Information Sciences 373 (2016) 219–232

Contents lists available at ScienceDirect

Information Sciences

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

Visual speaker identification and authentication by joint spatiotemporal sparse coding and hierarchical pooling

Jun-Yao Lai[a], Shi-Lin Wang[a,*], Alan Wee-Chung Liew[b], Xing-Jian Shi[c]

[a] School of Information Security Engineering, Shanghai Jiaotong University, Shanghai, China
[b] School of Information and Communication Technology, Gold Coast Campus, Griffith University, QLD4222, Queensland, Australia
[c] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China

ARTICLE INFO

Article history:
Received 22 September 2015
Revised 5 August 2016
Accepted 5 September 2016
Available online 6 September 2016

Keywords:
Visual speaker identification
Visual speaker authentication
Sparse coding
Hierarchical pooling

ABSTRACT

Recent research shows that lip shape and lip movement contain abundant identity-related information and can be used as a new kind of biometrics in speaker identification or authentication. In this paper, we propose a new lip feature representation for lip biometrics which is able to describe the static and dynamic characteristics of a lip sequence. The new representation captures both the physiological and behavioral aspects of the lip and is robust against variations caused by different speaker position and pose. In our approach, a lip sequence is first divided into several subsequences along the temporal dimension. For each subsequence, sparse coding (SC in short) is adopted to characterize the minutiae of the lip region and its movement in small spatiotemporal cells. Then max-pooling based on a hierarchical spatiotemporal structure is performed on the SC codes to generate the final feature for each of the subsequence. Finally, the entire lip sequence is represented by a set of features corresponding to each subsequence in it. Experiments are carried out on a dataset with 40 speakers and compared with three state-of-the-art approaches. From the experimental results, it was observed that the proposed feature achieved high identification accuracy (an accuracy of 99.96%) and very low authentication error (a Half Total Error Rate (HTER) of 0.46%), and outperformed the other approaches investigated. Moreover, even with random variations caused by different speaker position and pose, the proposed feature still provides good identification (an accuracy of 99.18%) and authentication results (a HTER of 2.34%) and has much lower performance degradation compared with the other approaches investigated. Finally, even when there is only one training sample per speaker, the proposed feature still achieves high discriminative power (an accuracy of 98.39% and HTER of 2.62%).

1. Introduction

In the past decade, many kinds of biometric features have been proposed for person identification/recognition in various security systems [14,15,22,23,28,32]. Compared with Personal Identity Number (PIN), biometric feature based person identification method provides higher level of security as well as better convenience. Recent research shows that human lip alone is able to provide sufficient information related to the identity of its owner. The lip is a twin biometric in the sense that it

* Corresponding author. Address: No. 800, Dong Chuan Rd., School of Information Security Engineering, Shanghai Jiaotong University, 200240, Shanghai, China. Fax: +8621-3420-5025.

E-mail addresses: wsl@sjtu.edu.cn (S.-L. Wang), a.liew@griffith.edu.au (A.W.-C. Liew).

http://dx.doi.org/10.1016/j.ins.2016.09.015
0020-0255/© 2016 Elsevier Inc. All rights reserved.

contains both physiological feature and behavioral feature [9,12,18,22,29]. The appearance (texture and shape) of human lip region is unique and can be regarded as a physiological feature [9,12]. On the other hand, lip movement during utterance is influenced by a speaker's talking habit and differs from person to person, and can be regarded as a behavioral feature [18,29].

Compared with traditional biometric features, lip feature has the following advantages. First, lip biometric is video based, which improves the robustness of a security system by guaranteeing "liveness" [8,35]. Second, the price and utility of a video camera is more accessible than expensive sensors used in some biometric systems. Finally, lip biometric can be easily integrated with other biometrics features such as face, voice, etc. to construct a multi-biometric system to provide a very high level of security [10]. These factors suggest that lip biometric has a good potential of high user acceptance and industrial practicability. The main challenge of using lip biometric is how to extract useful features from lip video clip effectively. The feature extraction scheme is supposed to maintain sufficient information to identify the speaker and at the same time to be robust to variations caused by a speaker's pose and position towards the camera, etc.

Various methods have been proposed to verify a person's identity based on lip biometrics. These methods can be roughly classified into two categories: lip model based methods [5,16,20,24,31] and lip region based methods [3,19,27].

Lip model based methods require localization of human lips and thus are highly robust to variations in speaker's pose and distance to the camera. In [16], Luettin et al. used the Active Shape Model (ASM) to obtain lip shape and intensity profile features. These features were then input into an HMM with Gaussian mixtures to perform speaker identification. The classification achieves an accuracy of 97.9% on a relative small dataset (12 speakers and two video clips for each speaker). In [31], Wark et al. adopted the lip contour profiles as the visual feature and both the visual and audio information are integrated by a multi-stream HMM to perform speaker authentication. Their method achieved an accuracy of over 80% on the M2VTS dataset [24]. In [5] Broun et al. combined teeth and tongue information with lip geometric feature and adopted a polynomial-based model as the classifier. The verification performance of their scheme is characterized by a Half Total Error Rate (HTER) of 6.3% on the XM2VTS dataset [20]. In [27], Singh et al. extracted geometric features from lip contour and reduced them by the Minimum Redundancy Maximum Relevance (MRMR) method. An accuracy of 95.9% was obtained on a dataset containing 20 people and a vocabulary of ten English digits from zero to nine. In our previous work [29], we performed a comprehensive study of the discriminative power of the physiological and behavioral features of human lip. In our experiments, lip texture as the most discriminative physiological feature achieves an accuracy of 91.71%, while the optical flow of lip region as the best behavioral feature achieves an accuracy of 93.45% on a dataset with 40 speakers uttering the same password [29]. Obviously the performance of the above lip model based methods is highly dependent on the accuracy and robustness of the lip model extraction step.

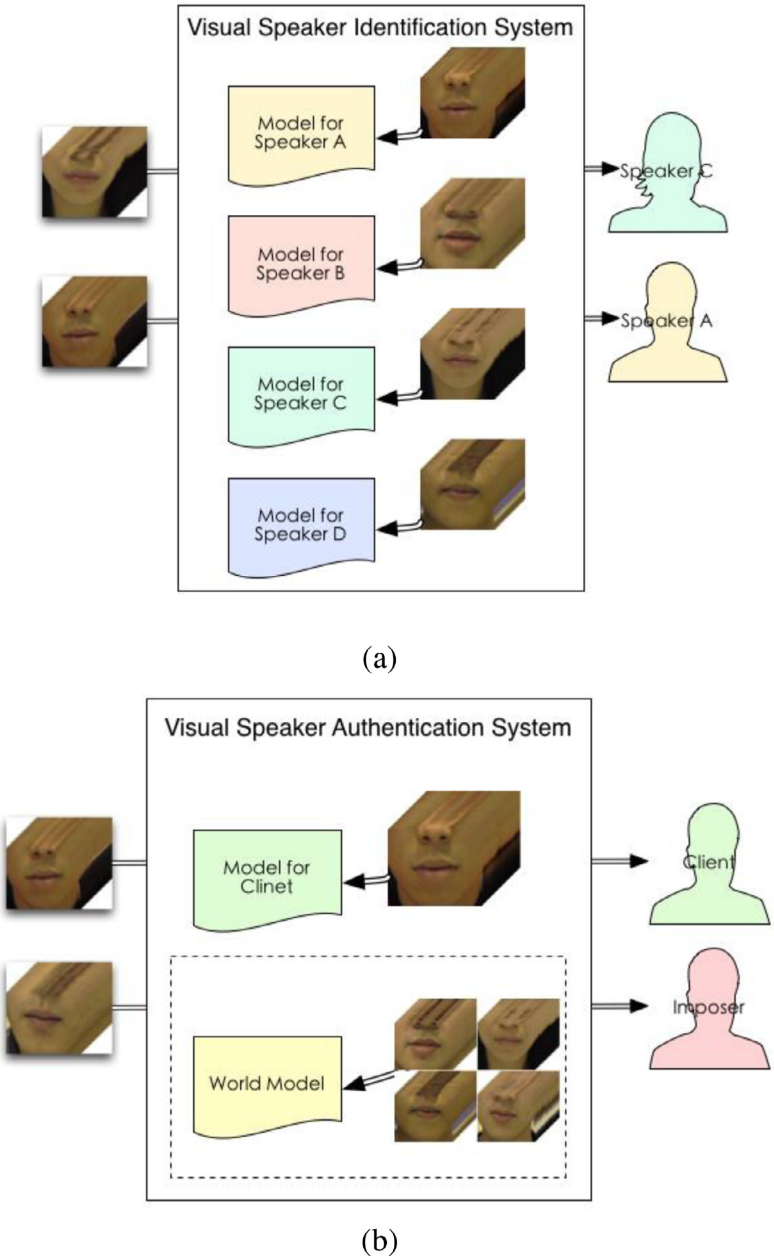
On the other hand, lip region based method extracts feature directly from a region in the image which contains the lip. As a consequence, the features contain both the lip region information and facial region information around the mouth (such as various kinds of mustache, moles, etc.). In [3] Bakry et al. used nonlinear mapping to initially parameterize human lip as a point in latent space. The latent space was then factorized using Kernel Partial Least Squares to reduce the dimension before being used as a speaker identification feature. Their scheme achieves an accuracy of 42.82% on AVLetters dataset [19] and 62.34% on OuluVs dataset [34]. In [13], Liu et al. used both the lip region based information (texture of mouth region and surrounding area) and lip model based information (geometric feature of lip contour) as features, and adopted HMM for speaker verification. They carried out experiments on a dataset especially established for lip-password speaker verification containing 46 speakers and achieved an Equal Error Rate (EER) of 3.91%. In [7] Chan et al. introduced Local Ordinal Contrast Pattern (LOCP) Histograms to represent the lip region. The Three Orthogonal Planes (TOP) was also adopted to combine the spatial and temporal information of the utterance video clip. Their experiments are based on the XM2VTS dataset and they achieved a very low HTER of 0.36%. The above approaches have demonstrated that the rough lip region can provide very high discriminative ability even without an accurate lip model. However, the extraction of speaker identity related information that is robust to variations caused by speaker's positions and poses remains a challenging task for the lip region based methods.

In order to achieve high performance without the need of an accurate lip contour, we propose a lip region based approach in this paper for visual speaker identification and authentication. The major contributions of our work can be summarized as follows. First, we propose a novel lip feature representation method based on sparse coding and hierarchical pooling. Sparse coding can effectively characterize the minutia of lip region and its movements. The hierarchical pooling improves the robustness against variations of speaker's pose and position. Second, the proposed scheme represents the lip image sequence in a spatiotemporal manner, which makes it suitable for lip biometric based on both physiological and behavior information.

The paper is organized as follows. Section II briefly introduces the problems of lip based speaker identification and authentication. Section III presents the proposed feature extraction scheme. Section IV presents the experiment results of the proposed scheme in comparison with three state-of-the-art approaches. Finally, Section V draws the conclusion.

2. Lip based visual speaker identification and authentication

Visual speaker identification and authentication are the two major applications in lip biometric. The term "visual" refers to the visual information source, i.e. visual images of the speaker's lip and its movements during utterance. Visual speaker identification, also called visual speaker recognition, aims to determine an unknown speaker's identity. Generally speaking,



(a)

(b)

Fig. 1. Lip biometric system for (a) speaker identification; (b) speaker authentication.

a visual speaker identification system usually works as follows (see Fig. 1(a)). In the training step, the lip sequences for each candidate are adopted to train the corresponding model for the speaker. In the identification step, the test lip sequence will be assigned to the speaker model with the highest probability or likelihood based on a certain similarity measure.

On the other hand, visual speaker authentication, also called visual speaker verification, aims to use lip biometric to verify whether the speaker is the one who he/she claims to be. In a visual speaker authentication system, lip sequences of all the clients (authorized users) are used to pre-train their corresponding models. In the verification stage, the test lip sequence will be assigned to the claimed identity if the similarity between the sequence and the corresponding client model is above a preset threshold. In this sense, authentication performs a 1 to 1 match while identification performs a 1 to N (N is the number of speakers in the database) match. Note that in many authentication systems, a world model is adopted which is pre-trained by the lip sequences of other speakers except the client. Then the verification is performed as a two-class classification procedure. A sketch of the visual speaker verification system is given in Fig. 1(b) and the major differences between the above two lip biometrics applications are briefly summarized in Table 1.

Table 1
Differences between visual speaker identification and authentication.

	Visual Speaker Identification	Visual Speaker Authentication
Functionality	To identify who the speaker is	To verify the claimed identity of the speaker
Test mode	1 to N match, output the best matching speaker's label	1 to 1 match, output whether to accept or reject the verification request
Typical application scenario	Identity retrieval	Access control

3. Proposed lip feature extraction scheme

For both the visual speaker identification and authentication systems, the most critical issue is how to extract lip features which are representative for each speaker and discriminative against other speakers. In this paper, a lip feature representation with high discriminative power is proposed. The overall feature extraction scheme is shown in Fig. 2. In this scheme, the lip sequence is first divided into a series of overlapping subsequences using a preset time window. The T image frames, each of size M by N, in the subsequence are stacked together to form a 3D spatiotemporal cube of dimension $M \times N \times T$. The spatiotemporal cube is then partitioned into small spatiotemporal cells and sparse coding (SC in short) is applied to each cell to capture its spatiotemporal characteristics. Then a hierarchical structure is used to describe each subsequence at various scales and locations. For each layer of the hierarchical structure, the corresponding feature is generated from all the SC codes of the spatiotemporal cells in that layer by max-pooling. The feature for the subsequence is then obtained by concatenating the features from all layers. Finally, the entire lip sequence is represented by a set of features obtained from all the subsequences.

3.1. Sparse coding on spatiotemporal cells

Sparse coding [2,11,21] is a generative model for signal analysis. In sparse coding, the input signal is described as a linear combination of elementary signals in a pre-built dictionary. An important attribute in sparse coding is the sparsity, i.e., there are a limited number of elementary signals used in the description of the input signal. In our approach, the sparse coding technique is adopted to extract representative features describing every minutiae of the entire lip sequence.

We denote a lip sub-sequence with an image size of M by N and a time window of T frames by a spatiotemporal cube. A series of densely overlapping (with the step size of one pixel or one frame) cubic cells with size $b \times b \times b$ can be extracted from the cube as illustrated in Fig. 3. Each cell describes the local characteristics of a small lip region and its short-time variation, which contain speaker-identity related information. Then sparse coding is employed to extract such identity related information in an efficient manner from each cell. Given a specific cell, let I be the vector composed of all the grayscale values in the cell (with the vector size of b^3). Then the corresponding sparse code w for the cell can be obtained by,

$$w = argmin_c \|I - Dc\|_{\ l^2} \ subject\ to\ \ \|c\|_{l^0} \le s \tag{1}$$

where D is the pre-trained dictionary containing all the K representative elements (K is the size of the dictionary) and s is the sparsity value. $\|\cdot\|_{l2}$ and $\|\cdot\|_{l0}$ denote L-2 and L-0 norm distance, respectively. In our approach, the final coding result w is obtained by minimizing the reconstruction error, i.e. $\|I - Dc\|_{l^2}$, using the Orthogonal Matching Pursuit (OMP) algorithm [25]. A brief flowchart of the OMP algorithm is given in Fig. 4.

The dictionary used in sparse coding should reflect the underlying characteristics of the input signal. In the proposed scheme, a series of cells randomly selected from the lip sequences of each speaker in the training set are used to construct the dictionary. We use K-SVD [1] to generate the dictionary.

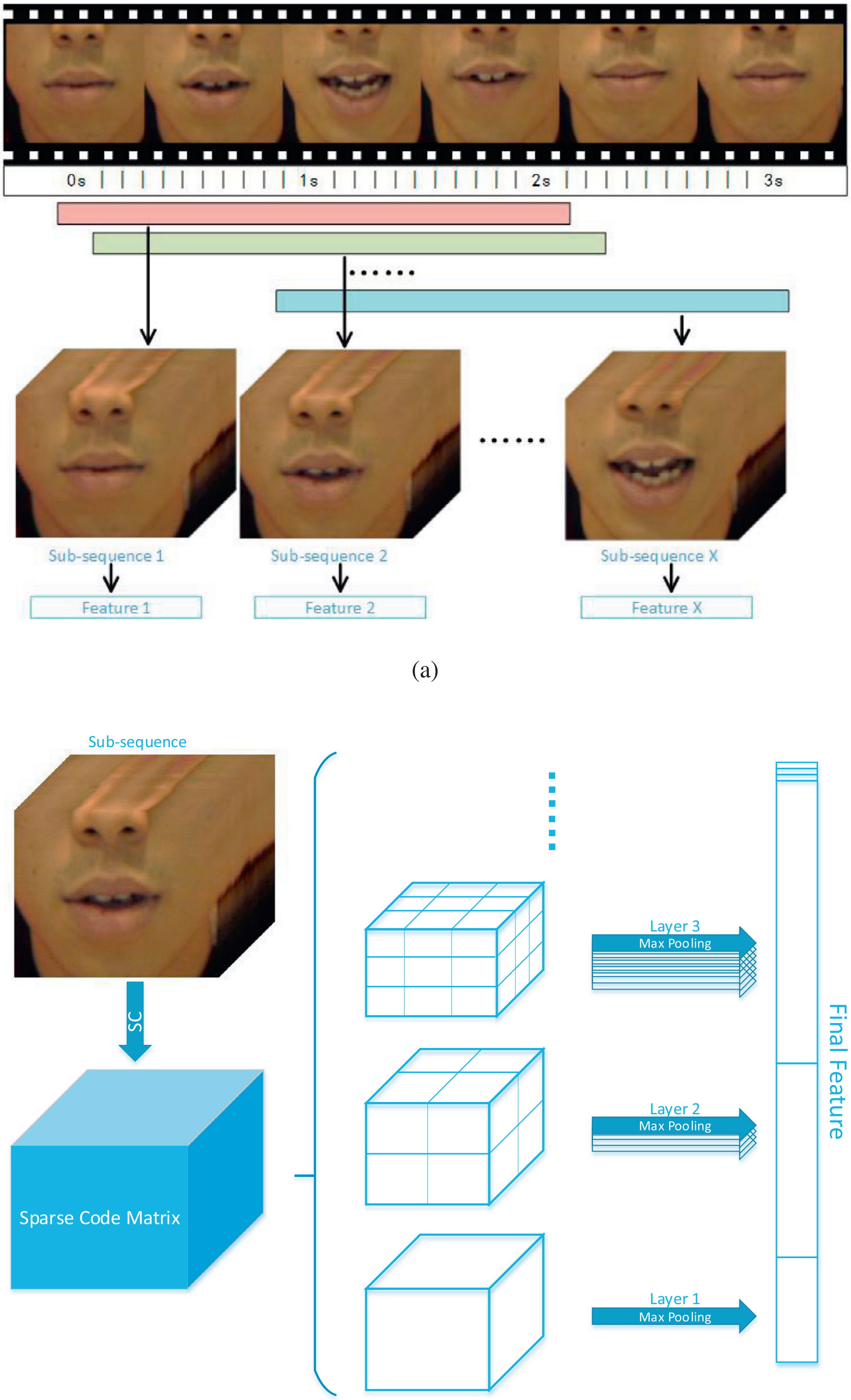
Let Φ be the training sample matrix, i.e. $\Phi = [I_1, I_2, \ldots, I_\eta]$, where η is the total number of training samples, and $W = [w_1, w_2, \ldots, w_\eta] = [w^1, w^2, \ldots, w^\eta]^T$ is the corresponding sparse code matrix obtained by the OMP algorithm [25], where the subscript k in w_k and I_k indicates the column index and the superscript k in w^k indicates the row index. The K-SVD algorithm aims to minimize the overall reconstruction error in (2) by an iterative optimization approach.

$$\min_{D,W}\left\{\left\|\Phi - DW_F^2\right\|\right\} \ subject\ to\ \forall i,\ 1 \le i \le \eta,\ \ \|w_i\|_{l^0} \le s \tag{2}$$

where $\|A_F\|$ means the Frobenius norm of matrix A, which is defined as $A_F = \sqrt{\sum_{ij} A_{ij}^2}$. The K-SVD algorithm runs as follows.

(1) Initialize D by a matrix whose columns are randomly selected input signals with L-2 normalization.
(2) Compute the sparse code matrix W using the current dictionary D using the OMP algorithm [25].
(3) For each column $k = 1, 2, \ldots, K$ in D (denoted by d_k), select the candidate training sample set using this atom, i.e., $\Delta_k = \{I_i | w^k(i) \neq 0,\ 1 \le i \le \eta\}$. Calculate the representation error matrix E_k by (3).

$$E_k = \Phi - \sum_{j \neq k} d_j w^j \tag{3}$$



(a)

(b)

Fig. 2. The subsequence division (a) and the feature extraction (b) scheme of the proposed approach.

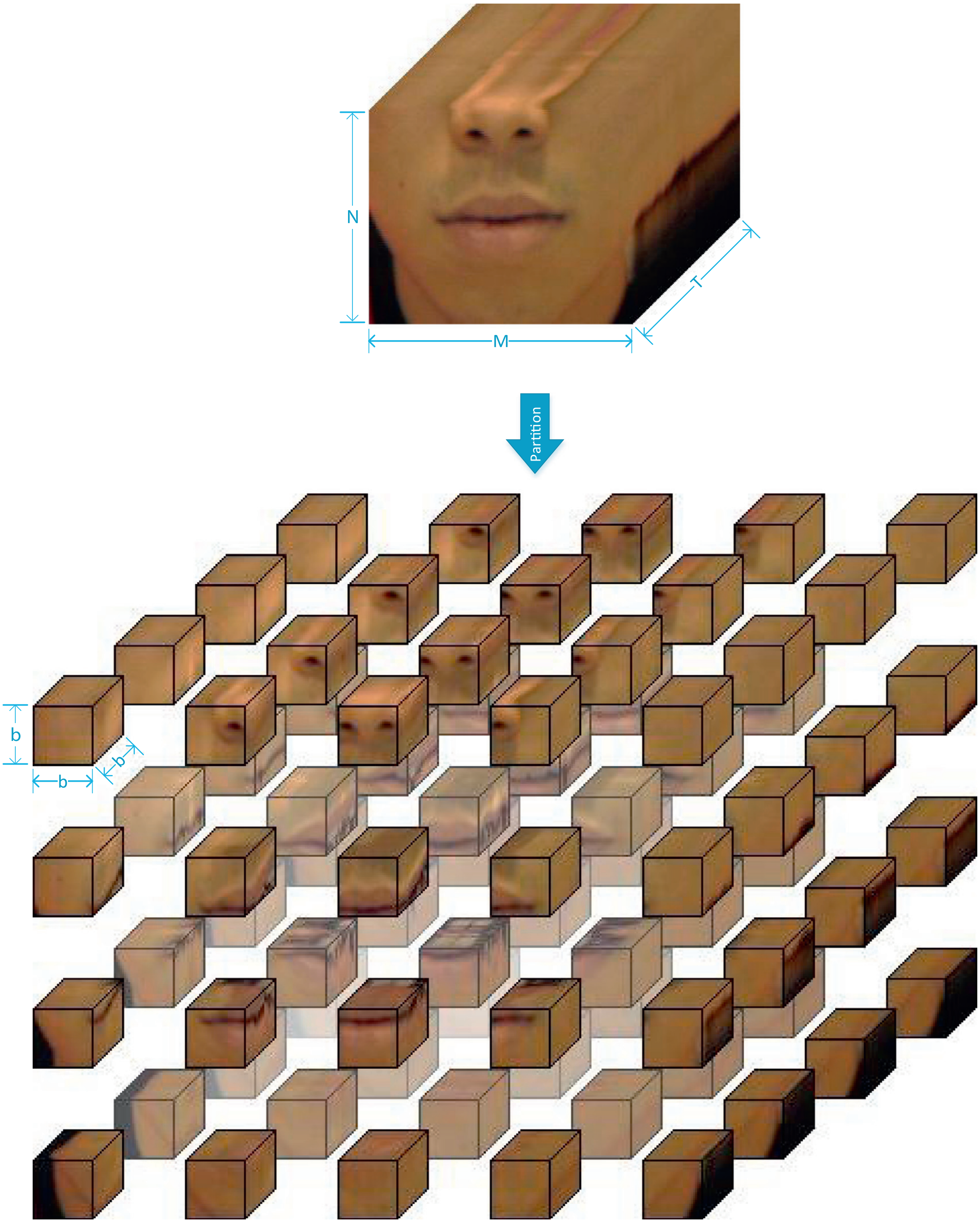


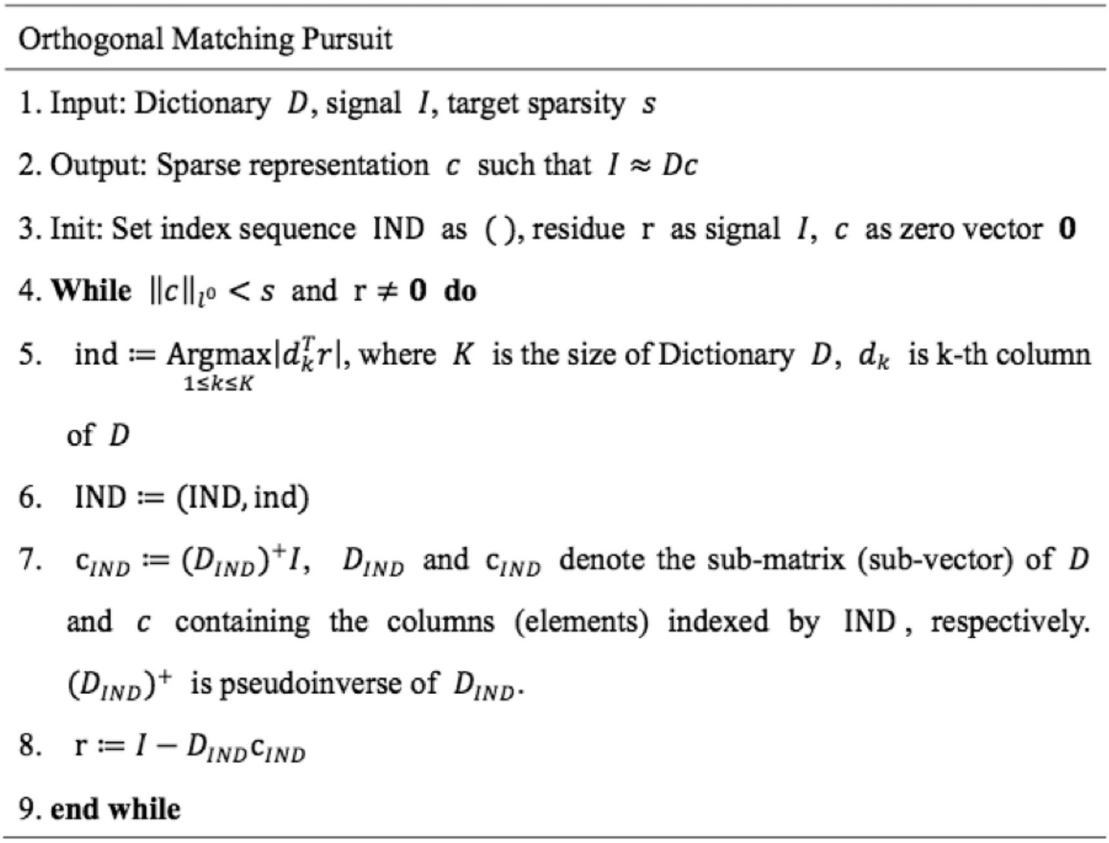
Fig. 3. Cell partitioning from a spatiotemporal cube of a subsequence.

Note that when calculating E_k, only the columns related to the samples in Δ_k are considered and other columns are set to zero in Φ and W.

(4) Apply the singular vector decomposition (SVD) on E_k by (4), where U and V are $b^3 \times b^3$ and $\eta \times \eta$ unitary matrices, respectively, and Σ is a diagonal matrix composed of singular values of E_k. Update d_k to be the first column of U.

$$E_k = U \Sigma V^T \tag{4}$$

(5) Repeat step 2 to step 4 until the average reconstruction error is smaller than a preset threshold (1 in our experiments), i.e. $\|\Phi - DW\|_F^2/\eta \le 1$, or the maximum number of iteration (100 in our experiments) is reached. Then the optimal dictionary D is obtained from the training samples.



Orthogonal Matching Pursuit

```
1. Input: Dictionary $D$, signal $I$, target sparsity $s$
2. Output: Sparse representation $c$ such that $I \approx Dc$
3. Init: Set index sequence IND as ( ), residue r as signal $I$, $c$ as zero vector $\mathbf{0}$
4. While $\|c\|_{l^0} < s$ and $r \neq \mathbf{0}$ do
5.    ind := $\mathrm{Argmax}_{1 \le k \le K} |d_k^T r|$, where $K$ is the size of Dictionary $D$, $d_k$ is k-th column
      of $D$
6.    IND := (IND, ind)
7.    $c_{IND} := (D_{IND})^{+} I$,  $D_{IND}$ and $c_{IND}$ denote the sub-matrix (sub-vector) of $D$
      and $c$ containing the columns (elements) indexed by IND, respectively.
      $(D_{IND})^{+}$ is pseudoinverse of $D_{IND}$.
8.    $r := I - D_{IND} c_{IND}$
9. end while
```

Fig. 4. Flowchart of the OMP algorithm.

Note that in each iteration, the total reconstruction error decreases monotonically, which guarantees convergence to a local minimum. Interested readers can refer to [1] for detailed convergence proof of the K-SVD algorithm. Finally, with the trained dictionary D by K-SVD, a 3D matrix is generated for the sub-sequence where each element is the sparse code for the cell in the corresponding position in the spatiotemporal cube.

3.2. A hierarchical representation and feature generation by spatiotemporal max-pooling

The 3D SC code matrix cannot be used directly to discriminate a speaker's identity due to the extremely high dimensionality of the resulting feature. For dense cell coverage of the spatiotemporal cube, the dimension of the feature vector of each subsequence is in the vicinity of $M \times N \times T \times K$, where K is the dictionary size of the sparse code. Such high dimension cannot be handled by most classification algorithms. In addition, such representation is not invariant to translation and time delay.

Instead, we propose a hierarchical spatiotemporal structure for feature generation from the 3D SC code matrix as shown in Fig. 5. The 3D SC code matrix for the lip subsequence is described at multiple hierarchical layers in the spatiotemporal domain. In each layer, the spatiotemporal cube of the subsequence is divided into a number of spatiotemporal sub-cubes. The spatiotemporal extent covered by each sub-cubes decreases, i.e., becoming more localized, as the layer index increases. A spatiotemporal sub-cube of the k-th layer is denoted by C_i^k, where i indicates its spatiotemporal position in this layer.

For each sub-cube, a representative feature is adopted to describe its characteristics. In our approach, max-pooling is applied on the SC codes of all the cells within the spatiotemporal sub-cube to generate the feature. Compared with other image statistics, such as histograms, max-pooling is supported by biophysical evidence in visual cortex and has shown good performance in many image classification algorithms [26,33]. Denote the SC code of each cell in a spatiotemporal sub-cube C by x_i, $i \in C$. The max-pooling result of the representative feature for this sub-cube is simply the component-wise maxima over all the SC codes within it [26], i.e.,

$$F(C) = \max_{i \in C} [\max(x_i(1), 0), \cdots, \max(x_i(K), 0), \ \max(-x_i(1), 0), \cdots, \max(-x_i(K), 0)] \tag{5}$$

In Eq. 5, i ranges over every cell in the sub-cube and $x_i(j)$ denotes the j-th component of the SC code x_i. Note that the positive and negative components of the sparse codes are split into separate features to allow separate treatment of positive

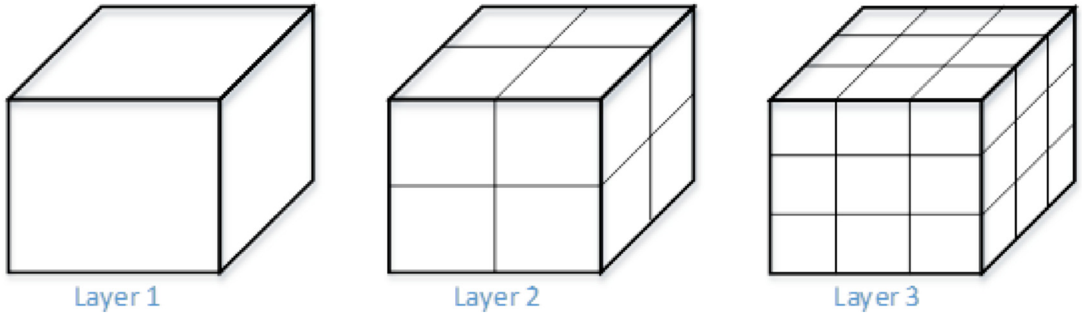


Fig. 5. The hierarchical spatiotemporal structure.

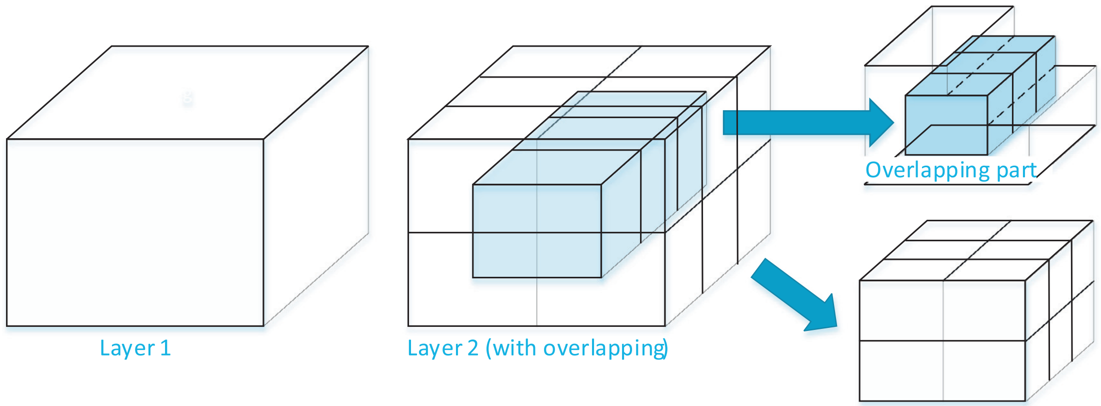


Fig. 6. The proposed alternative spatiotemporal structure.

and negative responses respectively. Then, the feature of the lip sub-sequence is obtained by concatenating features of every sub-cubes in each layer by,

$$F_j = \left[F_j\left(C_1^1\right), \cdots, F_j\left(C_{n_1}^1\right), F_j\left(C_1^2\right), \cdots, F_j\left(C_{n_2}^2\right), \cdots, F\left(C_1^k\right), \cdots, F_j\left(C_{n_k}^k\right)\right] \tag{6}$$

where the number of sub-cubes in the k-th layer is denoted by n_k and the subscript j is the index of the starting frame of the subsequence.

The proposed hierarchical structure with spatiotemporal max-pooling can effectively capture both global and local spatiotemporal patterns of the lip sequence. The proposed structure is similar to the spatial pyramid representation for image classification [26], where pooling over the sub-cubes in the same layer provides spatiotemporally localized descriptions, i.e. locally dominant features, for different locations within the cube at a specific scale, while pooling over smaller sub-cubes in different layers provides a description of successively more globally-to-locally dominant features of the entire cube. Moreover, for each sub-cube, the spatiotemporal max-pooling guarantees robustness to local spatial translations and time delays. Therefore, the final feature set is translation invariant at different granularity of spatiotemporal extent.

It should be noted that the dimensionality of the final feature vector for a subsequence is $\sum_{l=1}^{L} l^3 \times K$, where L is the number of layers used, and it will increase greatly with the increase of L (L is set to 3 in our approach as shown in Fig. 5). In order to reduce the dimensionality, an alternative hierarchical structure is proposed as shown in Fig. 6. In this structure, the sub-cubes in layer 2 overlap in the spatial domain, where another sub-cube is located in the center region. Since the mouth region is usually located in the center of the lip image, the above structure actually gives additional attention to the mouth region in the final feature. Also, note that the temporal extent of layer 2 in Fig. 6 is equal to the temporal extent of layer 3 in Fig. 5. In the alternative structure depicted in Fig. 6, we have 1 and 15 sub-cubes in layer 1 and 2, respectively.

Finally, the entire lip sequence is represented by a set of features, each extracted from a sub-sequence, i.e. $\{F_1, F_2, \ldots\}$. In our algorithm, the SC codes for all the cells in the entire sequence is calculated once and stored for constructing the features for all the subsequence. In addition, the features for the current subsequence can be obtained efficiently with the knowledge of the features for the previous subsequence. With the above steps, the calculations for the feature set are greatly reduced.

Different from previous approaches [6,7,16], a set of features rather than one feature is adopted to represent the lip sequence, which leads to slight differences in the training, evaluation (for visual speaker authentication only) and testing procedures. In training and evaluation, the samples are obtained feature-wise, i.e. each feature is treated as a sample.

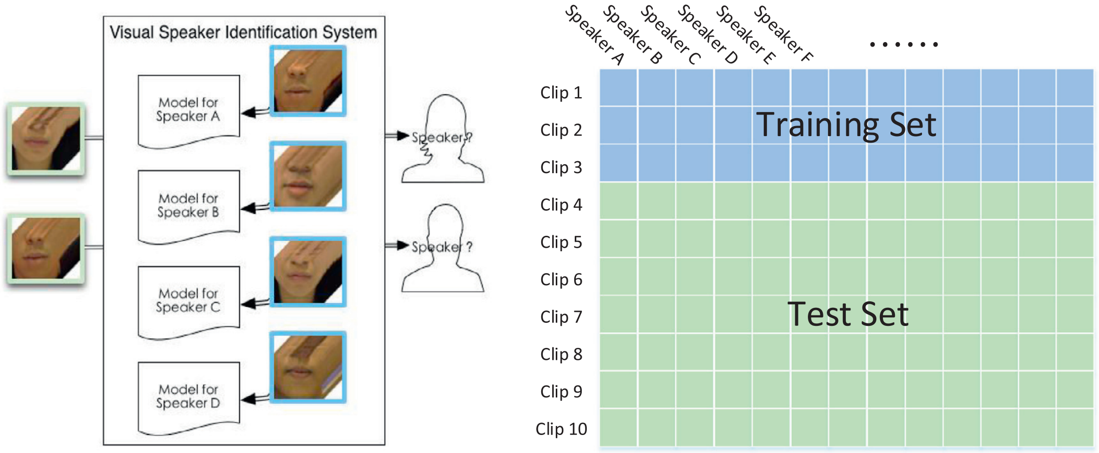


Fig. 7. The training and test set partitioning for visual speaker identification.

Hence, one lip sequence will generate more than one samples for training and evaluation. In testing, the final identification/authentication result is obtained by voting over the corresponding results of all the features in the test sequence.

4. Experiments and discussions

4.1. Experimental setup

To evaluate the performance of the proposed lip feature extraction scheme in visual speaker identification and authentication, the dataset in [29] is used in our experiment. The dataset is composed of 40 speakers (29 male and 11 female) and each speaker is asked to repeat the phrase “3725″ for ten times. This phrase is selected because it covers a wide variation of lip shapes and movement. Each utterance lasted for 3 seconds and consists of 90 frames with a spatial resolution of 220×180. Linear SVM [4] is adopted as the classifier for its superior performance in dealing with high dimensional sparse features. In our experiment, each lip image is resized to 55×45 to reduce the computational complexity and the SC dictionary is constructed from 50,000 cells randomly selected from the sequences in the training set.

In the visual speaker identification experiments, the dataset is partitioned into training set and testing set as shown in Fig. 7. For each speaker, three utterances are randomly selected as the training data and the rest are used for testing. The identification performance is measured by the identification accuracy, i.e.,

$$Accuracy = \frac{\text{No. of testing samples being correctly classified}}{\text{No. of all the testing samples}} \tag{7}$$

In order to avoid any bias in the selection of training samples, the final accuracy is obtained by averaging the results of ten random trials.

In the visual speaker authentication experiments, we employed a protocol which is similar to Lausanne protocol [17], which works as follows (as shown in Fig. 8):

(1) Client training samples: 3 clips per speaker.
(2) World model training samples: 30 clips from ten other speakers randomly selected from the dataset.
(3) Client evaluation samples: 3 clips per speaker.
(4) World model evaluation samples: the remaining 70 clips from the ten speakers.
(5) Testing set: the remaining 4 clips from client and 290 clips from the other 29 speakers as imposters.

Two measures, i.e. the equal error rate (EER) and the half total error rate (HTER), are adopted to evaluate the authentication performance. EER is obtained when the false accept rate (FAR) is equal to the false rejection rate (FRR) by tuning the threshold of the trained classifier using the evaluation set. However, as indicated in [20], the equal error rate cannot reflect the expected system performance when the testing set is unseen. To overcome this shortcoming, the Half Total Error Rate (HTER) is adopted as a more reasonable measure for speaker authentication. In our experiments, the EER and HTER are calculated as follows.

(i) Obtain the linear SVM with soft classification for each client based on the training set.
(ii) Tune the threshold of each client’s SVM to obtain EER using the evaluation set. The threshold is recorded for the client.
(iii) Based on the classifier obtained in step i) and the threshold obtained in step ii), perform classification using the testing set. HTER is calculated by HTER = (FAR+FRR)/2.

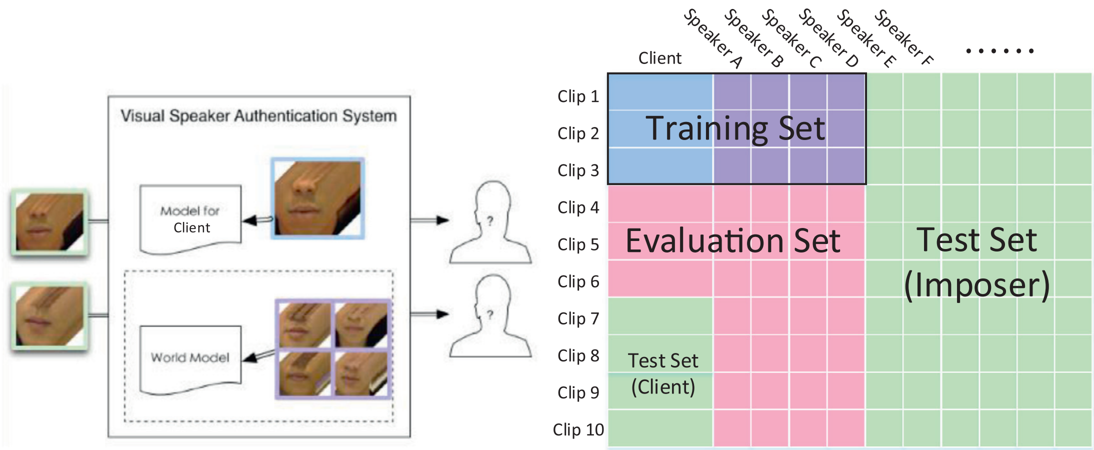


Fig. 8. The training, evaluation and test set partitioning for visual speaker authentication.

Table 2
Speaker identification results in % using various parameters.

T *a*	6	12	24	48	72
Size of the cubic cell = 3					
128	98.21	98.61	97.75	98.54	98.96
256	98.54	99.04	99.07	99.04	99.21
512	99.21	99.18	99.46	99.07	99.43
Size of the cubic cell = 4					
128	98.54	98.04	98.18	98.54	98.54
256	98.86	98.79	99.36	99.00	99.25
512	98.96	99.04	99.32	99.46	99.68
Size of the cubic cell = 5					
128	98.14	98.68	98.04	98.21	99.18
256	98.18	98.68	99.11	98.82	99.07
512	98.93	99.14	99.54	99.61	99.61

4.2. Parameter setting

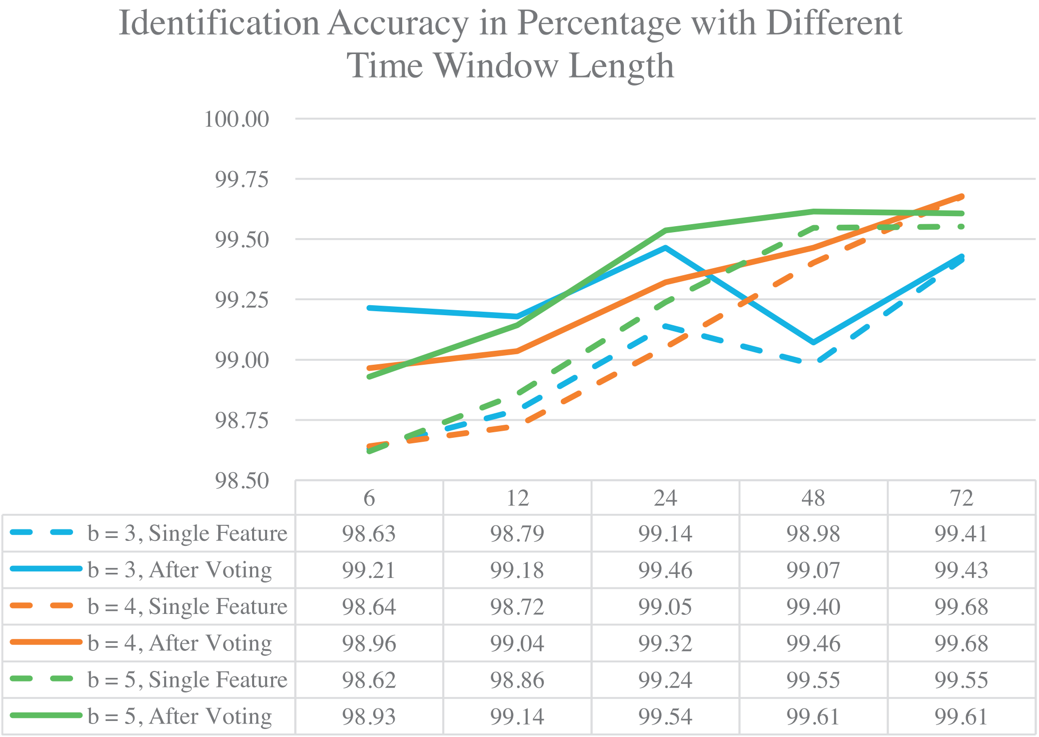
In the proposed feature extraction approach, three parameters need to be set, i.e. the size of the SC dictionary *K*, the size of the cubic cell *b*, and the time window length *T*. In this sub-section, we analyze the visual speaker identification and authentication performance with various parameter settings.

In sparse coding of spatiotemporal cells, the size of the dictionary affects the accuracy of the sparse representation. Generally speaking, a larger dictionary lead to a smaller reconstruction error at the cost of higher feature dimension. The size of the spatiotemporal cell *b* is another important parameter in the proposed feature, which determines the level of detail described by the cell. In our experiments, three dictionary sizes, from 128 to 512, and three cell sizes, from $3 \times 3 \times 3$ to $5 \times 5 \times 5$, are investigated.

In our feature representation, the input sequence is divided into a number of subsequences, each of duration *T* frames. A subsequence with small *T* depicts short duration lip movement, and its feature could differ quite substantially from other subsequences due to the dynamic nature of the content. In contrast, a subsequence with large *T* depicts the long-term lip variations during utterance. Five time window lengths, from 6 frames to 72 frames, are investigated in our experiments to see their effect.

Table 2 shows the visual speaker identification performance with various settings of dictionary size *K*, cell size *b* and time window length *T*. From the Table, the following observations can be made.

First, the discriminative power of the proposed feature usually increases (resulting in higher identification accuracy) with the increase of dictionary size *K*. A smaller dictionary size would give larger reconstruction error, resulting in the loss of some identity-related information. For the proposed hierarchical structure of Fig. 5, the dimension of the feature is $72 \times K$ (with 36 sub-cubes and separate features for positive and negative responses). When *K* is greater than 512, the feature dimension will exceed $512 \times 72 = 36{,}864$, which results in very high computational complexity and large memory cost during the classifier training process. Hence, identification/authentication results for features with dimension exceeds 40,000 is not considered in our experiments.



	6	12	24	48	72
b = 3, Single Feature	98.63	98.79	99.14	98.98	99.41
b = 3, After Voting	99.21	99.18	99.46	99.07	99.43
b = 4, Single Feature	98.64	98.72	99.05	99.40	99.68
b = 4, After Voting	98.96	99.04	99.32	99.46	99.68
b = 5, Single Feature	98.62	98.86	99.24	99.55	99.55
b = 5, After Voting	98.93	99.14	99.54	99.61	99.61

Fig. 9. Identification accuracy in % using different time window length with the dictionary size of 512.

Second, as mentioned in Section 3.2, the final identification result of a testing lip sequence is derived by voting over the classification results of the features of each sub-sequence. To see how the time window length of the subsequence affects the identification result, Fig. 9 shows the classification results based on a single subsequence (before voting) and the entire lip sequence (after voting) for a dictionary size of 512. The trend is similar for other dictionary size. It can be seen that when T is small, (e.g. $T=6$), the identification results using the entire sequence through voting are better than that of the single feature. However, the gain in identification accuracy becomes smaller with the increase of T and almost disappear when $T=72$. When T is small, the content of the utterance represented by the subsequence varies greatly, which causes difficulties for the classifier. In this case, voting over the results of the overlapping subsequences was able to improve the final identification result. When T is large, e.g. $T=72$ in a sequence with a total of 90 frames, the content represented by each subsequence tends to be very similar to the entire utterance. Since there is little difference in content among all the subsequences, voting becomes ineffective.

Third, as observed in Table 2 and Fig. 9, the three selections of cell size achieve similar identification results which demonstrate that any small spatiotemporal cells from $3\times3\times3$ to $5\times5\times5$ can achieve reliable results. Hence, the parameters of the optimum feature using the hierarchical structure in Fig. 5 can be set as follows: $K=512$, $b=4$, $T=72$.

As mentioned in Section 3.2, in order to increase the dictionary size while keeping an acceptable feature dimension, the alternative hierarchical structure shown in Fig. 6 can be employed, whose feature dimension equals to 32 times (16 sub-cubes) the size of the dictionary. The identification and authentication results using the original and alternative hierarchical structures are listed in Table 3. Table 3 shows that comparing with the original structure, the alternative structure provides comparable discriminative power with less than half the feature dimension. On the other hand, with comparable overall feature dimension, the alternative structure with a dictionary size of 1024 outperforms the original structure with a dictionary size of 512. The above experimental results demonstrated the effectiveness of the proposed alternative structure. The parameters of the optimum feature using the alternative structure can be set as: $K=1024$, $b=3$ and $T=72$.

4.3. Performance comparison with the State-of-the-art approaches

In order to provide a comprehensive evaluation of the proposed approach, three approaches widely used in visual speaker identification/authentication, namely Luettin et al.'s approach (Luettin's in short) [16], Cetingul et al.'s approach (Cetingul's in short) [6] and Chan et al.'s approach (Chan's in short) [7], are adopted for comparison. Luettin's and Cetingul's approaches are lip model based. In Luettin's approach, the lip shape and the intensity profile along the lip contour points are adopted as lip features. In Cetingul's approach, both the lip intensity and the lip motion features are used. To eliminate error due to erroneous model extraction, the extracted lip models from these two approaches are checked manually to ensure they match

Table 3
Identification and authentication results in % using the original (Fig. 5) and alternative (Fig. 6) spatiotemporal structures for $T = 72$.

K		128	256	512	768	1024
Size of the cubic cell = 3						
Original Spatiotemporal Structure	Accuracy	98.96	99.21	99.43	N/A	N/A
	EER/HTER	0.10/2.39	0.00/1.37	0.01/1.22	N/A	N/A
Alternative Spatiotemporal Structure	Accuracy	98.79	99.21	99.57	99.75	**99.96**
	EER/HTER	0.11/2.31	0.02/1.99	0.02/1.43	0.01/0.93	**0.00/0.46**
Size of the cubic cell = 4						
Original Spatiotemporal Structure	Accuracy	98.54	99.25	99.68	N/A	N/A
	EER/HTER	0.01/2.05	0.00/1.83	0.01/1.03	N/A	N/A
Alternative Spatiotemporal Structure	Accuracy	98.57	99.07	99.57	99.46	99.79
	EER/HTER	0.02/1.95	0.02/1.99	0.02/1.15	0.00/0.78	0.00/0.69
Size of the cubic cell = 5						
Original Spatiotemporal Structure	Accuracy	99.18	99.07	99.61	N/A	N/A
	EER/HTER	0.08/1.87	0.04/1.94	0.01/1.40	N/A	N/A
Alternative Spatiotemporal Structure	Accuracy	98.71	99.18	99.61	99.79	99.61
	EER/HTER	0.02/2.07	0.00/1.73	0.00/1.08	0.01/0.97	0.00/0.76

Table 4
Performance comparison (in %). The best results are highlighted.

	The proposed feature		Chan's	Cetingul's	Luettin's
	1	2			
Accuracy	**99.96**	99.71	98.61	95.54	88.33
EER	**0.00**	0.01	0.20	1.65	8.28
HTER	**0.46**	1.03	1.56	6.75	16.61

Table 5
Robustness tests (in %) for the four approaches investigated.

	The proposed feature		Chan's	Cetingul's	Luettin's
	1	2			
Accuracy	**99.18**	96.29	90.50	95.48	88.45
EER	**0.11**	0.51	2.55	2.08	8.31
HTER	**2.34**	5.32	8.00	7.16	17.73

the actual lip contour in our dataset. The hidden Markov model (HMM) is employed as the classifier in both approaches. On the other hand, similar to our approach, Chan's approach is lip region based and no prior lip contour information is required. In their approach, the Local Ordinal Contrast Pattern Histograms with a three orthogonal planes (TOP) configuration are adopted as features.

Table 4 lists the identification/authentication results for all the approaches. For our approach, two kinds of features are investigated, where feature 1 and 2 denotes the optimum feature obtained using the alternative spatiotemporal structure (with $K = 1024$, $b = 3$ and $T = 72$) and that using the original structure (with $K = 512$, $b = 4$ and $T = 72$), respectively. From the Table, it is observed that the region-based approaches (Chan's and ours) outperform the two model-based approaches (Luettin's and Cetingul's). Moreover, the proposed features can achieve better performance than Chan's approach with an identification accuracy improvement of 1.35% and an authentication EER/HTER improvement of 0.20%/1.1%.

4.3.1. Robustness against variations

In practical visual speaker identification/authentication applications, the captured lip sequences are usually affected by variations such as: i) distance between the speaker and camera; ii) position of the mouth in the captured window; iii) rotations caused by various head poses. In order to test the performance of the approaches to these variations, a series of robustness tests are performed. An affine transformation with translation (to simulate the position variations), rotation (to simulate the head pose variations) and scaling (to account for the distance variations) are applied to the original lip sequence to generate lip sequences with variation.

Specifically, for each lip sequence in the original dataset, an affine transformation with evenly distributed random parameters is applied to generate new lip sequences, where the translation, rotation, and scaling parameters ranged from -20 to 20 pixels (in both horizontal and vertical coordinates), -5^o to 5^o, and 1 to 1.2, respectively. All the lip sequences after the transformation are collected to form a new lip sequence dataset and the same experiments are performed on the new dataset.

Table 5 shows the identification and authentication results using all the four approaches after random affine transformations. From the table, it is observed that the performance of the two lip model based approaches (Luettin's and Cetingul's)

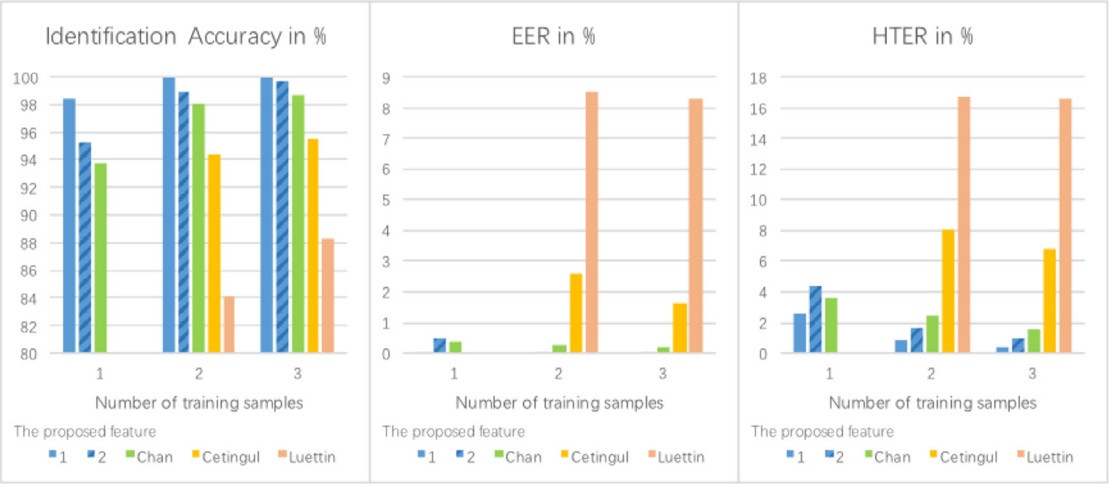


Fig. 10. The identification and authentication results in % with limited training data.

is quite similar to that obtained from the original dataset. It is because the lip model based features do not change much after affine transformation and are therefore insensitive to these variations. On the other hand, the lip region based features in Chan's approach are sensitive to the affine transformation, resulting in significant performance degradation (an 8% drop in accuracy). In our approach, the hierarchical structure and max-pooling performed on the sparse codes (especially for feature 1 with large dictionary size) of the densely overlapping cells make the proposed feature highly robust against translation, rotation and scaling and thus the proposed feature can achieve an accuracy improvement of 3.7% and an EER/HTER improvement of 1.97%/4.82% compared with those of the second best approach.

4.3.2. Identification & authentication results using limited training samples

During training, a speaker may become impatient when he/she is asked to repeat the utterance many times. Hence, approaches that work with a small number of training samples are preferred. In Fig. 10, we show the visual speaker identification and authentication results with limited number of training samples. Note that since HMM is a relatively complex statistical model, using limited training data is insufficient to optimize the model parameters [30]. In our experiment, HMM cannot even be initialized when using only one training sample and thus the results in Cetingul's and Luttin's approaches are not available. From the figure, it is observed that the proposed feature with alternative hierarchical structure always outperforms the other approaches with various selections of number of training samples. Moreover, when only one training sample is available, the proposed approach can still provide very reliable identification/authentication results (an accuracy of 98.39% and an HTER of 2.62%), which demonstrated the effectiveness of the proposed approach in case of very limited training data.

5. Conclusions and future works

In this paper, a visual speaker identification and authentication scheme based on joint spatiotemporal sparse coding and hierarchical pooling is presented. The lip sequence is first divided into a series of subsequences along the temporal dimension and each subsequence is considered as a spatiotemporal cube. The cube is then partitioned into many densely overlapping 3D spatiotemporal cells. Sparse coding is applied to describe the content of these cells, which resulted in a 3D SC matrix been generated for each subsequence. Max-pooling with a preset hierarchical structure is then performed on the 3D SC matrix to obtain the final feature for the subsequence, and the entire lip sequence is represented by a set of features corresponding to the subsequences in it. Experiment results have demonstrated that the proposed algorithm provides very high identification and authentication results and outperforms the other existing approaches we investigated. Moreover, we show that the proposed approach is robust against variations caused by speaker's positions and poses, and can even work well when only one training sequence is available.

Even though the proposed algorithm is demonstrated to achieve very reliable identification and authentication results, there are still issues to be solved. First, how does the performance change with the increase of number of speakers? Since the number of speakers in dataset [29] is relatively small, performance evaluations on a larger dataset will be interesting. Second, since the robust tests in Section 4.3.1 employ artificial data simulating translation, rotation and scaling, evaluations on real data with various poses will be the next step of our research. Third, most of the existing research in visual speaker identification and authentication are performed in a "prompt-text" scenario, where the speech content is constrained to be isolated word/sentence. How to extend it to a speech-independent, continuous speech scenario remains a more challenging task because in such scenario, the extracted features are required to be highly related to the speaker's talking style and independent to the speech content. This will be another important topic for our further research.

The availability of a good publicly available dataset will facilitate research in this field. Considering the limitations of current dataset [29], we plan to construct a new dataset for visual speaker identification and authentication in the near future. The new dataset is expected to be much bigger than the current dataset and will be released to the research community once completed.

Acknowledgement

The work described in this paper is fully supported by NSFC Fund (No. 61271319).

References

[1] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[2] L An, X.J. Chen, S.F. Yang, B Bhanu, Sparse representation matching for person re-identification, Inf. Sci. 355 (2016) 74–89.
[3] A. Bakry, E. Ahmed, Mkpls, Manifold kernel partial least squares for lipreading and speaker identification, in: Proc. 2013 IEEE Conference on Computer Vision and Pattern Recognition IEEE, 2013, pp. 684–691.
[4] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proc. ACM Workshop on Computational Learning Theory, 1992, pp. 144–152.
[5] C.C. Broun, X. Zhang, R.M. Mersereau, M. Clements, Automatic speechreading with application to speaker verification, in: Proc. 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002, p. 1.
[6] H.E. Cetingul, Y. Yemez, E. Erzin, A.M. Tekalp, Discriminative analysis of lip motion features for speaker identification and speech-reading, IEEE Trans. Image Process. 15 (10) (2006) 2879–2891.
[7] C.H. Chan, B. Goswami, J. Kittler, W. Christmas, Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication, IEEE Trans. Inf. Forensics. Secur. 7 (2) (2012) 602–612.
[8] G. Chetty, Biometric liveness detection based on cross modal fusion, in: Proc. 12th International Conference on Information Fusion, 2009, pp. 2255–2262.
[9] M. Choraś, The lip as a biometric, Pattern Anal. Appl. 13 (1) (2010) 105–112.
[10] R.W. Frischholz, U. Dieckmann, BiolD: a multimodal biometric identification system, Computer 33 (2) (2000) 64–68.
[11] Y.J. He, D.Y. Chen, G.L. Sun, J.Q. Han, Dictionary evaluation and optimization for sparse coding based speech processing, Inf. Sci. 310 (2015) 77–96.
[12] Y.F. Liu, C.Y. Lin, J.M. Guo, Impact of the lips for biometrics, image processing, IEEE Trans. Image Process. 21 (6) (2012) 3092–3101.
[13] X Liu, Y.M. Cheung, Learning multi-boosted HMMs for lip-password based speaker verification, IEEE Trans. Inf. Forensics. Secur. 9 (2) (2014) 233–246.
[14] F Liu, J.H. Tang, Y Song, Y Bi, S Yang, Local structure based multi-phase collaborative representation for face recognition with single sample per person, Inf. Sci. 346 (2016) 198–215.
[15] L Liu, P Fieguth, G Zhao, M Pietikäinen, D Hu, Extended local binary patterns for face recognition, Inf. Sci. 358 (2016) 56–72.
[16] J. Luettin, N. Thacker, S.W. Beet, Speaker identification by lipreading, in: Proc. the 4th International Conference on Spoken Language Processing (ICSLP"96), 1, 1996, pp. 62–65.
[17] J. Luettin, G Maître, Evaluation Protocol for the Extended M2VTS Database (XM2VTSDB), IDIAP, 1998.
[18] J.S. Mason, J.D. Brand, The role of dynamics in visual speech biometrics, in: Proc. 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 4, 2002 IV-4076-IV-4079.
[19] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 198–213.
[20] K. Messer, J. Matas, J. Kittler, J Luettin, G. Maitre, XM2VTSDB: The Extended M2VTS Database, in: Proc. The Second International Conference on Audio and Video-Based Biometric Person Authentication, 1999, p. 964.
[21] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609 6583.
[22] J. Ortega-Garcia, J. Bigun, D. Reynolds, J. Gonzalez-Rodriguez, Authentication gets personal with biometrics, IEEE Signal Process. Mag. 21 (2) (2004) 50–62.
[23] D. Peralta, M. Galar, I. Triguero, et al., A survey on fingerprint minutiae-based local matching for verification and identification: Taxonomy and experimental evaluation, Inf. Sci. 315 (2015) 67–87.
[24] S. Pigeon. The M2VTS Database, Laboratoire de Telecommunications et Teledection, Place du Levant, 1996.
[25] R. Rubinstein, M. Zibulevsky, M. Elad, Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit, CS Technion. 40 (8) (2008) 1–15.
[26] T. Serre, W. Lior, P. Tomaso, Object recognition with features inspired by visual cortex, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 994–1000.
[27] P. Singh, L. Vijay, M.S. Gaur, Speaker identification using optimal lip biometrics, in: Proc. the 5th IAPR International Conference on Biometrics (ICB), 2012, pp. 472–477.
[28] C.M. Travieso, J.R. Ticay-Rivas, J.C. Briceño, M. del Pozo-Bañosa, J.B. Alonsoa, Hand shape identification on multirange images, Inf. Sci. 275 (11) (2014) 45–56.
[29] S.L. Wang, W.C. Liew, Physiological and behavioral lip biometrics: a comprehensive study of their discriminative power, Pattern Recog. 45 (9) (2012) 3328–3335.
[30] S.L. Wang, W.H. Lau, S.H. Leung, An automatic lipreading system for spoken digits with limited training data, IEEE Trans. Circuits Syst. Video Technol. 18 (12) (2008) 1760–1765.
[31] T. Wark, S. Sridharan, V. Chandran, The use of temporal speech and lip information for multi-modal speaker identification via multi-stream HMMs, in: Proc. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000, pp. 2389–2392.
[32] Y. Xu, Q. Zhu, Z.Z. Fan, D. Zhang, J.X. Mi, Z.H. Lai, Using the idea of the sparse representation to perform coarse-to-fine face recognition, Inf. Sci. 238 (2013) 138–148.
[33] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1794–1801.
[34] G.Y. Zhao, M. Barnard, M. Pietikäinen, Lipreading with local spatiotemporal descriptors, IEEE Trans. Multimedia 11 (7) (2009) 1254–1265.
[35] Z.Y. Zhu, Q.H. He, X.H. Feng, Y.X. Li, Z.F. Wang, Liveness detection using time drift between lip movement and voice, in: Proc. 2013 International Conference on Machine Learning and Cybernetics (ICMLC), 2013, p. 2.