RESOLUTION LIMITS ON VISUAL SPEECH RECOGNITION

Helen L. Bear, Richard Harvey, Barry-John Theobald, Yuxuan Lan

School of Computing Sciences , University of East Anglia, Norwich, NR4 7TJ, UK.

helen.bear@uea.ac.uk, r.w.harvey@uea.ac.uk, b.theobald@uea.ac.uk, y.lan@uea.ac.uk

ABSTRACT

Visual-only speech recognition is dependent upon a number of factors that can be difficult to control, such as: lighting; identity; motion; emotion and expression. But some factors, such as video resolution are controllable, so it is surprising that there is not yet a systematic study of the effect of resolution on lip-reading. Here we use a new data set, the Rosetta Raven data, to train and test recognizers so we can measure the affect of video resolution on recognition accuracy. We conclude that, contrary to common practice, resolution need not be that great for automatic lip-reading. However it is highly unlikely that automatic lip-reading can work reliably when the distance between the bottom of the lower lip and the top of the upper lip is less than four pixels at rest.

1. INTRODUCTION

A typical lip-reading system has a number of stages: first, the data are pre-processed and normalised; second, the face and lips are tracked; third, visual features are extracted and classified. In practice many systems find tracking challenging, which affects the overall recognition performance. However, the tracking problem is not insurmountable and it is now realistic to track talking heads in outdoor scenes filmed with shaky hand-held cameras [2], so we focus on feature extraction using Active Appearance Models (AAMs) [4]. We select AAMs since they have been shown to have robust performance on a number of datasets ([8, 9, 10, 11] for example) and out perform other feature types [6].

2. DATASET AND FEATURE EXTRACTION

An AAM is a combined model of shape and appearance trained to fit to a whole video sequence [4]. Training creates a mean model and a set of modes, which may be varied to create shape and appearance changes. In training, a small number of frames are identified and manually landmarked. These models are Procrustes-aligned and the mean and covariance of the shape are computed. The eigenvectors of the covariance matrix give a set of modes of variation, which are used to deform the mean shape. For appearance a mesh shape-normalizes the images via a piecewise affine transform so the

pixels of all images are aligned. We then compute the mean and the eigenvectors of their covariance. Concatenating the shape and appearance features forms the feature vector for training and testing. Having built a model on a few frames, it is fitted to unseen data using inverse compositional fitting [8]. The Rosetta Raven data are four videos of two North American talkers (each talker in two videos), reciting Edgar Allen Poe's 'The Raven'. The poem was published in 1845 and, recited properly, the poem has trochaic octameter [13], but this does not appear to have been followed by the talkers in this dataset. Figure 3(a) shows example frames from the high-definition video of the two talkers. The database summarised in Table 1 was recorded at 1440×1080 non-interlaced resolution at 60 frames per second. The talkers wore no make-up.

Video	Train Images	Fit Images	Duration
Talker1 - 1	10	21,648	00:06:01
Talker1 - 2	10	21,703	00:06:02
Talker2 - 1	11	31,858	00:08:52
Talker2 - 2	11	33,328	00:09:17

Table 1: Frame images from each video

All four videos were converted into a set of images (one per frame) with ffmpeg using image2 encoding at full high-definition resolution (1440×1080).

To build an initial model we select the first frame and nine or ten others randomly. These *key frames* are hand-labelled with a model of a face and lips. This preliminary model is then fitted, via inverse compositional fitting [8] to the remaining frames (Table 1 lists total frames for each video). At this stage therefore we have tracked and fitted full face talker dependent AAMs on full resolution lossless PNG frame images as in Figure 1.

These models are then decomposed into sub-models for the eyes, eyebrows, nose, face outline and lips (this allows us to obtain a robust fit from the full face model but process only the lips). Figure 2 shows both talker's lips sub-model. Next, the video frames used in the high-resolution fitting were down-sampled to each of the required resolutions (Table 2) by nearest neighbor sampling and then up-sampled via bilinear sampling (Figure 3) to provide us with 18 sets of frames. These new frames are the same physical size as the origi-

Fig. 1: Showing full face mesh for talker T1 (left) and T2 (right)

nal (1440 × 1080) but contain far less information due to the downsampling.

1440 × 1080	960 × 720	720 × 540	360 × 270
240 × 180	180 × 135	144 × 108	120 × 90
90 × 67	80 × 60	72 × 54	65 × 49
69 × 45	55 × 42	51 × 39	48 × 36
45 × 34	42 × 32		

Table 2: Resolutions

We are most interested in the affect of low resolution on the loss of lip-reading information rather than, say the affect it would also have on the tracker (many AAM trackers lose track quite easily at low resolutions and we do not wish to be overwhelmed with catastrophic errors due to tracking problems which can often be solved in other ways [12]). Consequently the shape features in this experiment are unaffected by the downsample whereas as the appearance features vary (a useful benchmark as it will turn out).

Fig. 2: Showing lip-only mesh for talker T1 (left) and talker T2 (right)

For talker1 (T1), we retain 6 shape and 14 appearance parameters and for talker2 (T2), 7 shape and 14 appearance parameters. The number of parameters was chosen to retain 95% of

the variance in the usual way [4].

3. RECOGNITION METHOD

vID	Phones	vID	Phones
v01	/p/ /b/ /m/	v10	/i/ /ih/
v02	/f/ /v/	v11	/eh/ /ae/ /ey/ /ay/
v03	/th /dh/	v12	/aa/ /ao/ /ah/
v04	/t/ /d/ /n/ /k/ /g/ /h/ /j/ /ng/ /y/	v13	/uh/ /er/ /ax/
v05	/s/ /z/	v14	/u/ /uw/
v06	/l/	v15	/oy/
v07	/r/	v16	/iy/ /hh/
v08	/sh/ /zh/ /ch/ /jh/	v17	/aw/ /ow/
v09	/w/	v18	silence

Table 3: Phone to viseme mapping

To produce the ground truth we listen to each recitation of the poem and produced a ground truth text (some recitations of the poem were not word-perfect). This word transcript is converted to an American English phone level transcript using the CMU pronunciation dictionary [3]. However not all phones are visible on the lips, so we select a mapping from phones to *visemes* (which are the visual equivalent of phonemes). Here, the viseme mapping is based upon Walden's trained consonants [14] and Montgomery et al's vowel [7] classifications as illustrated in Table 3. Viseme recognition is selected over phoneme recognition as, on a small data set, it has the benefits of reducing the number of classes needed (the model for each class forms a single recogniser) and increasing the training data available for each viseme classifier. Note that not all visemes are equally represented in the data as is shown by the viseme counts in Figures 4 and 5.

For each talker, a test fold is randomly selected as 42 of the 108 lines in the poem. The remaining lines are used as training folds. Repeating this five times gives five-fold cross-validation. Visemes cannot be equally represented in all folds. For recognition we use Hidden Markov Models (HMMs) implemented in the Hidden Markov Toolkit (HTK) [15]. An HMM is initialised using the 'flat start' method using a prototype of five states and five mixture components and the information in the training samples. We choose five states and five mixtures via [9]. We define an HMM for each viseme plus silence and short-pause labels (Table 3) and re-estimate the parameters four times with no pruning. We use the HTK tool `HHEd` to tie together the short-pause and silence models between states two and three before re-estimating the HMMs a further two times. Then `HVite` is used to force-align the data using the word transcript [1].

---

[1] We use the `-m` flag with `HVite` with the manual creation of a viseme version of the CMU dictionary for word to viseme mapping so that the force-alignment produced uses the break points of the words.

Fig. 3: (a) 1440×1080-Original resolution image for T1 & T2, (b) 60×45-T1 downsampled, and (c) 1440×1080-T1 restored
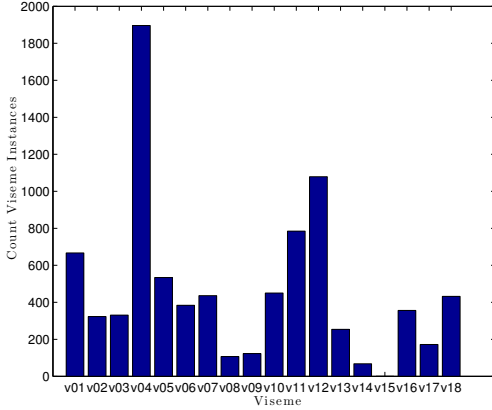

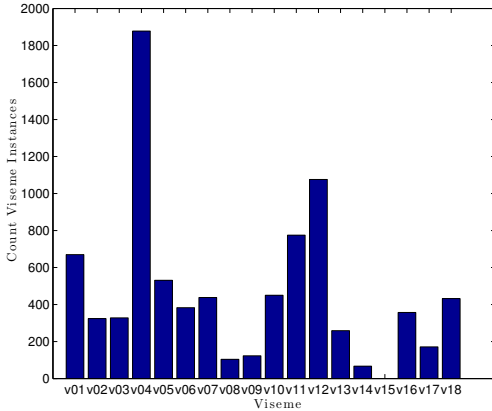**Fig. 4**: Visemes present in both T1 videos

Fig. 5: Visemes present in both T2 videos

The HMMs are now re-estimated twice more, however now we use the force-aligned viseme transcript rather than the original viseme transcript used in the previous HMM re-estimations. To complete recognition using our HMMs we require a word network. We use `HLStats` and `HBuild` to make both a Unigram Word-level Network (UWN) and a Bi-gram Word-level Network (BWN). Finally `HVite` is used with the different network support for the recognition task and

`HResults` gives us the correctness and accuracy values.

4. RESULTS

Recognition performance of the HMMs can be measured by both correctness, C, and accuracy, A,

$$C = \frac{N - D - S}{N}, \quad A = \frac{N - D - S - I}{N},$$

where S is the number of substitution errors, D is the number of deletion errors, I is the number of insertion errors and N the total number of labels in the reference transcriptions [15]. We use accuracy as a measure rather than correctness since it accounts for all errors including insertion errors which are notoriously common in lip reading. An insertion error occurs when the recognizer output has extra words/visemes missing from the original transcript [15]. As an example one could say "Once upon a midnight dreary", but the recognizer outputs "Once upon upon midnight dreary dreary". Here the recognizer has inserted two words which were never present and it has deleted one.

Figure 6 shows the acurracy, A, versus resolution for an UWN. The x-axis is calibrated by the vertical height of the lips of each talker in their rest position. For example, at the maximum resolution of 1440×1080 talker T1 has a lip-height of approximately 26 pixels in the rest position whereas T2 has a lip-height of approximately 17 pixels. The worst performance is from talker T2 using shape-only features. Note that the shape features do not vary with resolution so any variation in this curve is due to the cross-fold validation error (all folds do not contain all visemes equally). Nevertheless the variation is within an error bar. The poor performance is, as usual with lip-reading, a standard error dominated by insertion errors (hence the negative A values). The usual explanation for this effect is that shape data contains a few characteristic shapes (which are easily recognised) in a sea of indistinct shapes - it is easier for a recogniser to insert garbage symbols than it is to learn the duration of a symbol which has indistinct start and end shapes due to co-articulation. Talker T1 has more distinctive shapes so scores better on the shape feature.
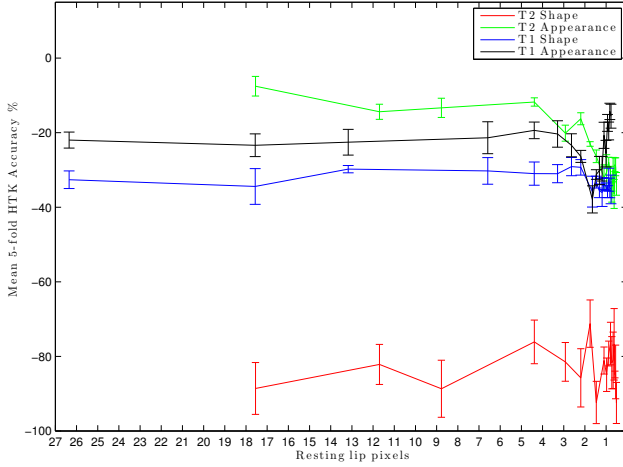
Fig. 6: Mean viseme recognition accuracy supported by UWN at 18 degraded resolutions shown by vertical resting lip height in pixels. Error bars show ± one standard error.

However it is the appearance that is of more interest since this varies as we downsample. At resolutions lower than four pixels it is difficult to be confident that the shape information is effective. However the basic problem is a very low error rate (shown in Figure 6) therefore we adopt a more supportive word model.
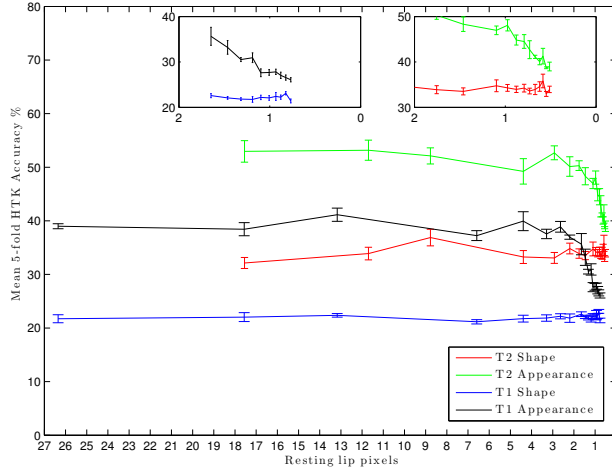

Fig. 7: Mean viseme recognition accuracy supported by BWN at 18 degraded resolutions shown by vertical resting lip height in pixels. Error bars show ± one standard error.

Figure 7 shows the recognition accuracy versus resolution (represented by the same x-axis calibration in Figure 6) for a BWN. It also includes two sub-plots which zoom the right-most part of the graph. Again the shape models perform worse than the appearance models but, looking at the zoomed

plots, appearance never becomes as poor as shape performance even at very low resolutions. As with the UWN accuracies, there is clear inflection point at around four pixels (two pixels per lip) and by two pixels the performance has declined noticeably.

Rest	Talker 1			Talker 2		
Pixels	Ins	Del	Sub	ins	Del	Sub
> 4	69.8	667.0	259.6	114.2	467.8	284.6
< 4	61.0	729.2	271.0	106.0	464.4	300.0

Table 4: Error rates for insertions, deletions and substitutions where the pixels are more than four covering the lips at rest (where recognition is still reliable), and less than four pixels where recognition performance falls. Values are averaged over all five folds.

Table 4 shows the deletion, insertion and substitution error rates for the recognition performance of resolutions which are just above and below the four pixels at rest. We see that the insertion errors are significantly lower than both deletions and substitutions so we are confident that our accuracy scores are accurate insertions despite negative accuracy scores being achieved with the Unigram Word Network support in Figure 6.

5. CONCLUSIONS

We have shown that the performance of simple visual speech recognizers has a threshold effect with resolution. For successful lip-reading one needs a minimum four pixels across the closed lips. However the surprising result is the remarkable resilience that computer lip-reading shows to resolution. Given that modern experiments in lip-reading usually take place with high-resolution video ([16] and [1] for example) the disparity between measured performance (shown here) and assumed performance is very striking.

Of course higher resolution may be beneficial for tracking but, in previous work we have been able to show other factors believed to be highly detrimental to lip-reading such as off-axis views [5] actually have the ability to improve performance rather than degrade it. We have also noted that previous shibboleths of outdoor video, poor lighting and agile motion affecting performance can all be overcome [1]. It seems that in lip-reading it is better to trust the data than conventional wisdom.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] R. Bowden, S. Cox, R. Harvey, Y. Lan, E.-J. Ong, G. Owen, and B.-J. Theobald. Recent developments in automated lip-reading. In *SPIE Security+ Defence*, pages 89010J–89010J. International Society for Optics and Photonics, 2013.

[2] R. Bowden, S. J. Cox, R. W. Harvey, Y. Lan, E.-J. Ong, G. Owen, and B. Theobald. Is automated conversion of video to text a reality? In C. Lewis and D. Burgess, editors, *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII*, volume SPIE 8546, pages 85460U–85460U–9. SPIE, 2012.

[3] Carnegie Mellon University. CMU pronounciation dictionary, 2008.

[4] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681 –685, Jun 2001.

[5] Y. Lan, B.-J. Theobald, and R. Harvey. View independent computer lip-reading. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 432–437. IEEE, 2012.

[6] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, and R. Bowden. Improving visual features for lip-reading. In *Proceedings of International Conference on Auditory-Visual Speech Processing*, volume 201, 2010.

[7] D. Massaro. *Perceiving Talking Faces*. The MIT Press, 1998.

[8] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[9] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(2):198 –213, feb 2002.

[10] E. Ong and R. Bowden. Robust lip-tracking using rigid flocks of selected linear predictors. In *8th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG2008)*, pages 247–254, 2008.

[11] E. Ong, Y. Lan, B. Theobald, R. Harvey, and R. Bowden. Robust facial feature tracking using selected multi-resolution linear predictors. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1483–1490. IEEE, 2009.

[12] E.-J. Ong, Y. Lan, B. Theobald, R. Harvey, and R. Bowden. Robust facial feature tracking using selected multi-resolution linear predictors. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1483–1490, Sept 2009.

[13] P. F. Quinn. The critical mind of Edgar Poe: Claude Richard. Edgar Allan Poe: Journaliste et critique. *Poe Studies-Old Series*, 13(2):37–40, 1980.

[14] B. E. Walden, R. A. Prosek, A. A. Montgomery, C. K. Scherr, and C. J. Jones. Effects of training on the visual recognition of consonants. *Journal of Speech, Language and Hearing Research*, 20(1):130, 1977.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchec, and P. Woodland. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.

[16] Z. Zhou, X. Hong, G. Zhao, and M. Pietikäinen. A compact representation of visual speech data using latent variables. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1):181–187, Jan 2014.