions on Image Processing,* vol. 15, no. 10, pp. 2879-2891, 2006.

[5] Y. Liu, C. Lin, and J. Guo, "Impact of the Lips for Biometrics," *IEEE Transactions on Image Processing,* vol. 21, no. 6, pp. 3092-3101, 2012.

[6] X. Liu and Y. Cheung, "Learning Multi-Boosted HMMs for Lip-Password Based Speaker Verification," *IEEE Transactions on Information Forensics and Security,* vol. 9, no. 2, pp. 233-246, 2014.

[7] Z. Yi, L. Quan-jie, l. Hua, and Z. Li, *Intelligent Wheelchair Multi-modal Human-machine Interfaces in Lip Contour Extraction Based on PMM*. 2010, pp. 2108-2113.

[8] E. Petajan, "Automatic Lipreading to Enhance Speech Recognition," *[No source information available],* 01/01 1984.

[9] M. Kass, A.Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," *Intern. Jour. of Computer Vision,* vol. 1, pp. 321-332, 01/01 1987.

[10] R. Nath, F. Rahman, S. Nath, S. Basak, S. Audin, and S. Fattah, *Lip contour extraction scheme using morphological reconstruction based segmentation*. 2014, pp. 1-4.

[11] L. Yan, H. Ye, Y. Wang, and Y. Chang, "A lip localization method based on HSV transformation in smart phone environment," *International Conference on Signal Processing Proceedings, ICSP,* vol. 2015, pp. 1285-1290, 01/19 2015.

[12] K. D. Lee, M. J. Lee, and S.-Y. Lee, *Extraction of frame-difference features based on PCA and ICA for lip-reading*. 2005, pp. 232-237 vol. 1.

[13] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Networks,* vol. 2, no. 1, pp. 53-58, 1989/01/01/ 1989.

[14] L. Dong, S. W. Foo, and Y. Lian, "A Two-Channel Training Algorithm for Hidden Markov Model and Its Application to Lip Reading," *EURASIP Journal on Advances in Signal Processing,* vol. 2005, no. 9, p. 347367, 2005/06/21 2005.

[15] X. Liu and Y.-m. Cheung, "An Exemplar-Based Hidden Markov Model with Discriminative Visual Features for Lipreading," *Proceedings - 2014 10th International Conference on Computational Intelligence and Security, CIS 2014,* pp. 90-93, 01/20 2015.

[16] Y.-L. Lay, T. Chung Ho, H.-J. Yang, C.-S. Lin, and C.-Z. Lai, "The application of extension neuro-network on computer-assisted lip-reading recognition for hearing impaired," *Expert Systems with Applications,* vol. 34, pp. 1465-1473, 02/01 2008.

[17] Y. L. Lay, B. J. Yang, C. S. Lin, and B. F. Lee, "Lip Language Recognition for Specific Words," *Indian Journal of Science & Technology,* vol. 5, no. 11, pp. 3565-3572, 2012.

[18] M. Carrasco, J. Lopez, and S. Maldonado, "A second-order cone programming formulation for nonparallel hyperplane support vector machine," *Expert Systems with Applications,* vol. 54, no. jul., pp. 95-104, 2016.

[19] X. Wu, J. Yang, and S. Wang, "Tea category identification based on optimal wavelet entropy and weighted k-Nearest Neighbors algorithm," *Multimedia Tools and Applications,* vol. 77, no. 3, pp. 3745-3759, 2018.

[20] Y. Zhang, S. Wang, Z. Dong, P. Phillip, and J. Yang, "Pathological Brain Detection in Magnetic Resonance Imaging Scanning by Wavelet Entropy and Hybridization of Biogeography-Based Optimization and Particle Swarm Optimization," *Progress in Electromagnetics Research,* vol. 152, pp. 41-58, 2015.

[21] W. Li, P. Yi, Y. Wu, L. Pan, and J. Li, "A New Intrusion Detection System Based on KNN Classification Algorithm in Wireless Sensor Network," *Journal of Electrical & Computer Engineering,* vol. 2014, pp. 1-8, 2014.

[22] S. D. Zenzo, "Pattern recognition: a statistical approach: Devijver P A and Kittler J Prentice-Hall, Englewood Cliffs, NJ, USA (1982) pp 448. 24.95," *Image and Vision Computing,* vol. 3, no. 2, pp. 87-88, 1985.

[23] Y. Gao, C. Xue, R. Wang, and X. Jiang, "Chinese fingerspelling recognition via gray-level co-occurrence matrix and fuzzy support vector machine," *ICST Transactions on e-Education and e-Learning,* vol. 7, no. 20, p. 166554, 2020.

[24] X. X. Zhou *et al.*, "Detection of abnormal MR brains based on wavelet entropy and feature selection," *IEEJ Transactions on Electrical and Electronic Engineering,* vol. 11, no. 3, pp. n/a-n/a, 2016.

[25] S. H. Wang *et al.*, "Single slice based detection for Alzheimer's disease via wavelet entropy and multilayer perceptron trained by biogeography-based optimization," *Multimedia Tools & Applications,* 2016.

[26] S. Wang *et al.*, "Wavelet Entropy and Directed Acyclic Graph Support Vector Machine for Detection of Patients with Unilateral Hearing Loss in MRI Scanning," *Frontiers in Computational Neuroscience,* vol. 10, no. 4, 2016.

[27] M.-J. Lee, D.-W. Choi, S. Kim, H.-M. Park, S. Choi, and C.-W. Chung, "The direction-constrained k nearest neighbor query," *GeoInformatica,* vol. 20, no. 3, pp. 471-502, 2016/07/01 2016.

[28] E. Mangalova and O. Shesterneva, "K-nearest neighbors for GEFCom2014 probabilistic wind power forecasting," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1067-1073, 2016/07/01/ 2016.

[29] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician,* vol. 46, no. 3, pp. 175-185, 1992/08/01 1992.

[30] K. Q. Weinberger and L. K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *J. Mach. Learn. Res.,* vol. 10, pp. 207–244, 2009.

[31] N. B. A, S. C. A, S. S. M. B, and S. D. B, "A parameter independent fuzzy weighted k -Nearest neighbor classifier," *Pattern Recognition Letters,* vol. 101, pp. 80-87, 2018.

[32] M. Wand, J. Koutnk, and J. Schmidhuber, "Lipreading with long short-term memory," *Int. Conf. Acustics, Speech Signal Process,* pp. 6115-9, 2016.

[33] K. Thangthai, R. Harvey, S. Cox, and B.-J. Theobald, "Improving Lip-reading Performance for Robust Audiovisual Speech Recognition using DNNs," *The 1st*

Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing Vienna, Austria, pp. 127-131, 2015.

[34] M. H. Rahmani and F. Almasganj, "Lip-reading via a DNN-HMM Hybrid System Using Combination of The Image-based and Model-based Features," *3rd International Conference on Pattern Recognition and Image Analysis (IPRIA 2017),* pp. 195-199, 2017.

[35] H. A. Vrooman *et al.*, "Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification," *Neuroimage,* vol. 37, no. 1, pp. 71-81, 2007.