IJCSI International Journal of Computer Science Issues, Volume 14, Issue 4, July 2017
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.20943/01201704.17

Person Identification From Text Independent Lip Movement Using the Longest Matching Segment Method

Paul C. Brown, Ji Ming, Daryl Stewart

Institute of ECIT, Electronics and Computer Engineering Cluster, Queen′s University Belfast, Belfast BT7 1NN, UK

Abstract

The use of lipreading as a standalone modality for biometric classification continues to gain ground but is still presented with several real world challenges. The paper presents a novel form of video temporal modelling using the Longest Matching Segment (LMS) method on a given baseline training model. LMS uses a Vector Quantization (VQ) model to encode full training video dynamics by mapping it to a frame sequence of maximum likelihood codewords. The model is applied to person identification from text independent lip movement on segmented test sets of the CMU-PIE, VidTIMIT and XM2VTS talking datasets and identification is based on the class with the longest matching segment. The results show that LMS improves the conventional VQ models especially when combined with dynamic delta features. Combined with magnitude 2D-FFT (Mag-2D-FFT) features, the system delivers comparable accuracies to full face recognition.

Keywords— *Lip Movement, Vector Quantisation, Longest Matching Segment, Person Identification*

1. Introduction

From the beginnings in the latter half of the 20th century lipreading systems have made significant progress in theoretical research and practical implementation. At first it was as an important aid to audio speech recognition [1, 2, 3]. Over time it became more established in human and speech recognition. Now machine-based lipreading systems are significantly better than the human equivalent [4], and there is a wide range of applications in public security, banking, medicine, and law enforcement. One such is LipVerify, an easy-to-use biometric authentication solution developed by Liopa that validates user identity and performs visual phrase recognition via any smartphone, tablet, laptop or desktop from viseme profiles [5] The standalone use of lip motion as a human biological trait began near the turn of the 21st century. Kittler et al in 1997 showcased the importance of lip boundary geometric features in identity recognition [6]. Yamamoto et al in 1998 showed significant performances with lip movement captured from sensors around the mouth [7]. CetingÃijl et al in 2004 observed improved speaker identification with speaker dependent lip motion features [8, 9]. CetingÃijl et al in 2006 further showed that explicit lip motion information was useful for speaker identification and speech-reading [10]. Faraj et al in 2009 achieved 80% person recognition accuracy with Support Vector Machines (SVM), a Radial Basis Function (RBF) kernel and optical flow features on the full XM2VTS dataset [11]. Bakry et al in 2013 presents a Manifold Kernel Partial Least Squares (KPLS) for lipreading and speaker identification on the AVLetters and OuluVS databases [12].

On the other hand the inherent low quality of real world video is prone to weakly constrained factors and limited cooperation leaving room for more robust improvements. The encoding of more accurate temporal dynamics is a significant contributor to recognition accuracy. It can be grouped as fixed or variable, short or long-range, dynamic features captured during feature extraction, or dynamic models built into the training model. Scholars have adopted several means of encoding the different forms for person identification. The most common form of dynamic features are fixed short range delta (velocity) and delta-delta (acceleration) features [13], but there are other types. Edwards et al improved person identification by actively learning how individual faces vary through in video [14]. FrÃűba et al in 2000 uses lip motion features derived from the inter-frame optical flow power spectrum to evaluate sensor calibration in a biometric person recognition framework [15]. Lee et al applied a pattern classification algorithm called LDG (Locality Discriminant Graph) to the temporal filtering of visual speech in 2007 [16]. Zhang et al proposed a two-stage, space-time discriminant analysis to extract lip motion features in 2012 [17].

Hidden Markov Models (HMMs) are among the most widely used dynamic models for a visual-only recognition systems due to its strong time sequence ability [18]. Luettin et al obtained promising results in 1996, using spatial and temporal HMM analysis with lip contour and shape-based features on the Tulips database [19]. Jourlin et al extended this task to the M2VTS database using 37 speakers in 1997 [20]. Liu et al in 2003 used Adaptive HMMs to learn temporal statistics of a subject model for video-based face recognition [21]. Shipilova et al in 2006 presented a person recognition problem using Luettin's setup with HMMs and GMMs and achieved better lip-based recognition results than face-based or voice-based methods [22], a work later improved by Faraj in [23]. Seymour et al in 2008 used HMMs with DCT and DWT fea-

tures for visual-only identification [13]. In 2012 Mehraj et al presented a comparative review of various lip based biometric techniques using HMMs trained with DCT and lip contour features on the VidTIMIT corpus [24].
Deep learning has delivered step improvements over HMMs in lipreading classification. Techniques such as Deep Autoencoders (DAE) [25] and Deep Boltzmann Machines (DBMs) [26] were used between 2011-12 in cross-modality unsupervised feature learning to improve classification performance on the AVLetters and CUAVE databases.

2. Modelling the Video Temporal Dynamics for Person Identification

LMS has been successfully used in statistical training to model long range audio speech dynamics. Jafari et al used LMS with frame-based GMMs for text-independent speech recognition [27]. Srinivasan et al used the same setup to distinguish clean speech corrupted with nonstationary noise [28]. LMS has been combined with HMMs for speechreading [29], where it is argued to be an improvement opportunity with Deep Neural Networks (DNNs). However a text independent system can be suitably modelled as a collection of discrete frame templates with no time sequence information, or a VQ model. Since LMS is under study, the baseline model is fixed throughout, and VQ is proven in face and speech recognition [30, 31, 32, 33] this paper integrates LMS with a baseline or conventionall VQ model. The ease of computation makes VQ a suitable framework to evaluate long-range LMS dynamics against conventiona VQ training models with static and dynamic delta features in person identification.
Let lip motion dynamics be the time series of variations locked between frame events, containing the unique temporal discrimination that can be interpreted as biometric signatures. For lipreading person identification fixed short range dynamic features and a baseline VQ training model are applied to the formulation of long-range inter frame dynamics using the LMS method.

2.1. Training the Lipreading Image Sequence

Given static feature f_i of lip ROI frame i, delta features Δf_i and acceleration features $\Delta\Delta f_i$ facilitate the mathematical removal of constant inter-frame bias via the regression formula [34].
The conventional VQ model is trained with static and dynamic features using the Linde-Buzo-Gray (LBG) technique [35], with Euclidean distance replaced by Cosine Similarity. VQ clusters I_X-frame training video $X = \{x_i : i = 1, 2, ..., I_X\}$ of each class m into a codebook of N finite vector mean codewords $\Phi_m = [\phi_{n,m} w_n :: n = 1, ..., N, N <= I_X]$. The VQ model for the entire training set Φ is the set of M codebooks such that $\Phi = \{\Phi_m, m = 1, ..., M\}$.
Video-based LMS training considers a lipreading database of M classes containing I_X-frame training video X, and M N-component baseline training models Φ_m. LMS models the full dynamics of each video X per class m such that any segment of any length up to the complete video can be used as a whole unit for identification, and the same class produces a greater number, length, and similarity of matching segments over all possible segment lengths than different classes. LMS maps frame x_i to the codeword $\phi_{n,m}$ with the maximum cosine similarity, resulting in a time series model of the maximum-similarity VQ components $\phi_{n_{X,i},m}$. The resulting frame sequence model can be formulated with the following time series of indices:

$$(n_X, m) = ([(n_{X,i}, m), i = 1, 2, ..., I_X] \tag{1}$$

The indices $(n_{X,i}, m)$ denote maximum likelihood codeword sequence of lip ROI templates. This model offers both a smooth representation of the short-time spectra and a video-length representation of the temporal dynamics.

2.2. Identifying the Longest Matching Segments

Given T_Y-frame test video $Y = \{y_t : t = 1, 2, ..., T_Y\}$ where y_t is a frame at time t, the likelihood of test frame y_t associated with training frame x_i is based on cosine similarity measure $\psi(x_i, y_t)$ defined as:

$$\psi(x_i, y_t) = \frac{x_i . y_t}{|x_i| . |y_t|} \tag{2}$$

Rewriting $\psi(x_i, y_t)$ in exponential form proportional to equation 2, by raising it to a large positive scalar λ gives:

$$\zeta(x_i, y_t) = \lambda^{\psi(x_i, y_t)} \tag{3}$$

The function $\zeta(x_i, y_t)$ takes the form of the likelihood of x_i associated with y_t, similar to probability distribution $p(x_i|m)$ [36].
Let Y contain segment $Y_{t:\tau} = \{Y_\epsilon : \epsilon = t, t+1, ..., \tau\}$, and frame sequence model (n_X, m) contain a same length training segment indexed by $(n_{X,u:v}, m) = [(n_{X,i}, m) : i =, u, u+1, ..., v]$. LMS performs identification by seeking the longest matching segment in all frame sequence models $(n_{X,i}, m)$ of all speakers, forcing a match to the codeword sequences of training video X. Assuming independence between frames, the likelihood of $Y_{t:\tau}$ associated with $(n_{X,u:v}, m)$ is the likelihood function $p(Y_{t:\tau}|n_{X,u:v}, m)$ as a function of exponential similarity $\zeta(Y_{t:\tau}, X_{n_{X,u:v},m})$ or:

$$\begin{aligned} p(Y_{t:\tau}|n_{X,u:v}, m) &= \zeta(Y_{t:\tau}, X_{n_{X,u:v},m}) \\ &= \prod_{\epsilon=t}^{\tau} \lambda^{(\psi(Y_\epsilon, X_{n_{X,\epsilon},m}))} \end{aligned} \tag{4}$$

where i_ϵ is the most-likely frame map path between test frames Y_ϵ and the training frame sequence model (n_{X,i_ϵ}, m) so that $i_t = u$ and $i_\tau = v$.

Assuming equal prior similarity ρ for all video segments S that may match $Y_{t:\tau}$, the similarity between the test and training segments is measured by the following posterior probability formulation $P(n_{X,u:v}, m|Y_{t:\tau})$ that derives the longest matching segment:

$$P(n_{X,u:v}, m|Y_{t:\tau}) = \frac{p(Y_{t:\tau}|n_{X,u:v}, m)\rho}{p(Y_{t:\tau})}$$
$$= \frac{p(Y_{t:\tau}|n_{X,u:v}, m)\rho}{\sum\limits_{S\in Database}(p(Y_{t:\tau}|S)\rho) + \sum\limits_{S\notin Database}(p(Y_{t:\tau}|S)\rho)}$$
$$= \frac{p(Y_{t:\tau}|n_{X,u:v}, m)}{\sum_{m'}\sum_{X'}\sum_{u',v'} p(Y_{t:\tau}|n_{X',u':v'}, m') + p(Y_{t:\tau}|\xi_{t:\tau})} \quad (5)$$

Applying the exponential similarity $\zeta(Y_{t:\tau}, X_{n_{X,u:v},m})$ to Equation 5 creates a posterior similarity formulation, written as:

$$P(n_{X,u:v}, m|Y_{t:\tau}) = \frac{\zeta(Y_{t:\tau}, X_{n_{X,u:v},m})\rho}{\zeta(Y_{t:\tau})}$$
$$= \frac{\zeta(Y_{t:\tau}, X_{n_{X,u:v},m})\rho}{\sum\limits_{S\in Database}(\zeta(Y_{t:\tau}, S)\rho) + \sum\limits_{S\notin Database}(p(Y_{t:\tau}|S)\rho)}$$
$$= \frac{\zeta(Y_{t:\tau}, X_{n_{X,u:v},m})}{\sum_{m'}\sum_{X'}(\sum_{u',v'}\zeta(Y_{t:\tau}, X_{n_{X',u':v'},m'}) + p(Y_{t:\tau}|\xi_{t:\tau}))} \quad (6)$$

The denominator first term is the sum of the similarity of all training segments from all locations in all videos of all the classes that are likely to match $Y_{t:\tau}$. The denominator second term $(p(Y_{t:\tau}|\xi_{t:\tau}))$ is the Universal Segment Model (USM) representing the similarity that $Y_{t,\tau}$ is not seen in any corpus segments in the training model. It is calculated as the frame-by-frame product of either the sum of the maximum class component similarities normalized over all classes, or the sum of the class component similarities normalized over all components, or:

$$p(Y_{t:\tau}|\xi_{t:\tau}) = \begin{cases} \prod_{\epsilon=t}^{\tau} \frac{\sum\limits_{m=1}^{M} w_m(\max\limits_{1\le n\le N}(\zeta(Y_\epsilon, \phi_{n,m})))}{M} \\ \prod_{\epsilon=t}^{\tau} \frac{\sum\limits_{m=1}^{M} w_m(\sum\limits_{n=1}^{N}(\zeta(Y_\epsilon, \phi_{n,m})))}{MN}. \end{cases} \quad (7)$$

A more robust formulation takes into account both the forward and reverse context of the test video. The forward VQ-based LMS is the posterior similarity formulation applied to test segment $Y_{t:\tau}$ and training segment $(n_{X,u:v}, m)$ in their exact frame sequences as shown in Equation 6. The reverse VQ-based LMS represents the posterior similarity formulation of test segment $Y_{\tau_r:t}$, and training segment $(n_{X,v_r:u}, m)$ in their reverse frame sequences, where $u - v_r = v - u$ and $t - \tau_r = \tau - t$, given by:

$$P(n_{X,v_r:u}, m|Y_{\tau_r:t}) = \frac{\zeta(Y_{\tau_r:t}, X_{n_{X,v_r:u},m})\rho}{\zeta(Y_{\tau_r:t})}$$
$$= \frac{\zeta(Y_{\tau_r:t}, X_{n_{X,v_r:u},m})\rho}{\sum\limits_{S\in Database}(\zeta(Y_{\tau_r:t}, S)\rho) + \sum\limits_{S\notin Database}(\zeta(Y_{\tau_r:t}|S)\rho)}$$
$$= \frac{\zeta(Y_{\tau_r:t}, X_{n_{X,v_r:u},m})}{\sum_{m'}\sum_{X'}(\sum_{v'_r,u'}\zeta(Y_{\tau_r:t}, X_{n_{X',v'_r:u'},m}) + p(Y_{\tau_r:t}|\xi_{\tau_r:t}))}. \quad (8)$$

Central LMS is computed over segment lengths $3, 5, 7, ..., 2t+1, ..., T$ by the following:

$$P_{central}(n_{X,i_\epsilon}, m|Y_\epsilon) = \begin{cases} P(n_{X,u:v}, m|Y_{t:\tau}). & \\ \quad P(n_{X,v_r:u}, m|Y_{\tau_r:t}) & \text{if } \tau \le T \text{ and } \tau_r \ge 1 \\ P(n_{X,u:v}, m|Y_{t:\tau}) & \text{if } \tau_r < 1 \\ P(n_{X,v_r:u}, m|Y_{\tau_r:t}) & \text{if } \tau > T. \end{cases} \quad (9)$$

2.3. Recognition Formulation using the Longest Matching Segment Method

For a conventional VQ model, assuming the codewords $\phi_{n,m}$ are independent, person identification is based on the class with the maximum product of the best single-frame cosine similarity ψ or exponential cosine similarity ζ between the T-frame test video and codeword, defined as:

$$Score_{VQ} = \begin{cases} arg \max\limits_{1\le m\le M}\left(\prod_{t=1}^{T}\max\limits_{1\le n\le N}(\psi(y_t, \phi_{n,m}))\right) \\ arg \max\limits_{1\le m\le M}\left(\prod_{t=1}^{T}\max\limits_{1\le n\le N}(\zeta(y_t, \phi_{n,m}))\right). \end{cases} \quad (10)$$

The LMS method uses the central posterior similarity formulation $P_{central}$ for classification. Person identification is based on the class m with the longest matching segment [28], or the class m containing segment $n_{X,u:v}, m$ with the maximum central posterior similarity score $P_{central}$. This score is computed from equations 6 and 8, over all possible segment lengths $(3, 5, 7, , ..., 2t+1, ..., \min(I_X, T_Y))$, based on the exponential likelihood function of Equation 4, or:

$$Score_{LMS} = \max_{1\le m\le M}\left(\max_{1\le\epsilon\le\min(I_X,T_Y)}(P_{central}(n_{X,i_\epsilon}, m|Y_\epsilon))\right). \quad (11)$$

If the selected class m is the same as that of the test video it scores a match. Overall system accuracy is achieved by collating all correct matches as a percentage of the total number of test cases per test set.

3. Person Identification Experiments

Multiple text-independent tasks are carried out to evaluate the impact of LMS temporal dynamics on lipreading person identification. Test results show there is very little to

learn from full test videos in the chosen CMU PIE [37], VidTIMIT [38] and the XM2VTS [39] talking datasets, so small video segments are used instead to simulate limited training data. Each dataset is preprocessed into cropped and normalized grayscale 48×64 lip ROI frames, further downsized to 12×16-pixels. Each frame is decoded into 191-length Mag-2D-FFT, features that are preferred due to their superior performance over benchmark imaged-based types [29]. Four similarity-based VQ models per feature type are trained with resolutions N=8, 16, 32, 64 using a similarity-based Linde-Buzo-Gray (LBG) method.

LMS training maps each frame to the maximum similarity codeword, encoding temporal dynamics the full training video length. Person identification on the VQ and LMS models are performed and the LMS histogram plotted in Figure 1, revealing an optimum segment length of 3-frames per database, meaning the longer the test segment the more potential longest matching segment candidates ($^nC_3$) and the more accurate the model. The XM2VTS has a second peak at 19-frames, on account of more variations per class from larger training data. There is limited benefit from LMS to a 3-frame test video which only has one longest matching segment candidate. If the test video is under 3 segments the LMS model is likely to struggle..

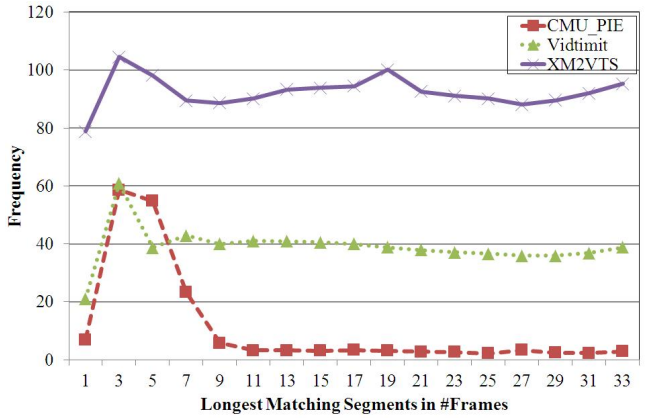

Fig. 1: LMS Histogram of the frequency of longest matching segment lengths on the CMU-PII, VidTIMIT, and XM2VTS databases.

3.1. Person Identification Using the CMU-PIE Database

The frontal (c27) is selected for training and three-quarter (c05) profiles selected for test for all 68 speakers of the CMU PIE database. The maximum VQ resolution is 60 codewords for the 60-frame training videos. Perfect scores are regularly achieved with full video, so each test video divided into 3x30-frame evenly overlapping segments to get 204 test cases. The similarity-based person identification accuracies in Figure 2 show that the novel introduction of variable long-range LMS dynamics with static features delivers performance improvement to a VQ-based text-independent person identification model with static features on the CMU-PIE database given limited training data and slight variations in pose.

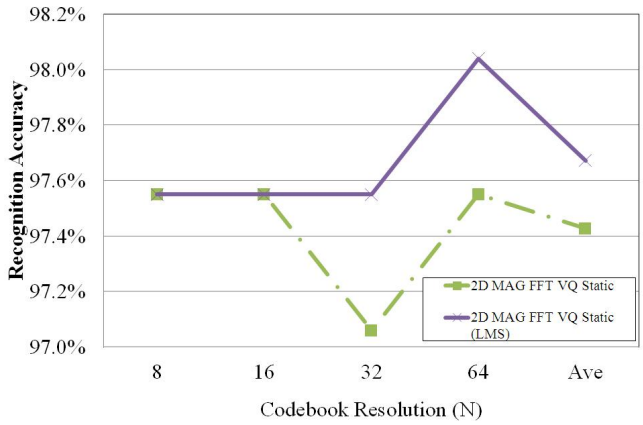

Fig. 2: VQ vs VQ-LMS person identification accuracies on the CMU-PIE database using Mag-2D-FFT features

3.2. Person Identification Using the VidTIMIT Database

In the VidTIMIT talking dataset the first eight utterances of each speakers are used for training, while the last 2 are divided into 60-frame equally overlapping segments to complete a 161 member test set, as perfect scores are regularly achieved with the full video. The Similarity-based person identification accuracies are shown in Figure 3. The results show that the more training data with greater discriminative content in VidTIMIT increases accuracy and consistency with resolution, and LMS temporal models also return distinct improvements on the equivalent baseline VQ models with static features.

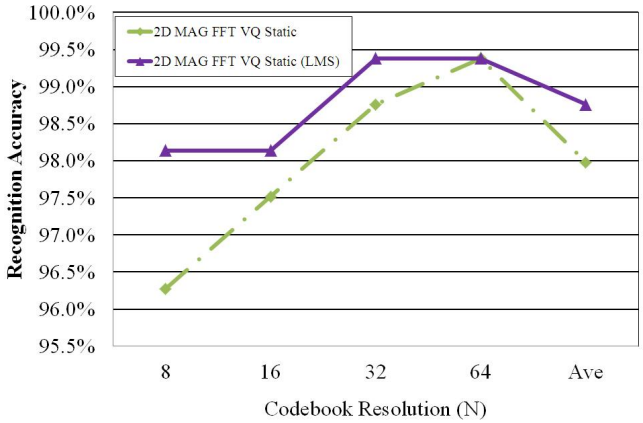

Fig. 3: VQ vs LMS person identification accuracies on the VidTIMIT database for Mag-2D-FFT features

3.3. Person Identification using the XM2VTS Database

Preprocessing extracts 176×144 frames from which normalized 12×16 grayscale lip ROIs are cropped from the significantly larger XM2VTS database. From the four recorded sessions of two utterances each, the first three are used for training and the last for test. Silence is not removed from the database. The first 100-frames of each test video forms a test segment resulting in 590 test cases. 382-length dynamic features comprising of Mag-2D-FFT static and delta features are introduced for benchmarking against face recognition. The Similarity-based person identification accuracies are shown in Figure 4. The results show stable class discrimination with consistent trending across resolution N where LMS recognition accuracies are not just superior; they tends towards face recognition benchmarks with averages of 90.04% static vs 99.96% LMS and 91.10% delta.

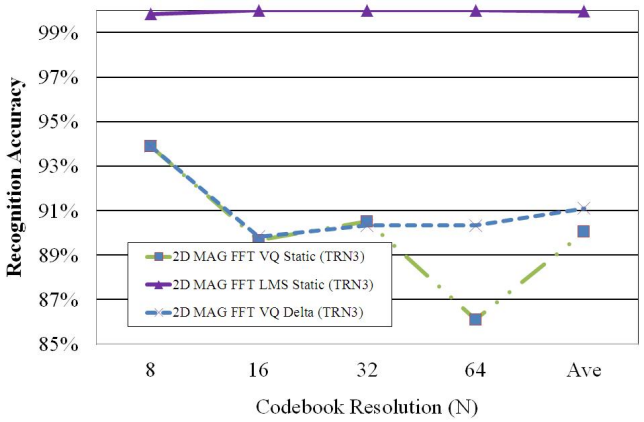

Fig. 4: VQ vs VQ-Delta vs VQ-LMS person identification accuracies on the XM2VTS database for Mag-2D-FFT features.

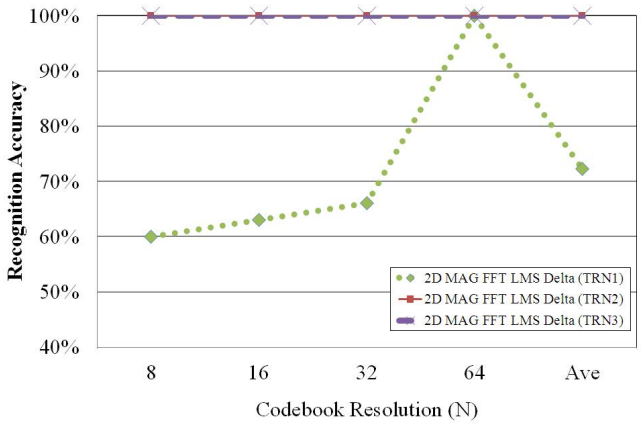

Fig. 5: Similarity-based person identification using VQ with LMS dynamics and Mag-2D-FFT-Δ features

The identification accuracies for a VQ-LMS-Delta setup are compiled in Figure 5 and their performances rated against state of the art lipreading systems in Table 1. One-utterance, two-utterance (protocol 2 [11]), and three utterance training models are developed for all 295 speakers. Once LMS training and delta features are combined the performance takes a significant leap towards face recognition benchmarks, achieving perfect scores on 2-video and 3-video models for all resolutions.

Table 1: Benchmarking person identification performance under near-matching to identical test conditions on the M2VTS and XM2VTS databases

CITE	METHOD	FEATURE	TRAIN/TEST	%REC
Paper	VQ	Mag-2D-FFT-Δ	1/4 (Session ID)	72.29
[20]	HMM	Shape,Intensity	Protocol 2	72.00
[11]	SVM-RBF	Optical flow	Protocol 2	80.00
[13]	HMM	DCT-$\Delta\Delta$	1-3/4 (Session ID)	1.03 EER
[13]	HMM	DWT Z-Score	1-3/4 (Session ID)	1.55 EER
Paper	VQ	Mag-2D-FFT-Δ	1-2/4 (Session ID)	100.00
Paper	VQ	Mag-2D-FFT-Δ	1-3/4 (Session ID)	100.00

4. Conclusion

In this paper LMS is successfully used to model variable long range lipreading dynamics on a given baseline training model for text-independent person identification on the CMU PIE, VidTIMIT, and XM2VTS databases. LMS extracts the complete temporal dynamics of the training videos by mapping it to a frame sequence of the maximum likelihood codewords per person. The results show that the novel introduction of LMS dynamics consistently delivers better person identification performance in the constrained conditions of the CMU-PIE and VidTIMIT databases, and balanced training conditions in the XM2VTS database. When trained with Mag-2D-FFT-Δ features on the XM2VTS database the LMS system matches full face recognition accuracies. This provides a significant real world advantage for faster state of the art recognition solutions that uses limited training data and smaller ROI. Overall LMS effectively encodes long range video dynamics that provides improvement over conventional VQ, improving VQ trained with dynamic delta features. LMS has been applied to VQ, HMM, GMM and arguably can be used with DNNs to further improve biometric temporal modelling. LMS is therefore a credible biometric detection tool given only lip movement with a considerable degree of semi-constrained robustness.

5. References

[1] Kenji Mase and Alex Pentland, "Automatic lipreading by optical-flow analysis," *Systems and Computers in Japan*, vol. 22, no. 6, pp. 67–76, 1991.

[2] Iain Matthews, Timothy F Cootes, J Andrew Bangham, Stephen Cox, and Richard Harvey, "Extraction of visual

features for lipreading," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 2, pp. 198–213, 2002.

[3] Gerasimos Potamianos, Chalapathy Neti, and Sabine Deligne, "Joint audio-visual speech processing for recognition and enhancement," in *AVSP 2003-International Conference on Audio-Visual Speech Processing*, 2003.

[4] Sarah Hilder, Richard Harvey, and Barry-John Theobald, "Comparison of human and machine-based lip-reading.," in *AVSP*, 2009, pp. 86–89.

[5] Liopa, "Liopa | lip based biometric authentication," 2015.

[6] Josef Kittler, YP Li, Jiri Matas, and MU Ramos Sánchez, "Combining evidence in multimodal personal identity recognition systems," in *Audio-and Video-based Biometric Person Authentication*. Springer, 1997, pp. 327–334.

[7] Eli Yamamoto, Satoshi Nakamura, and Kiyohiro Shikano, "Lip movement synthesis from speech based on hidden markov models," *Speech Communication*, vol. 26, no. 1, pp. 105–115, 1998.

[8] HE Cetingul, Yücel Yemez, Engin Erzin, and A Murat Tekalp, "Discriminative lip-motion features for biometric speaker identification," in *Image Processing, 2004. ICIP'04. 2004 International Conference on*. IEEE, 2004, vol. 3, pp. 2023–2026.

[9] Hasan Ertan Çetingül, Yücel Yemez, Engin Erzin, and A Murat Tekalp, "Robust lip-motion features for speaker identification.," in *ICASSP (1)*, 2005, pp. 509–512.

[10] H Ertan Cetingul, Yücel Yemez, Engin Erzin, and A Murat Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *Image Processing, IEEE Transactions on*, vol. 15, no. 10, pp. 2879–2891, 2006.

[11] Maycel Isaac Faraj and Josef Bigun, "Lip motion features for biometric person recognition," *Visual Speech Recognition: Lip Segmentation and Mapping*, 2009.

[12] Amr Bakry and Ahmed Elgammal, "Mkpls: manifold kernel partial least squares for lipreading and speaker identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 684–691.

[13] Rowan Seymour, ," in *Audio-Visual Speech and Speaker Recognition, PhD Thesis*. Queens University Belfast, 2008.

[14] Gareth J Edwards, Christopher J Taylor, and Timothy F Cootes, "Improving identification performance by integrating evidence from sequences," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.* IEEE, 1999, vol. 1.

[15] Bernhard Fröba, Constanze Rothe, and Christian Kublbeck, "Evaluation of sensor calibration in a biometric person recognition framework based on sensor fusion," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 512–517.

[16] Jong-Seok Lee and Cheol Hoon Park, "Temporal filtering of visual speech for audio-visual speech recognition in acoustically and visually challenging environments," in *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007, pp. 220–227.

[17] Zeliang Zhang and Xiongfei Li, "An effective parameter estimation algorithm of the visual language features.," *International Journal of Digital Content Technology & its Applications*, vol. 6, no. 4, 2012.

[18] Mihaela Gordan, Constantine Kotropoulos, and Ioannis Pitas, "Application of support vector machines classifiers to visual speech recognition," in *Image Processing. 2002. Proceedings. 2002 International Conference on*. IEEE, 2002, vol. 3, pp. III–129.

[19] J. Luettin, N.A. Thacker, and S.W. Beet, "Speaker identification by lipreading," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, Oct 1996, vol. 1, pp. 62–65 vol.1.

[20] P. Jourlin, J. Luettin, D. Genoud, and H. & Wassner, "Acoustic-labial speaker verification," *Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication*, pp. 319–326, 1997.

[21] Xiaoming Liu and Tsuhan Chen, "Video-based face recognition using adaptive hidden markov models," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. IEEE, 2003, vol. 1, pp. I–340.

[22] O. Shipilova, "Person recognition based on lip movements," 2006, [retrieved July 15, 2006 from http://www.it.lut.fi/kurssit/03-04/010970000/seminars/Shipilova.pdf].

[23] Josef Bigun Maycel-Isaac Faraj, "Synergy of lip-motion and acoustic features in biometric speech and speaker recognition," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1169–1175, September 2007.

[24] Mehraj Haider and Hussain Ajaz Mir, "Lip based recognition: A comparative analysis," *International Journal of Electronics & Communication Technology*, vol. 3, no. 4, pp. 153–157, 2012.

[25] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[26] Nitish Srivastava and Ruslan R Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[27] Ayeh Jafari, Ramji Srinivasan, Daniel Crookes, and Ming Ji, "A longest matching segment approach for text-independent speaker recognition," pp. 1469–1472, 9 2010.

[28] Ji Ming, Ramji Srinivasan, and Danny Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 822–836, 2011.

[29] Paul C Brown, "Visual-only person and word recognition from lip motion dynamics, phd thesis," August 2016.

[30] Tadahiro Ohmi, Koji Kotani, Qiu Chen, and Feifei Lee, "Face recognition algorithm using vector quantization codebook space information processing," *Intelligent Automation & Soft Computing*, vol. 10, no. 2, pp. 129–142, 2004.

[31] Mohamed Debyeche, Jean Paul Haton, and Amrane Houacine, "Improved vector quantization approach for discrete hmm speech recognition system.," *Int. Arab J. Inf. Technol.*, vol. 4, no. 4, pp. 338–344, 2007.

[32] Shishir Bashyal and Ganesh K Venayagamoorthy, "Recognition of facial expressions using gabor wavelets and learning vector quantization," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 7, pp. 1056–1064, 2008.

[33] Dr. H.B. Kekre, Tanuja Sarode, Prachi Natu, and Shachi Natu, "Performance comparison of face recognition using dct against face recognition using vector quantization algorithms lbg, kpe, kmcg, kfcg," 2010, vol. 4, pp. 377–389.

[34] Steve Young, Gunnar Evermann, and Mark Gales, ," in *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2009, vol. 3.4.

[35] HB Kekre, Sudeep D Thepade, Tanuja K Sarode, and Vashali Suryawanshi, "Image retrieval using texture features extracted from glcm, lbg and kpe," *International Journal of Computer Theory and Engineering*, vol. 2, no. 5, pp. 695, 2010.

[36] Jie Lin, Ji Ming, and D. Crookes, "Robust face recognition with partially occluded images based on a single or a small number of training samples," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 881–884.

[37] Terence Sim, Simon Baker, and Maan Bsat, "The cmu pose, illumination, and expression (pie) database," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. IEEE, 2002, pp. 46–51.

[38] Conrad Sanderson, ," in *The VidTIMIT Database*. IDIAP Com 02-06, 2002.

[39] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "Xm2vtsdb: The extended m2vts database," in *In Second International Conference on Audio and Video-based Biometric Person Authentication*, 1999, pp. 72–77.