Visual Speaker Identification with Spatiotemporal Directional Features

Guoying Zhao and Matti Pietikäinen

Center for Machine Vision Research,
Department of Computer Science and Engineering,
P.O. Box 4500 FI-90014 University of Oulu, Finland
http://www.cse.oulu.fi/CMV

Abstract. In this paper, a novel local spatiotemporal directional descriptor is proposed for speaker identification by analyzing mouth movements. For this new descriptor, the directional local binary pattern features in three orthogonal planes are coded. In addition, besides sign features, magnitude information encoded as weight for the bins with the same sign value is developed to improve the discriminative ability. Moreover, decorrelation is exploited to remove the redundancy of features. Experimental results on the challenging XM2VTS database show the effectiveness of the proposed representation for this problem.

Keywords: speaker recognition, dynamic features, local binary pattern.

1 Introduction

Identification of an individual using physical attributes such as speech, for example, is of increasing importance in the field of security. Speaker identification is to automatically identify who is speaking on the basis of speech signals and lip movements concerning individual information. A comprehensive review of speech and speaker recognition can be found in [7]. Another review on audio-visual biometrics combining voice, visual speech and face is presented in [2].

Most of the early research focused on using audio signal only for this task [5,7,9,19]. This technique has good results when used in acoustically noise free environments but has limited success in busy environments such as offices, airports, train stations, and factory floors or in the presence of multiple talkers. In speech recognition community it is well known that visual information from lip or mouth movements provides additional speech information, which can lead to improved speaker recognition performance.

The dynamic visual features are suggested based on the shape and intensity of the lip region because changes in the mouth shape including the lips and tongue carry significant phoneme-discrimination information. An intuitive understanding of features related to lip movement is that different people have different articulatory styles when speaking the same utterance. For instance, when uttering the same phoneme or word, some speakers tend to widely open the mouth while others may only keep it slightly opened; some speakers tend

M. Kamel and A. Campilho (Eds.): ICIAR 2013, LNCS 7950, pp. 1–10, 2013.
© Springer-Verlag Berlin Heidelberg 2013

to move the lips in a specific direction, and for some speakers, the teeth are always visible when uttering [14]. A desirable method is to extract features from the gray-level data directly rather than extract geometric features which commonly require accurate and reliable facial and lip points detection and tracking. The type of features are based on observing the whole mouth Region-of-Interest (ROI) as visual informative about the spoken utterance. The feature vectors are computed using all the video pixels within the ROI. The proposed approaches include Discrete Cosine Transform (DCT), Principal Component Analysis (PCA), and more recently some dynamic features, e.g. EdgeMap_LBP[18] and LOCP[4]. Zhao et al. proposed to combine spatiotemporal dynamic texture features of local binary patterns extracted from localized mouth regions and structural edge map features [18] extracted from the image frames for representing appearance characteristics (EdgeMap_LBP). Chan et al. used local ordinal contrast pattern with three orthogonal planes (LOCP-TOP) to represent both the appearance and dynamics features in visual speech [4]. Both approaches gave promising results. Even though they took spatiotemporal planes into account, 1) they did not consider the directional features; 2) it only used binary code (sign) information; 3) the correlation of pixels was not eliminated.

In this paper, we focus on the speaker recognition using only visual information. A novel dynamic descriptor called local spatiotemporal directional features (LSDF) is proposed to address the above-mentioned problems, which 1) considers the changes of six directions in three orthogonal planes, providing more details concerning spatiotemporal transition; 2) combines sign information and magnitude information where magnitude is utilized as weights for the pixels with same sign value; and 3) utilizes decorrelation to eliminate the correlation of pixels in spatiotemporal domain to achieve more discriminative representation.

2 Local Spatiotemporal Directional Features

Our features are built on the basis of the local binary pattern (LBP) operator which is a gray-scale invariant texture primitive statistic that has shown excellent performance in the classification of various kinds of textures [12] and face recognition [1], for example. For each pixel in an image, a binary code (sign features) is produced by thresholding its neighborhood with the value of the center pixel.

Later, a method for temporal texture recognition using spatiotemporal local binary patterns extracted from three orthogonal planes (LBP-TOP) was proposed [16]. With this approach the ordinary LBP for static images was extended to spatiotemporal domain and region-concatenated descriptors using LBP-TOP features were developed for facial expression recognition [16] and visual speech recognition [17]. Furthermore, the LBP features extracted from XT and YT plane are utilized to represent the motion of utterance by catching the transition information of micro textures for speaker identification in [18].

To extract more detailed spatiotemporal directional changes and combine both sign and magnitude information, novel Local Spatiotemporal Directional Features (LSTD) are presented in the following subsections.

2.1 Encoding Spatiotemporal Directional Changes

In XY plane, there are mainly X and Y directions. Likewise, in XT and YT planes, there are X (Y) and T directions. How the information changes in each direction would be important for mouth movement progress during speech. So here we propose local spatiotemporal directional features (LSTD), as shown in Fig. 1, where the sign information like traditional LBP is encoded and denoted as $SLSTD$.

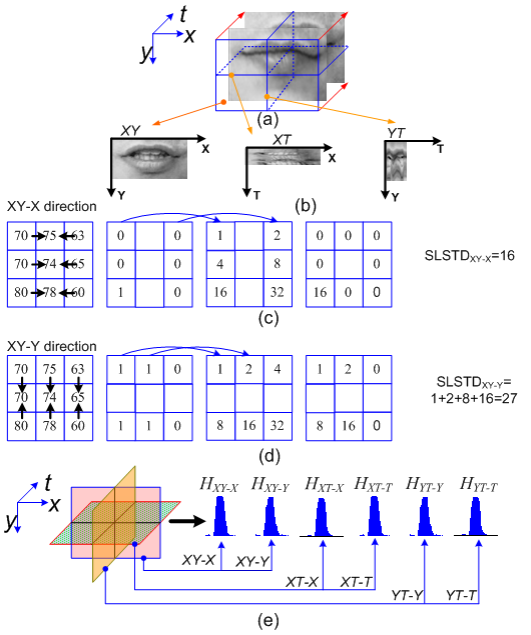

Fig. 1. Illustration of directional coding of sign information

Fig. 1 (a) shows the three planes in a speech video, where each plane contains two directions. Fig. 1 (b) shows the images from three planes. For each pixel, we can obtain a 3×3 neighboring area around it in each plane, which can be marked as P_0, P_1, $\cdots$, P_8 from topleft to bottomright. Then we could calculate the directional features as shown in Fig. 1 (c) and (d). For each row (column), a binary code is produced by thresholding its neighborhood with the value of the center pixel in this row (column). Fig. 1 (c) shows the calculation for the X direction and Fig. 1 (d) for Y direction. In this way, the information in spatiotemporal direction can be represented. Similarly, the X direction and T direction in XT plane and Y direction and T direction in YT plane can be obtained. Finally, the resulting histograms in each direction can be concatenated as a spatiotemporal directional features for a speech video, shown in Fig. 1 (e).

The whole feature representation contains three steps: sampling, transformation and quantization.

First is sampling. For each central pixel, we can get eight neighboring points and in total nine points in the calculation. The sampling distance of each direction can be changed. As Fig. 2 shows, in X direction, the sampling radius is three and in Y direction, the radius is one. So we can set R_x, R_y, and R_t with different values to represent the sampling radii in three directions. Then we could get P_0, P_1,$\cdots$, P_8 corresponding to, e.g. $I(x_c - R_x, y_c - R_y, t_c)$, $I(x_c, y_c - R_y, t_c)$, $I(x_c + R_x, y_c - R_y, t_c)$, $I(x_c - R_x, y_c, t_c)$, $I(x_c, y_c, t_c)$, $I(x_c + R_x, y_c, t_c)$, $I(x_c - R_x, y_c + R_y, t_c)$, $I(x_c, y_c + R_y, t_c)$, $I(x_c + R_x, y_c + R_y, t_c)$ in XY plane, where R_x, R_y are sampling radii in X direction and Y direction, respectively. Similarly we can have R_t as sampling radius in T direction of XT and YT planes. From this area the local spatial-temporal feature is calculated for central pixel $P_4 = I(x_c, y_c, t_c)$. It yields flexible sampling and multiresolution possible.

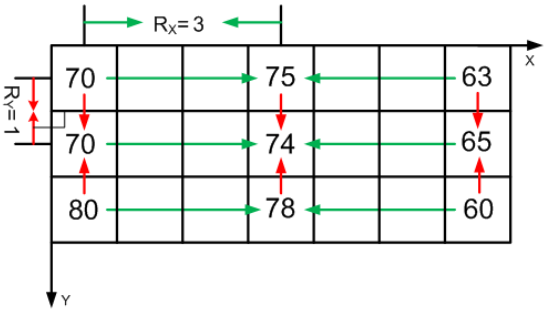

Fig. 2. Sampling in XY plane with radius three in X direction and radius one in Y direction

Second step is transformation. We represent it in a more general way:

$$F(x) = Wf(x) \tag{1}$$

W is 6-by-R^2 ($R = 3$ here) transformation matrix and $f(x) = [P_0, P_1, \cdots, P_8]^T$ is another vector containing all R^2 image pixels which are neighboring points to each central pixel $x = [x_c, y_c, t_c]$.

We here consider the changes in each direction, so the difference of its neighboring point against the middle point in each direction is calculated. Thus, for $XY - X$ ($XT - X$ or $YT - T$ direction,

$$W = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

and for $XY-Y$ ($XT-T$ or $YT-Y$) direction $W = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}$

Third step is quantization.

$$SLSTD_d(x_c, y_c, t_c) = \sum_{n=0}^{5} p_n 2^n, \tag{2}$$

calculates the sign feature for the pixel located at (x_c, y_c, t_c) for direction d, $d \in \{XY-X, XY-T, XT-X, XT-T, YT-Y, YT-T\}$.

$$p_n = \begin{cases} 1, if\ f_n \geq 0; \\ 0, otherwise. \end{cases} \tag{3}$$

where f_n is the nth component of $F(x)$. The quantized coefficients are represented as integer values between $0-63$ using binary coding.

Then, we can calculate the $SLSTD$ features for all pixels and get the sign histogram of LSDF ($HisSLSDF$):

$$HisSLSDF_d(k) = 1/(MNT) \sum_{x=R_x}^{M-R_x} \sum_{y=R_y}^{N-R_y} \sum_{t=R_t}^{T-R_t} I(SLSTD_d(x,y,t) = k) \quad k \in [0, K-1]; \tag{4}$$

where M and N are width and height of image, T is the utterance length, $K = 2^6$ is the number of different labels produced by the $LSTD$ operator, and

$$I\{A\} = \begin{cases} 1, if\ A\ is\ true; \\ 0, if\ A\ is\ false. \end{cases} \tag{5}$$

2.2 Combining Sign and Magnitude Information

Besides sign information, magnitude is also very useful for description which represents the contrast information, like in CLBP [8].

The differnce from sign coding is in the quantization step: Magnitude is calculated as mean of the absolute intensity difference for a certain direction, as shown in Fig. 3, which shows the value of the magnitude for $XY-X$ direction ($MLSTD_{XY-X}$) corresponding to Fig. 1 (c).

$$MLSTD_d(x_c, y_c, t_c) = 1/6 \sum_{n=0}^{5} |f_n|, \tag{6}$$

calculates the magnitude feature for the pixel located at (x_c, y_c, t_c) for direction d, where f_n is again the nth component of $F(x)$ in Eq. (1).

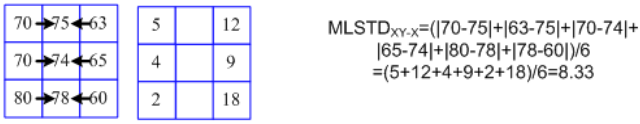

Fig. 3. Magnitude features in one direction

After obtaining $MLSTD$ for each pixel, the magnitude is used as an adaptive weight to adjust the contribution of the $SLSTD$ code in histogram calculation. The magnitude histogram of local spatiotemporal directional feature (HisMLSDF) is computed as:

$$HisMLSDF_d(k) = 1/(MNT)\sum_{x=R_x}^{M-R_x}\sum_{y=R_y}^{N-R_y}\sum_{t=R_t}^{T-R_t} w(SLSTD_d(x,y,t),k)$$
$$k \in [0,K]; d \in \{XY-X, XY-T, XT-X, XT-T, YT-Y, YT-T\}. \tag{7}$$

with

$$w(SLSTD_d(x,y,t),k) = \begin{cases} WLSTD_d(x,y,t), if\ SLSTD_d(x,y,t) = k; \\ 0, otherwise. \end{cases} \tag{8}$$

where d denotes the index for spatiotemporal direction, which can be $XY-X$, $XY-Y$, $XT-X$, $XT-T$, $YT-Y$ and $YT-T$.

Here the magnitude is encoded in different way to CLBP [8]. 1) Magnitude is not encoded into binary pattern, but continuous values used as weight for the pixels with the same sign feature. So in its design, it has been well related to the sign part; 2) the mean value of the magnitude difference is directly used, without thresholding according to the global mean of magnitude difference like CLBP, which makes our coding much more computationally efficient.

To furthur combine sign and magnitude features, $HisSLSDF$ and $HisMLSDF$ are concatenated as final description.

2.3 Decorrelation

If the samples to be quantized are statistically independent, the information is maximally preserved in scalar quantization. So before quantization the coefficients should be decorrelated [13].

Assuming Gaussian distribution, independence can be achieved using a whitening transform:

$$G(x) = V^T F(x) = V^T(Wf(x)) \tag{9}$$

where $V$is an orthonormal matrix derived from the singular value decomposition (SVD) of the covariance matrix of the transform coefficient vector $F(x)$ and V can be solved in advance. For details, please refer to [13].

Then we use $G(x)$ for independent sign and magnitude feature calculation:

$$\widetilde{SLSTD_d} = \sum_{n=0}^{5} q_n 2^n, \tag{10}$$

$$q_n = \begin{cases} 1, if\ g_n \geq 0; \\ 0, otherwise. \end{cases} \tag{11}$$

where g_n is the nth component of $G(x)$ after decorrelation. The quantized coefficients are represented as integer values between $0 - 63$ using binary coding.

In similar way,

$$\widetilde{MLSTD}_d = 1/6 \sum_{n=0}^{5} |g_n|, \tag{12}$$

calculates the decorrlated magnitude feature for the pixel located at (x_c, y_c, t_c) for direction d.

Finally, two histograms $His\widetilde{SLSTD}$ and $His\widetilde{MLSTD}$ after decorrelation of these $\widetilde{SLSTD}$ and $\widetilde{MLSTD}$ values from all image positions are composed and concantenated (as Eqs. (4) and (7)), as a 64×2-dimensional feature vector in classification.

3 Experiments and Analysis

Experiments were carried out for evaluating the performance of the proposed method on challenging XM2VTS Audio-Visual database. The database consists of video data recorded from 295 subjects in four sessions, spaced monthly. The first recording per session of the sentence "Joe took fathers green shoe bench out" was used for this research.

Boosted Haar features [15] are used for automatic coarse face detection and 2D Cascaded AdaBoost [11] is applied for localizing eyes in the detected faces. Because the face images in the database are of good quality and almost all of them are frontal faces, detection of faces and eyes is quite easy. The positions of the two eyes in the first frame of each sequence were given by the eye detector automatically and then these positions were used to determine the fine facial area and localize the mouth region using pre-defined ratio parameters [17] for the whole sequence. Fig. 4 demonstrats the obtained mouth region images from XM2VTS database. Usually lots of preprocessing [6] was done before the feature extraction and classification to make the task a bit easier, including that 1) the mouth ROI was identified manually, whereas it was localized automaticlly in our work; 2) the start and end of some sentences were clipped, mouth ROI was histogram equalized and the mean pixel value was subracted for dealing with varying lighting conditions across sessions, whereas no such preprocessing was done in our experiment. So our experimental conditions are much more difficult. Fig. 4 shows the mouth images from four sessions of the same person (top row) and mouth images from the last session of four different persons (bottow row). Because there is no preprocessing after automatic localization of the mouth area, the translations, rotations and scale variations can be seen from the first row and illumination or skin changes in the second row, which makes it more demanding for the discriminative capability of features.

Fig. 4. Localized mouth images from XM2VTS database

For feature extraction, each utterance sequence is divided into $1 \times 5 \times 1$ cuboids and the sampling radii are set as $R_x = R_y = R_t = 3$. We implemented all the other methods under comparison using the same cropped mouth images and test setups for a fair comparison.

In our experiments, the same experimental setup to [6] is utilized. The probe sequences used for testing were obtained from the fourth session. The galleries for training were formed from the first three sessions according to Table 1. Five trials were constructed to test how the performance of the proposed method varied when 1) the time difference between the gallery and the probe set varied between one and three months and 2) multiple sessions were used to form the gallery. When there are multiple traning samples, we get one additional sample by averaging all traning samples.

Table 1. Speaker identification results (%) on the XM2VTS database

Trial #	Train	Test	$\widetilde{HisSLSTD} + \widetilde{HisMLSTD}$	$\widetilde{HisSLSTD}$	$HisSLSTD$
1	1	4	**66.44**	63.05	52.88
2	2	4	72.20	66.10	63.05
3	3	4	**77.29**	72.54	67.46
4	2,3	4	**86.44**	84.07	78.31
5	1,2,3	4	**88.14**	86.78	79.66
Trial #	Train	Test	LOCP-TOP [4]	LBP-TOP [16]	EdgeMap_LBP [18]
1	1	4	62.03	57.97	50.85
2	2	4	**73.22**	67.46	64.07
3	3	4	74.24	69.15	65.42
4	2,3	4	84.07	81.02	78.31
5	1,2,3	4	86.44	84.07	79.66

For first three trials, there is only one training sample which was captured one (two or three) month ago for each subject. For the fourth and fifth trials, there are only two and three training samples for each subject. It makes the classification very challenging. Because there are only so few training samples for each class, the testing samples are classified or verified according to their difference with respect to the class using the k nearest neighbor method ($k = 1$). The dissimilarity between a sample and a model distribution is measured using the $L1$ distance.

From Table 1, we can see that our method achieved much better results than other spatiotemporal features including LOCP-TOP, LBP-TOP and EdgeMap_LBP features. Especially in the first trial, where the time difference between training sample and test sample is three months, our results are 4.41%, 8.47% and 15.59% higher than that from LOCP-TOP, LBP-TOP and EdgeMap_LBP features. When using multiple sessions in training, the recognition result (86.44% and 88.14%) for two and three tranining samples is better than using only one training sample. The first column shows the results of concatenating sign and magnitude features which is much better than the second column using only sign features, indicating the effective supplemental inforamtion from magnitude. The third column shows the results of sign features without decorrelation, which is inferior to the second column with decorrelation.

4 Discussion and Conclusion

A novel local spatiotemporal directional descriptor for speaker recognition using sole visual information was proposed, combining both sign and magnitude components of dynamic texture features for motion and appearance representation. The movements of mouth regions are described using local binary patterns and intensity contrast from six directions in three orthogonal planes. Decorrelation technique is utilized to get a more independent and discriminative representation.

Very challenging evaluation on XM2VTS database collected from 295 subjects is carried out: only a few training samples (one to three) are available and they were collected in different time sessions. Experiments show that the proposed method outperforms the state of the art.

In the future, we would evaluate the effect of parameters on bigger and more practical databases, and make extensive experiments, comparison and analysis. Moreover, it is of interest to combine visual and audio information to promote speaker recognition, and to apply our methodology to natural human-robot interaction in a smart environment where we need to cope with movement of the speaker's head during operation.

Acknowledgments. The financial support provided by the Academy of Finland and Infotech Oulu is gratefully acknowledged.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 28(12), 2037–2041 (2006)
2. Aleksic, P., Katsaggelos, A.: Audio-visual Biometrics. Proceedings of the IEEE 94, 2025–2044 (2006)
3. Cetingul, H., Yemez, Y., Erzin, E., Tekalp, A.: Discriminative Lip-motion Features for Biometric Speaker Identification. In: Proc. of ICIP (2004)

4. Chan, C.H., Goswami, B., Kittler, J., Christmas, W.: Local Ordinal Contrast Pattern Histograms for Spatiotemporal, Lip-Based Speaker Authentication. IEEE Trans. on Information Forensics and Security 7(2), 602–612 (2012)
5. Faundez-Zanuy, M., Satue-Villar, A.: Speaker Recognition Experiments on A Bilingual Database. In: Proc. of EUSIPCO (2006)
6. Fox, N., Gross, R., Chazal, P., Cohn, J., Reilly, R.: Person Identification Using Automatic Integration of Speech, Lip and Face Experts. In: Proc. of the ACM SIGMM Workshop on Biometrics Methods and Applications, pp. 25–32 (2003)
7. Furui, S.: Fifty Years of Progress in Speech and Speaker Recognition. Acoustical Society of America Journal 116(4), 2497–2498 (2004)
8. Guo, Z., Zhang, L., Zhang, D.: A Completed Modeling of Local Binary Pattern Operator for Texture Classification. IEEE Transactions on Image Processing 19(6), 1657–1663 (2010)
9. Liu, M., Zhang, Z., Hasegawa-Johnson, M., Huang, T.: Exploring Discriminative Learning for Text-Independent Speaker Recognition. In: Proc. of ICME, pp. 56–59 (2007)
10. Luettin, J., Thacher, N., Beet, S.: Speaker Identification by Lipreading. In: Proc. of International Conference on Spoken Language Proceedings, pp. 62–64 (1996)
11. Niu, Z., Shan, S., Yan, S., Chen, X., Gao, W.: 2d Cascaded Adaboost for Eye Localization. In: Proc. of ICPR, pp. 1216–1219 (2006)
12. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE PAMI 24(7), 971–987 (2002)
13. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) ICISP 2008. LNCS, vol. 5099, pp. 236–243. Springer, Heidelberg (2008)
14. Ouyang, H., Lee, T.: A New Lip Feature Representation Method for Video-based Bimodal Authentication. In: NICTA-HCSNet Multimodal User Interaction Workshop (2006)
15. Viola, P., Jones, M.: Rapid Object Detection Using A Boosted Cascade of Simple Features. In: Proc. of CVPR, pp. 511–518 (2001)
16. Zhao, G., Pietikäinen, M.: Dynamic Texture Recognition Using Local Binary Patterns with An Application to Facial Expressions. IEEE PAMI 29(6), 915–928 (2007)
17. Zhao, G., Barnard, M., Pietikäinen, M.: Lipreading with Local Spatiotemporal Descriptors. IEEE Transactions on Multimedia 11(7), 1254–1265 (2009)
18. Zhao, G., Huang, X., Gizatdinova, Y., Pietikäinen, M.: Combining Dynamic Texture and Structural Features for Speaker Identification. In: ACM Multimedia 2010 Workshop on Multimedia in Forensics, Security and Intelligence, pp. 93–98 (2010)
19. Zhou, X., Fu, Y., Liu, M., Hasegawa-Johnson, M., Huang, T.: Robust Analysis and Weighing on MFCC Components for Speech Recognition and Speaker Identification. In: Proc. of ICME, pp. 188–191 (2007)