TRACKING FACIAL MARKERS WITH AN ADAPTIVE MARKER COLLOCATION MODEL

Jon Barker

Department of Computer Science, University of Sheffield
211 Portobello St, Sheffield, S1 4DP, UK (Europe)
phone: +44 (0)114 22 21800, fax: +44 (0)114 22 2181, email: j.barker@dcs.shef.ac.uk
web: www.dcs.shef.ac.uk/∼jon

ABSTRACT

This paper presents a robust and computationally low-cost technique for tracking facial markers which exploits an adaptive marker collocation model to recover from tracking errors. Marker collocation statistics are estimated during periods where the markers are successfully tracked, and employed to estimate the position of missing markers during periods where the tracker fails to locate the full set. Evaluation experiments have been conducted on a small audio-visual corpus of connected digits in which the speaker was recorded with small white markers affixed to easily locatable points on the chin, lips, and nose. It is demonstrated that use of the marker collocation model makes the tracker robust in the face of marker occlusion.

1. INTRODUCTION

There is a great deal to be learnt about the interdependence of the acoustic and visual aspects of the speech signal [2]. One of the hindrances to studying this area is the lack of suitable data. The simple geometric visual properties that we might want to study, such as jaw position, mouth width etc, are notoriously difficult to extract from video. Although lip tracking is a well developed research area, state-of-the-art model-based techniques (such as those based on active shape and appearance modelling [4]) are often computationally expensive and are not sufficiently reliable to render precise and reliable ground-truth measurements of facial movements. For this reason a popular device is to avoid the problem by using easily trackable artificial markers that are physically attached to key points on the face of the speaker being recorded. If carefully designed a surprisingly small number of markers can be used to convey most of the relevant information present in the visual speech signal [5].

However, although markers can be accurately tracked for the most part using low cost algorithms, such algorithms will *occasionally* make unrecoverable tracking errors. This may happen for a number of reasons. For example, a marker attached to the upper lip boundary may disappear from view during lip protrusion; unless lighting is carefully controlled markers at the lip corners can be lost in shadow if the speaker turns his or her face slightly from the camera; tracking problems if there is an sudden unexpected movement (e.g. if the speaker coughs or sneezes) or if one or more markers are visually occluded. To ensure reliable tracking semi-automatic procedures may be employed in which a human operator monitors the performance of the tracker and intervenes where necessary. Not only is such an approach time consuming, the human element may introduce problems of consistency.

This paper proposes the use of marker collocation statistics as a means to impute the position of markers that are lost during tracking. The central idea is that due to the correlation in marker movements, if during tracking one or more markers are *known* to be missing in a given frame, then the positions of the reliably tracked markers can be employed to make a maximum-likelihood based estimate of the missing marker positions. Provided that these estimates are sufficiently reliable, the tracker will be able to 'pick up' the missing marker positions when they reappear in subsequent video frames. This idea is compatible with any tracking technique but is demonstrated here using a very simple and computationally inexpensive tracker.

The structure of the remainder of the paper is as follows. Section 2 describes the connected digit audio-visual speech data that has been employed for initial experiments. Section 3 presents the basic tracking technique. Section 4 explains the use of collocation statistics to impute missing marker positions. Experiments testing the robustness of the collocation statistics are presented in Section 5. The paper concludes with a brief discussion and the outline of plans for future work.

2. THE CONNECTED-DIGIT AV SPEECH DATA

Connected digit speech data was recorded from a single male native English speaker with a southern British accent. White, self-adhesive paper markers of roughly 3 mm square in dimension and with a dark central spot, were affixed to the face of the speaker at the positions indicated in Figure 1. The five markers around the lips and on the lower jaw were designed to capture speech information, whereas the markers on the nose bridge and nose tip are designed as fixed reference points. Two reference points are used, rather than one, to allow the possibility of some degree of head pose normalisation. The recordings were conducted in a quiet, day-lit room. The data was recorded in AVI format onto miniDV tape using a Canon MV650i digital video camera.

The speech material was composed of the following: i) 263 **three digit numbers** - these were generated by asking the speaker to read the last three digits of the timer that appeared on the video monitor; ii) 217 **six and seven digit telephone numbers** (the largest part of the corpus). The digit "0" was pronounced as "oh" throughout (as opposed to "zero"). In the telephone number section the subject had to read the numbers from a list. The subject was instructed to face towards the camera while looking down to read. Although awkward this seemed to work reasonably well introducing only a small change in head pose between parts i) and ii) of the data. In total the corpus consisted of 2200 digits spoken

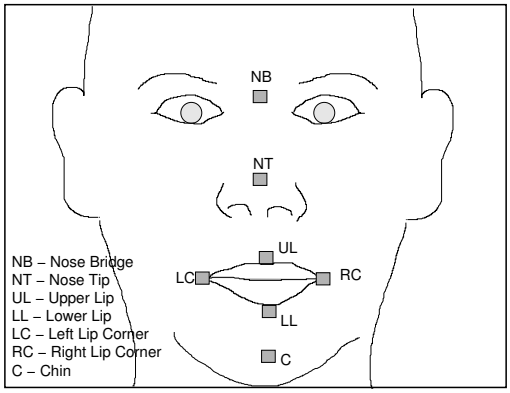

Fig. 1. The positions of the seven markers.

over a period of roughly 25 minutes in a single recording session.

The video was recorded at 25 frames/sec with half-frame interleaving and compressed using MPEG2 with a resolution of 720 by 480 pixels and a video data rate of 6,000 Kbits/s. Unless carefully applied, MPEG2 compression introduces significant artefacts, especially in moving high contrast image regions, such as the boundaries of the face markers (see Figure 2). One of the aims of this work was to develop techniques that work well with MPEG2 compressed data. Being able to work directly with compressed video is a great advantage when considering large multi-speaker corpora consisting of many hours of video.

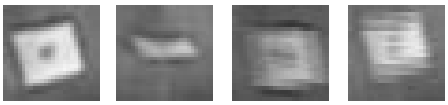

Fig. 2. Examples of the appearance of an MPEG2 compressed marker in various situations. From left to right: The UL marker while stationary; during lip protrusion; left-right head motion; up-down head motion.

3. TRACKING THE MARKERS

The technique developed for tracking the markers employs a simple, yet robust algorithm similar in operation to the CAMSHIFT face tracking algorithm [3]. The algorithm is initialised with the approximate dimensions, $l \times l$ pixels, of the marker to be tracked along with their positions in the first frame of video. Given the position in frame t the position in frame $t+1$ is estimated as follows. The tracker considers a region of interest (ROI) of dimension $(l+\Delta) \times (l+\Delta)$ pixels centred at the position of the marker in frame t. The luminance of each pixel in the ROI is calculated. The ROI is then converted into a black and white mask image by comparing the luminance to some threshold value, L_0. Pixels are set to 0 or 1 depending on whether their luminance is less than or greater than L_0 respectively. The threshold is tuned so that the marker appears as a region of 1's against a background of 0's. Finally the centre of the marker is calculated by computing the centre of mass of the mask image. In the instance where no pixels are labelled 1 the tracker returns the marker position as the centre of the region of interest.

With appropriate setting of the parameters Δ and L_0 this simple technique works surprisingly well. The value of Δ is a compromise. If the value is too small the region of interest may not capture the marker. This will happen if the marker moves more than $\Delta + l$ pixels either horizontally or vertically between frames. On the other hand, if Δ is large the ROI is more likely to contain bright objects other than the marker being tracked - if this happens the centre of mass of the ROI will no longer be a reliable indicator of the marker's centre. For the present data it was found that a value of 15 pixels was a good compromise. The luminance threshold, L_0 was tuned by displaying a binary thresholded image of the initial frame and slowly reducing L_0 until the whole of each marker was visible.

Naturally, this simple algorithm will occasionally break down. Problems occur when a marker either temporarily disappears from view or moves fast enough to leave the ROI. Fortunately, although the algorithm makes errors, it is generally possible to know when the errors have occurred. A simple tracking confidence measure can be computed based on the number of pixels, n, labelled as being above the luminance threshold. If the marker disappears from the ROI then the ratio, $\frac{n}{l \times l}$, will fall significantly below 1.0. If this ratio become small, tracker confidence is low, and the marker can be flagged as missing. Missing marker positions can then be estimated using collocation statistics as described in the next section.

4. ESTIMATING THE POSITION OF MISSING MARKERS

The position of the missing markers can then be estimated based on the position of the reliably tracked markers. This is a standard missing feature problem, solutions of which have been previously detailed for various statistical models in various domains - see for example [1].

A multivariate Gaussian marker collocation model is constructed from a segment of n frames of data that is known to be reliably tracked. The 2-D coordinates of the 7 markers in frame, t, can be represented by a 14 element feature vector, $\mathbf{x}_t$. Then the Gaussian marker collocation model can be written as,

$$f(\mathbf{x}_t) = \frac{1}{\sqrt{(2\pi)^N \mathbf{S}}} e^{-\frac{1}{2}(\mathbf{x}_t - \mu)' \mathbf{P} (\mathbf{x}_t - \mu)}$$

where N is the dimensionality of $\mathbf{x}_t$, the coprecision matrix, $\mathbf{P}$, is the inverse of the covariance matrix, $\mathbf{S}$. The mean vector, μ, and covariance matrix, $\mathbf{S}$, are estimated as,

$$\mu = \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t$$

$$\mathbf{S} = \frac{1}{n-1} \sum_{t=1}^{n} (\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)'$$

In any given frame the position of one or more of the markers may be unknown. Consider rearranging the elements of $\mathbf{x}_t$ so that it may be partitioned into present components, $\mathbf{x}_{tp}$, and missing component $\mathbf{x}_{tm}$, i.e. $\mathbf{x}_t = \{\mathbf{x}_{tp}, \mathbf{x}_{tm}\}$. (Note that for each missing marker there will be a pair of missing values in $\mathbf{x}$, i.e. both 2-D position coordinates.) The row and columns of the coprecision matrix, $\mathbf{P}$, can be likewise reordered so that,

$$\mathbf{P} = \begin{pmatrix} \mathbf{P_{pp}} & \mathbf{P_{pm}} \\ \mathbf{P_{mp}} & \mathbf{P_{mm}} \end{pmatrix}$$

and the mean vector, μ, is similarly reordered and partitioned,

$$\mu = \{\mu_p, \mu_m\}$$

By taking the derivative of $f(\mathbf{x}_t)$ with respect to $\mathbf{x}_{tm}$ and setting the result to zero and solving for $\mathbf{x}_{tm}$, it is shown that, given the present features, $\mathbf{x}_{tp}$, the maximum-likelihood estimate of the missing features, $\mathbf{x}_{tm}$, is,

$$\mathbf{x}_{tm} = -\mathbf{P_{mm}}^{-1}\mathbf{P_{mp}}(\mathbf{x}_{tp} - \mu_p) + \mu_m \quad (1)$$

So, if a frame occurs in which the tracking algorithm reports one or more marker positions to be missing, the coordinates of the missing markers are estimated using Equation 1 and tracking proceeds. For the tracker to recover from losing a marker, the markers at frame $t+1$ need to be founded within the region of interest that the tracker places around the estimated positions at frame t. This is only likely to occur if the missing marker positions have been accurately estimated.

5. EXPERIMENTS

The experiments reported here aim to establish the level of accuracy with which the missing markers can be estimated, and to measure how often large estimation errors occur.

5.1. Single Missing Markers

The first experiment examined the situation in which a *single* marker is missing. First, the complete corpus was reliably tracked using the basic tracking algorithm described in Section 3 along with human intervention to hand correct tracking errors as they occurred. The telephone numbers were split into two sets; a small set of 46 and a larger set of 171 utterance. The larger set was used as test data. Each of the seven markers in turn was considered to be missing at every frame. Using a marker collocation model the positions of the remaining six markers were used to estimate the position of the missing marker as described in Section 4. For each frame the distance between the actual tracked position of the marker and the estimated position was calculated. The mean of this distance was calculated across all frames. In order to test how well the estimation technique generalises, two different marker distribution models were employed. The first was calculated from the 3 digit string section of the corpus, and the second from the small set of telephone numbers. Results are shown in Table 1.

It can be seen that some markers are easier to estimate than others. Generally, the markers on the nose and upper lip are accurately predictable, whereas the lower lip, lip corners and chin markers are less so. The less predictable markers are those that exhibit a greater degree of movement relative to the position of the centre of the face. Tracking is generally successful for the distribution model trained on the telephone data, whereas errors are significantly larger for the model trained on the 3 digit sequences. As noted in Section 2, during the recording of the telephone data, the speaker is looking down to read the prompts and hence the head pose is slightly different. It can be seen that if a suitable collocation model is employed the errors are seldom very large, and will mostly be sufficiently small that the tracker can successfully proceed from the estimated positions.

The poor generalisation of the Gaussian marker collocation model is a significant point to consider in the design of the tracker. Better estimation results may be achieved by an online update of the models at regular intervals, e.g. estimating missing markers in one T second window of data with models based on the previous T second window. As the models are only updated every T seconds, the cost of recomputing the mean and covariance matrices will not be significant compared to the other frame rate processing. More crucially, the window needs to be sufficiently long to reliably train the models. For example, a short window may fall between utterances and hence contain no speech at all.

An online updated model was tested using the telephone number data with T set to 20 seconds - equivalent to about 6 telephone numbers. Results are shown in Table 2. It can be seen that the technique is effective in reducing the errors, with only the chin marker showing any frames with an estimation error greater than 10 pixels. Figure 3 provides some indication of the fidelity of the marker location estimation for three of the less predictable markers.

Trained 3-digit sequences, tested telephone							
Marker	NB	NT	UL	LL	C	LC	RC
Err (mean)	6.2	7.1	2.9	6.1	7.0	7.8	4.3
Err (sd)	1.0	1.4	1.7	1.8	4.8	2.5	1.5
% Error > 10	0	0.9	0.3	2.5	25	22	0
Trained telephone, tested telephone							
Marker	NB	NT	UL	LL	C	LC	RC
Error (mean)	2.8	2.9	2.0	3.7	3.7	2.0	1.6
Error (sd)	1.1	1.1	1.2	2.1	2.6	1.0	1.0
% Error > 10	0	0	0.1	1.2	3.5	0	0

Table 1. The mean and standard deviation of the marker estimation error measured in pixels when using either the 3 digit sequence (top) or the telephone numbers (bottom) as training data. The final row of each table shows the percentage of frames for which the estimation error is greater than 10 pixels.

20 second window training, tested telephone							
Marker	NB	NT	UL	LL	C	LC	RC
Error (mean)	0.9	0.8	1.4	2.4	3.4	1.7	1.8
Error (sd)	0.4	0.5	0.9	1.5	2.2	0.7	0.7
% Error > 10	0	0	0	0	0.2	0	0

Table 2. Results using the online updated model described in the Section 5.1.

5.2. Multiple Missing Markers

The next set of experiments used the online updated model technique, and tested the accuracy of marker estimation in situations where multiple markers are simultaneously missing. Table 3 shows again that the lower lip and chin markers are the hardest to estimate. The lower lip marker is particularly hard to estimate if the upper lip marker is also missing. However, even in this condition mean estimation error is still only 4.1 pixels. Finally, trials were conducted in which N markers, selected at random, were considered missing, and the reconstruction errors for each were recorded. Table 4 reports the overall average of the reconstruction errors collected from over 1,000 trials for each value of N. Note, even with 3 markers missing large errors are made on less than 1% of occasions. When all 7 markers are missing, reconstruction is equivalent to using the mean value of the training window which results in many large errors. Observing just a single marker may be sufficient to greatly improve the estimation.

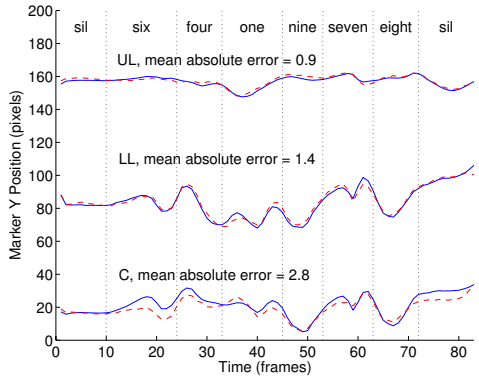

Fig. 3. Estimated marker heights (dashed) compared to actual values (solid) for the digit sequence "641978" using the 20 second windowing technique. Approximate word boundaries and mean absolute errors are indicated.

5.3. Testing the Full System

In the final set of experiments the full tracking system was evaluated both with and without the missing marker imputation component. In order to stress the tracker head movement and artifical visual occlusions were introduced. A 40 second segment of video was employed in which the head moves from side to side, up and down, and the subject speaks digit sequences. Either a vertically moving horizontal bar or a horizontally moving vertical bar was repeatedly scrolled across the image. The bar was 40 pixels wide and moved at a rate of 1 pixel a frame. The sequence was run 10 times for each direction with the bar at randomised initial positions. At the end of each run the number of markers lost by the tracker was counted. Results are shown in Table 5. In all but 2 of the 20 runs the tracker with imputation successfully locates all the markers at the end of the sequence. Demonstrations of the tracker running in this condition can be found on the author's website (www.dcs.shef.ac.uk/∼jon).

6. CONCLUSIONS AND FURTHER WORK

The paper has proposed a facial marker tracking system based on the combination of a simple tracking algorithm, and the use of marker collocation statistics to fix tracking errors as they occur. The tracking algorithm exploits the *temporal* continuity of the marker positions, and as such it is vulnerable to the temporary disappearance of markers. This weakness is compensated by the use of the highly correlated marker location statistics (i.e. *spatial* information). The paper demonstrates how by using complementary spatial and temporal information two simple techniques can be put together to form a robust whole.

The marker tracker has been developed in lieu of plans to record a large multi-speaker connected digit AV speech corpus. The initial tests, employing 30 minutes of MPEG2 compressed single speaker data, were conducted in part to test whether the planned set of recording conditions are sufficient for producing readily extractable marker tracks. The results are highly encouraging. Despite taking little care over optimising the lighting, and despite observable artefacts introduced by the MPEG compression, the tracking technique was proved sufficiently reliable to handle the resulting data.

Estimation errors for missing marker pairs							
	NB	NT	UL	LL	C	LC	RC
NB	—	1.9	0.9	0.8	0.8	0.8	0.9
NT	1.9	—	0.9	0.9	0.8	0.8	0.8
UL	1.4	1.8	—	2.4	1.5	1.7	1.7
LL	2.6	3.0	4.1	—	3.0	2.2	2.7
C	3.6	3.9	3.8	4.0	—	3.5	4.3
LC	1.6	1.8	2.1	1.6	2.6	—	1.8
RC	1.9	1.8	2.1	1.8	2.6	1.8	—

Table 3. Mean estimation errors when two markers are missing. Each row shows the estimation error for a given missing marker. The columns indicate the identity of the $2nd$ missing marker.

Estimation errors for N random missing markers							
N missing	1	2	3	4	5	6	7
Mean Error	1.7	2.1	2.5	3.1	3.9	4.9	7.4
% Error $>$ 10	0.0	0.2	0.7	1.9	4.2	7.9	16

Table 4. Mean estimation errors for multiple random missing markers. The final row shows the percentage of reconstructions for which the error is greater than 10.

Vertically moving bar occlusion										
Run	1	2	3	4	5	6	7	8	9	10
Baseline	5	3	3	5	5	3	2	1	3	1
+Imputation	0	0	0	4	0	0	0	0	7	0
Horizontally moving bar occlusion										
Run	1	2	3	4	5	6	7	8	9	10
Baseline	2	5	6	6	3	2	4	1	1	0
+Imputation	0	0	0	0	0	0	0	0	0	0

Table 5. The number of markers lost during tracking in each of the 10 runs with the moving horizontal and vertical occlusion.

7. REFERENCES

[1] S. Ahmed and V. Tresp, "Some solutions to the missing feature problem is vision" In *Advances in Neural Information Proc. Sys.,* Vol. 5, pp 393–400, Morgan Kaufmann, San Mateo, CA (1993)

[2] L. E. Bernstein, D. Burnham and J.-L. Schwartz, "Issues in audiovisual spoken language processing (when, where and how?)", in *Proc ICSLP*, Denver, Colorado, pp. 1445-1449 (2002)

[3] G. R. Bradski, "Computer vision face tracking for use in a perceptual user interface," In *Intel Technology Journal*, Q2, (1998).

[4] T. F. Cootes, G. J. Edwards and C. J. Taylor. Active Appearance Models. In *Proc. European Conf. on Computer Vision*, Vol. 2., pp. 484–498, (1998)

[5] L. D. Rosenblum and H. M. Saldaña, "An audiovisual test of kinematic primitives for visual speech perception," *Journal of Experimental Psychology: Human Perception and Performance* 22, pp.318-331, (1996)