Visual Speech Recognition Using Weighted Dynamic Time Warping

Kyungsun LEE†, Minseok KEUM†, David K. HAN††, Nonmembers, and Hanseok KO†††, Member

SUMMARY It is unclear whether Hidden Markov Model (HMM) or Dynamic Time Warping (DTW) mapping is more appropriate for visual speech recognition when only small data samples are available. In this letter, the two approaches are compared in terms of sensitivity to the amount of training samples and computing time with the objective of determining the tipping point. The limited training data problem is addressed by exploiting a straightforward template matching via weighted-DTW. The proposed framework is a refined DTW by adjusting the warping paths with judiciously injected weights to ensure a smooth diagonal path for accurate alignment without added computational load. The proposed WDTW is evaluated on three databases (two in the public domain and one developed in-house) for visual recognition performance. Subsequent experiments indicate that the proposed WDTW significantly enhances the recognition rate compared to the DTW and HMM based algorithms, especially under limited data samples.

key words: visual speech recognition, lip reading

1. Introduction

Automatic visual speech recognition (VSR) plays an important role for speech-interface-based smart TVs and voice command car navigation systems in extremely noisy environments. VSR is traditionally viewed as a supplement to speech recognition. In recent years, many researchers have studied VSR solely from visual channels[1]–[3]. However, robust VSR from visual channels still faces challenges considering the variations of lip shape and appearance related to speaking speeds and intensities [4]. Due to these variations, a large number of training data sets on isolated phrases are required for developing an effective VSR system. However, it is generally not practical to collect a sufficient amount of data from a subject by asking him/her to repeat phrases a large number of times. For these reasons, only a few VSRs based on HMM or DTW have recently been demonstrated on a predefined-phrase dataset. The objective of this letter, therefore, is to propose a new DTW variant with an adaptive weight-based distance measure to overcome the limited database issue.

2. Proposed Method

2.1 Feature Extraction

An overview of the proposed algorithm is shown in Fig. 1. First, a whole face is detected from input image, and the eyes are located from the detected face [5], [6]. Positions of the eyes are used to localize the lip region. If it is successful in detecting the lip region, the associated lip feature is calculated. For this approach, a combination of HOG and LBP features are used. The histogram of oriented gradients (HOG) [7] feature is known as the distribution of intensity gradients and edge orientations, which is employed to describe local shape information. While the Local Binary Pattern (LBP) operator[8] is an exceptional texture descriptor, it is highly discriminative and its key advantages, namely its invariance to monotonic gray level changes and computational efficiency, makes it suitable for demanding image analysis tasks. Thus, an augmented feature vector combining the HOG feature with the cell-structured LBP feature is employed in this Letter as the feature for VSR training. HOG alone performs poorly when the mouth region is cluttered with noisy edges. LBP is complementary in this aspect. It can filter out noise using the concept of uniform pattern [8]. Hence, appearance of a lip is well captured if the edge/local shape information is combined with the texture information.

2.2 Visual Voice Activity Detection (VVAD)

For effective word classification, visual voice activity detection (VVAD) is an important step that is required to accurately segment the voice activity interval. To decide the interval, a local variance of histogram (LVH) as a statistical measure of lip motion change is proposed here. The measure of Local Variance is as follows [8]:

Fig. 1 Overview of proposed VSR framework.
used as a visual cue for detecting speech as follows:

consequently, the number of pixels with high value of LVH is
in high variance bins. An example of a closed and opened
of an open mouth would result larger values of histogram
hibiting low intensity values, it is hypothesized that LVH
exhibiting high intensity values and the gap in-between ex-
an open mouth consists of separated lips and teeth regions

Fig. 2  Plot of local variance histogram and feature extraction result: (a) LVH of closed mouth (b) LVH of open mouth (c) Distribution of the number of high bin pixels \( x[n] \) of LVH in video sequence.

\[
VAR_{PR} = \frac{1}{P} \sum_{P=0}^{P-1} (g_P - \mu)^2, \text{ where } \mu = \frac{1}{P} \sum_{P=0}^{P-1} g_P
\]  

(1)

where \( g_P \) corresponds to the gray value of \( P \) equally spaced pixels on a circle of radius \( R \). For the experimental validation, \( P \) was set at 8 and \( R \) was at 1. By observing that an open mouth consists of separated lips and teeth regions exhibiting high intensity values and the gap in-between exhibiting low intensity values, it is hypothesized that LVH of an open mouth would result larger values of histogram in high variance bins. An example of a closed and opened mouth underscoring the observation is shown in Fig. 2. Consequently, the number of pixels with high value of LVH is used as a visual cue for detecting speech as follows:

\[
c[n] = \sum_{i=T}^{M} H_n[i], \quad x[n] = |c[n] - c[n-1]|
\]

(2)

where \( c[n] \) denotes the number of pixels of the lip region at the \( n \)-th video frame whose LVH bin is above a threshold \( T \) and \( H_n[i] \) is its bin at \( t = 0, \ldots, M \). The threshold \( T \) can be computed according to the Neyman-Pearson theorem for each video sequence and is dependent on the distribution of silent frames. Parameter \( x[n] \) which is a difference of frame sequence \( n, n-1 \) is used for dynamic lip motion feature. Finally, LVAD is done by applying the results of edge detection filter to a state transition diagram proposed by Q. Li, et al. [9].

2.3 Classification

Hidden Markov Model (HMM) [10] has been used successfully in many applications where temporal patterns change dynamically over time. However, HMM requires a lot of samples for training. On the contrary, Dynamic Time Warping (DTW) completely eludes the speech modelling step which requires training [11]. The main advantage of DTW is that it performs well with just a few templates. DTW is used to compute a distance between two time series. A time series is a list of samples taken from a signal, ordered by the time that the respective samples were obtained. DTW solves the discrepancies between two matching time series by minimizing the local distance, thereby recovering optimal alignments between sample points in the two time series. However, the conventional DTW calculates the distance of all points between two series with equal weight of each point regardless of the phase difference between a reference point and a testing point. This may lead to misclassification in VSR. Relative significance depending on the phase difference between points should be considered. Therefore, this letter proposes an approach for VSR based on a novel distance measure, coined as “weighted dynamic time warping (WDTW)”, which essentially assigns nearer neighbors with smaller weights depending on the phase difference between a reference point and a testing point [12].

The WDTW-distance between two time series \( X : (x_1, x_2, \ldots, x_m) \) and \( Y : (y_1, y_2, \ldots, y_n) \) with lengths \( m \) and \( n \) can be calculated in a dynamic programming approach as follows.

\[
D(i, j) = \min \left\{ \begin{array}{l}
n\cdot D(i, j-1) \\
\cdot D(i-1, j) \\
\cdot D(i-1, j-1) + w_{|i-j|} \cdot d(x_i, y_j) \end{array} \right.
\]

(3)

The particular choice of recurrence equation and “local” distance function \( d(\cdot, \cdot) \) varies with the application. Using the given three values \( D(i, j-1), D(i-1, j) \) and \( D(i-1, j-1) \) in the calculation of \( D(i, j) \) realizes a local continuity constraint [Fig. 3 (a)], which ensures smooth time warping (e.g. no samples left out in warping). The Sakoe-Chiba band constraint [11] is used to ensure this path stay close to the diagonal of the matrix which contains the \( D(i, j) \) (Fig. 3 (b)). A weight, \( w_{|i-j|} \), in Eq. (3) is a function of the distance between time series points. The logistic function proposed in Eq. (4) is of a modified form with its bounds controlled by assigned parameter as follows.

\[
w(i) = \frac{w_{\max}}{1 + \exp(-s(i - m_c))}
\]

(4)

where \( m_c \) is the midpoint of a sequence and \( w_{\max} \) is the de-
sired upper bound for the weight parameter, and \( s \) is an empirical constant that controls the slope of the function.

In order to proceed with the WDTW algorithm, a local distance function needs to be defined that compares the feature sets. For this purpose, a squared Euclidean distance from images \( I \) and \( J \) is employed as follows.

\[
d(x_i, y_j) = \sum_{k=1}^{N} (f_k(I, i) - f_k(J, j))^2
\]

where \( f_k \) is the extracted \( k \)-th feature set of \( N \)-vectors from images. The optimal distance between \( X \) and \( Y \) is normalized by length \( C \) of the warping path \((i_1, j_1), \ldots, (i_k, j_k)\):

\[
D(X, Y) = \frac{1}{C} \sum_{k=1}^{K} w_{|i_k-j_k|}d(x_{i_k}, y_{j_k})
\]

Finally, best alignment path is the one attaining the smallest distance between \( X \) and \( Y \) utterance pair.

3. Experimental Results

Relevant experiments were conducted to verify effectiveness of the proposed method in visual speech recognition task. Two in the public domain databases (e.g. OuluVS, AVLetters) and one developed in-house database (e.g. KUVSR) were used in subject-dependent setting as follows:

(1) **OuluVS database** [2]: It consists of ten different everyday phrases. Each phrase is uttered by 20 subjects up to five times. The video sequences were set to 25 fps and each frame was at resolution of 720 x 576 pixels. The dataset contains sequences of images for mouth area with average resolution of 120 x 70 pixels.

(2) **AVLetters database** [1]: It contains uttering scenes of 10 subjects. Each speaker repeats every English letter (A~Z) exactly three times, with a total of 780 video sequences. The speaker was requested to start and end utterance of every letter in a neutral state (mouth closed). No head motion/rotation is allowed from speakers. Every frame is a 60 x 80 pixel image of the mouth area.

(3) **KUVSR database**: This is our own in-house visual speech dataset collection for VSR performance evaluation. Video data were captured with a 640 x 480 pixel resolution at 30 fps. The database contains the recording of 17 people with each person uttering 10 Korean words. Each word was repeated 50 times. These words are listed in Table 1 and Fig. 4.

For speaker-dependent experiments, *leave-one-utterance-out* strategy was utilized for cross validation. Hence, one sequence is left for test and the remaining sequences are used for training. For feature descriptors, \( LBP_{u}^{2 \times 8} \) and \( HOG_{3 \times 3} \) were used to extract the visual features [7], [8]. The images were divided into \( 1 \times 5 \) blocks and each block was overlapped with its neighboring blocks to counter the vertical misalignment as described by [2]. An overlap ratio of 70% of the original non-overlapping block size was selected empirically. The recognition rate of each algorithm based on DTW, HMM, and the proposed WDTW was obtained by increasing the number of data samples from the KUVSR database. Figure 5 shows the resulting average recognition rate per number of data samples. It can be seen that the proposed WDTW based method achieved higher recognition rate than the DTW and HMM based methods with just a few data samples. As shown in Fig. 5, HMMs need around 30 samples to achieve the same VSR rate as that of WDTW. This is an important result that can indicate which method should be selected when it comes to the amount of samples available in user’s dataset. In addition, the computing time required for training and recognition was measured with 30 samples using a PC with 2.66GHz CPU and 8G RAM. As shown in Table 2, the average time needed to train an HMM for each class was 51.3ms. On the contrary, since WDTW has no training stage, no time is required. Recognition time is lower on HMM than WDTW due to the fact that the recognition times using WDTW is directly proportional to the number of comparisons. How-

![Fig. 4 Sample images of KUVSR dataset](image)

![Fig. 5 Recognition rate according to number of data samples](image)

| Table 1 | Phrases included in the KUVSR dataset |
|---------|-------------------------------------|
| **Pronunciation** | **Meaning** | **Pronunciation** | **Meaning** |
| /ah-dong/ | Child | /mood/ | Mode |
| /eum-sung/ | Voice | /na-i/ | Age |
| /gu-gi/ | Google | /oh-rak/ | Entertainment |
| /jung-ji/ | Stop | /si-jak/ | Start |
| /moe-il/ | Mail | /li-vi/ | Television |
ever, the recognition time on WDTW is acceptable for most applications. The performance was also compared with the recent visual speech recognition works (e.g. [1]–[3]) using two public databases: OuluVS and AVLetters, as shown in Table 3. Overall, the results confirm that the proposed approach outperforms significantly better than the conventional approaches especially with limited number of training samples.

### 4. Conclusion

In this Letter, a novel time series classification algorithm is proposed for VSR in data limited scenarios. The HOG-LBP combination was employed for local feature descriptor and LVH was developed as visual voice activity detector. WDTW was adopted as the basis for new distance measures for time sequence data mapping. The effectiveness of WDTW compared to the DTW and HMM was verified through representative VSR performance experiments. These evaluations revealed that the proposed WDTW based method indeed improved the recognition performance using public data sets consistently under small data sample cases. Based on the VSR result, having 30 samples was determined as the tipping point for WDTW-to-HMM transition with respect to VSR rate performance.

### Acknowledgments

This research was funded and supported by Samsung Electronics Co., Ltd.

### References

[1] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey, “Extraction of visual features for lipreading,” IEEE Trans. Pattern Anal. Mach. Intell., vol.24, no.2, pp.198–213, 2002.
[2] G. Zhao, M. Barnard, and M. Pietikainen, “Lipreading with local spatiotemporal descriptors,” IEEE Trans. Multimedia, vol.11, no.7, pp.1254–1265, 2009.
[3] E.-J. Ong and R. Bowden, “Learning temporal signatures for lip reading,” IEEE Conference on Computer Vision Workshops, Univ. of Surrey, pp.958–965, Guildford, UK, 2011.
[4] J.L. Newman, B.J. Theobald, and S.J. Cox, “Limitations of visual speech recognition,” Int. Conf. on Audio-Visual Speech Process, pp.1–4, 2010.
[5] P. Viola and M.J. Jones, “Robust real-time face detection,” Int. J. Comput. Vision, vol.57, no.2, pp.137–154, 2004.
[6] Z. Niu, S. Shan, S. Yan, X. Chen, and W. Gao, “2d cascaded adaboost for eye localization,” International Conference on Pattern Recognition, pp.1216–1219, 2006.
[7] N. Dalal and B. Triggs, “Histogram of oriented gradients for human detection,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.886–893, 2005.
[8] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Anal. Mach. Intell., vol.24, no.7, pp.971–987, 2002.
[9] Q. Li, J. Zheng, A. Tsai, and Z. Zhou, “Robust endpoint detection and energy normalization for real-time speech and speaker recognition,” IEEE Trans. Speech and Audio Process., vol.10, no.3, pp.146–157, 2002.
[10] L. Rabiner, “A tutorial on Hidden Markov Models and selected applications in speech recognition,” Proc. IEEE, vol.77, no.2, pp.257–286, 1989.
[11] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” IEEE Trans. Acoust. Speech Signal Process., vol.26, no.1, pp.43–49, 1978.
[12] Y.-S. Jeong, M.K. Jeong, and O.A. Omitaomu, “Weighted dynamic time warping for time series classification,” Pattern Recognition, vol.44, no.9, pp.2231–2240, 2011.