Lip reading from scale-space measurements

Richard Harvey, Iain Matthews, J. Andrew Bangham and Stephen Cox
School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, UK.
{rwh,iam,jab,sjc}@sys.uea.ac.uk

Abstract

Systems that attempt to recover the spoken word from image sequences usually require complicated models of the mouth and its motions. Here we describe a new approach based on a fast mathematical morphology transform called the sieve. We form statistics of scale measurements in one and two dimensions and these are used as a feature vector for standard Hidden Markov Models (HMMs).

1. Introduction

Although the visual cues of speech alone are unable to discriminate between all phonemes ([b] and [p] for example) the incorporation of visual with acoustic information leads to a more robust recogniser. Degradation of one channel via, for example, interfering noise or cross-talk for audio, or occlusion for video, may be compensated, to some extent, by information from the other. In some cases information in each channel is complementary. For example, the phonemes [m] and [n] are acoustically similar but visually dissimilar.

Two problems have emerged in audio-visual speech recognition: audio-visual integration and visual feature extraction. The first problem has been addressed elsewhere [1, 20, 23]. Here, ways to extract visual features are examined. The aim is to reduce the dimensionality of the image to a small representative set of features. The usual approach is to assume a model for the lips that is described by as few parameters as possible and to fit the model parameters to the image sequence data [8, 10]. Unfortunately whether one uses active shape models [19], deformable templates [14] or dynamic contours [15], tracking the lips is a hard problem and furthermore such models remove the possibility of learning any other visual cues that may be significant.

Figure 1 shows an example of a parametric model (training data from a point distribution model (PDM) [17] trained over three talkers saying the letters: A, P, M, E, U and shows the variations corresponding to the largest three eigenvalues in a principal component decomposition of the shapes. The inner contour outlines the dark void bounded by lips and/or teeth. The outer contour corresponds to the outside margin of the lips. The first two modes of this sixteen-point model, which contain 70% and 17% of the variation, appear to show changes in the area and aspect ratio of the mouth. The third mode accounts for 2% of the total variation and appears, like the remaining modes, to be more related to variability in annotation of the training images.

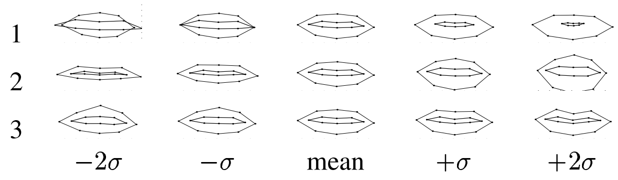

Figure 1. First three modes of variation at ± 2 standard deviations about the mean.

This observation leads to two questions. Firstly, were it possible to derive mouth area and aspect-ratio measurements directly from the grey-levels, how well would the system perform? Secondly, how much of the detail is lost due to this particular choice of model?

A desirable alternative is to discard the parametric model and attempt to extract features from the grey-level data directly [9]. This is the approach discussed in this paper but we seek to overcome the obvious problems of high-dimensionality and lack of robustness by adopting a scale-space approach. We examine the use of one- and two-dimensional scale-space primary vision systems to simplify the image and provide features that may be used to train a Hidden Markov Model. Section 2 describes the scale-space primary vision systems and in Section 3 a number of methods for extracting visual features are examined. Section 5 shows how these features might be combined with audio features to form an improved recogniser.

2. Scale-space primary vision systems

The properties of a scale-space decomposition are well known [16, 28] and have been closely connected with the properties of a diffusion system [18, 26]. However the scale-space properties are not unique to diffusion and there are several types of mathematical morphology systems that have some, if not all, of the desirable scale-space properties [4]. The systems used here are the one- and two-dimensional variants of types of *sieve* [3–7].

The sieve may be defined in any number of dimensions by defining the image as a set of connected pixels with their connectivity represented as a graph [13], $G = (V, E)$ where the set of vertices, V, are pixel labels and E, the set of edges, represent the adjacencies. Defining $C_r(G)$ as the set of connected subsets of G with r elements allows the definition of $C_r(G, x)$ as those elements of $C_r(G)$ that contain x.

$$C_r(G, x) = \{\xi \in C_r(G) | x \in \xi\} \tag{1}$$

Morphological openings and closings, over a graph, may be defined as

$$\psi_r f(x) = \max_{\xi \in C_r(G,x)} \min_{u \in \xi} f(u) \tag{2}$$

$$\gamma_r f(x) = \min_{\xi \in C_r(G,x)} \max_{u \in \xi} f(u) \tag{3}$$

The effect of an opening[1] of size one, ψ_1, is to remove all maxima of area one (γ_1 would remove minima of size one). Applying ψ_2 to $\psi_1 f(x)$ will now remove all maxima of area two and so on. The $\mathcal{M}$ and $\mathcal{N}$ operators are defined as $\mathcal{M}^r = \gamma_r \psi_r$ and $\mathcal{N}^r = \psi_r \gamma_r$. Sieves, and filters in their class such as alternating sequential filters with flat structuring elements, depend on repeated application of such operators at increasing scale. This cascade behaviour is key, since each stage removes maxima or minima of a particular scale. The output at scale r is denoted by $f_r(x)$ with

$$f_1 = \mathcal{Q}^1 f = f \text{ and } f_{r+1} = \mathcal{Q}^{r+1} f_r \tag{4}$$

where $\mathcal{Q}$ is one of the γ, ψ, $\mathcal{M}$ or $\mathcal{N}$ operators. An illustration of the sieve is provided elsewhere [4]. The differences between successive stages of a sieve, called *granule functions*, $d_r = f_{r+1} - f_r$, contain non-zero regions, called *granules*, of only that scale.

In one-dimension the graph, (1), becomes an interval

$$C_r(x) = \{[x, x + r - 1] \,|\, x \in \boldsymbol{Z}\} \tag{5}$$

where $\boldsymbol{Z}$ is the set of integers and C_r is the set of intervals in $\boldsymbol{Z}$ with r elements and the sieves so formed give decompositions by length.

[1] Openings and closings are duals so inverting the colour map swaps their operation – a common cause of confusion.

Figure 2. Two-dimensional decomposition of the image shown in the top left panel.

The sieves described here differ in the way they process extrema. Opening-sieves remove maxima and closing-sieves remove minima. The M- and N-sieves remove minima then maxima or maxima then minima. The algorithm operates by parsing the signal to build lists of extrema. Removal of extrema then amounts to merging lower length extrema sub-lists into higher ones. A further refinement is to process the maxima and minima as they occur. In one-dimension such removal is equivalent to recursive median filtering and the sieve so formed is called an m-sieve. The two-dimensional m-sieve uses the same algorithm but does not give a recursive median operation. Both inherit the ability to robustly reject noise in the manner of medians and do so more effectively than diffusion based schemes [12]. Moreover, the sieve-based algorithms have low order complexity and use integer-only arithmetic and so are highly appropriate for a real-time implementation described later. For these reasons, here, we discuss only results from the sieve.

For a p by q image there are potentially $p \times q$ granule functions so it is convenient to sum them into channels. This is shown in Figure 2. The top right panel shows extrema characterised by having an area between 50 and 100 pixels. Likewise, increasing areas, 100 to 200 and 200 to 2000, are shown in the lower panels. Greylevel 128 represents zero in the granularity domain and dark regions are negative granules. The mouth is particularly prominent in the lower right panel.

Figure 3 shows an example of the operation of a one-dimensional m-sieve. On the top is a plot of the intensity along a vertical scan line taken through the nose of the person shown in the top left panel of Figure 2. Beneath is a plot of the granule functions versus scale. The small-scale features appear as granules at lower scales. The absolute gran-

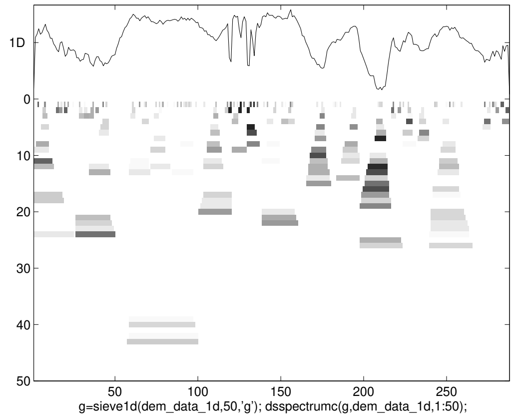

Figure 3. One-dimensional recursive median sieve decomposition for a vertical scanline (top trace) taken through the image of Figure 2.

ule amplitude is represented by greyscale density. Sharp-edged objects have a compact support in the scale-space.

3. Visual feature extraction

The granules are a mapping of the original image and contain all the information in the original [2, 4]. If pattern analysis is to be achieved in this domain it is helpful to reduce the overall dimensionality. The shape analysis in Figure 1 shows that the area of the centre part of the mouth accounts for most of the variance. This suggests the 2D sieve may be a useful tool for feature extraction.

Feature extraction method 1. Figure 2 indicates that the mouth may be isolated by a combination of approximate position within the face, and scale. The dark blob representing the mouth is particularly pronounced and therefore easy to track. In fact all the methods reported here work on sub-images of the type shown on the top of Figure 4, in which the mouth is the major feature. This is common practice in lip reading since full systems often include a separate head tracker [15]. The area is used as a single feature; it is an, easily calculated, alternative to fitting the second mode of the active shape model shown in Figure 1.

Feature extraction method 2. A possible alternative to the first two modes of the active shape model analysis might be estimates of the height and width of the mouth located in the previous method (lower right of Figure 2).

Feature extraction method 3. More information associated with the areas (but not shapes) of objects can be obtained from an area-histogram, see Figure 4 middle. This is a plot of the number of granules (ordinate) as a function of scale (abscissa). It is insensitive to both horizontal and vertical displacement of image features but, as shown in Figure 4, has variation as the lips move apart.

Figure 5 shows an utterance sequence "D-G-M". The top panel shows typical frames from the image sequence: a neutral frame and the centre frame from each utterance. The bottom panel shows the audio channel. The lower centre panel is the time varying area-histogram. The number of granules is plotted as an intensity, white represents a large number of granules and scale increases down the ordinate. The number of granules at a scale alters whenever the mouth opens and remains fairly stationary at other times. As the utterances are isolated letters the visual cues begin before, and end after, the acoustic signal.

Feature extraction method 4. The sieves defined in Section 2 may also be used to analyse a two-dimensional image by scanning in a given direction. In this case the vertical direction contains most of the lip motion that occurs during speech. The number of granules at a particular scale, a length histogram, indicates how many features of that length that the image contains and varies with the shape of the image features.

The top of Figure 4 shows images of a mouth and, at the bottom, their length histograms. These are formed by counting the number of granules in all vertical scan lines. This ignores their amplitude and gives improved insensitivity to lighting variation. The histogram of the granule count has a dimensionality equal to the number of analysis scales (sixty in this case). Its operation is best seen by viewing a motion sequence and Figure 5, upper centre panel, shows the time-varying scale histogram.

4. Data

A variety of sieve-based features were used on an audio-visual database consisting of ten talkers, five male (two with moustaches) and five female (none with moustaches) with each talker repeating each of the letters A to Z three times – a total of 780 utterances. Recording took place in a television studio under normal studio lighting conditions. Three cameras simultaneously recorded different views of the talker: full face, mouth only and a side view. All the recording was to tape. The full face images used here were recorded to SVHS quality. The output of a high quality tie-clip microphone was adjusted for each talker through a sound mixing desk and fed to all video recorders. Each talker was asked to return their mouth to the neutral position after each utterance and allowed to watch an autocue for the letters. No attempt at restraining was made but talkers were asked not to move their mouth out of the frame of the mouth-only camera.

All 780 full face sequences were digitised at quarter frame PAL (376×288) resolution and full frame rate (25Hz) using the standard frame grabber hardware of a Macin-

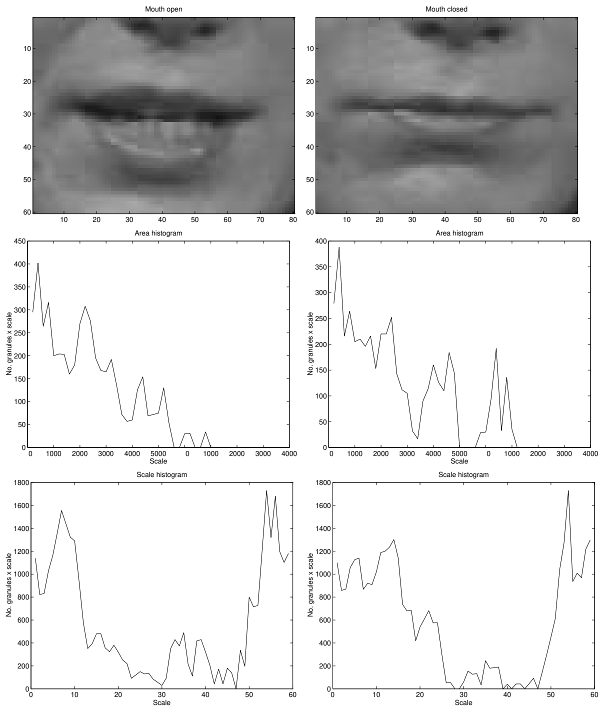

Figure 4. Open and shut mouths and their 2D (centre) and 1D (bottom) histograms.

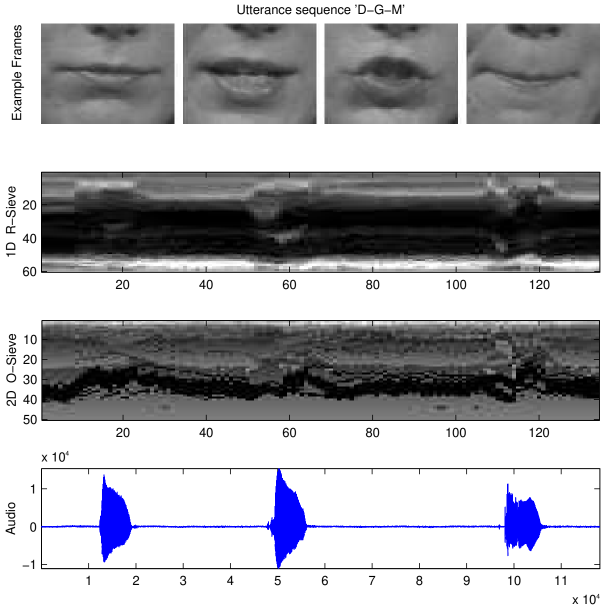

Figure 5. Utterance sequence "D-G-M". Top shows example image frames, upper centre the scale histogram, lower centre the area-histogram and bottom the waveform.

tosh Quadra 660AV (ITU-R BT.601 8-bit grayscale). Audio was simultaneously digitised using 16-bit resolution at 22.05kHz to ensure audio-visual time alignment. The database is available on four CD-ROMS.

Each utterance sequence was manually segmented using the video channel so that each image sequence began and ended with the talkers mouth in the neutral position. The audio data within this window was then hand labeled as silence–letter–silence.

5. Results

All recognition experiments were performed using the first two utterances from each of the ten talkers as a training set (20 training examples per utterance) and the third utterance from each talker as a test set (10 test examples per utterance). Classification was done using left to right HMMs [22], each state associated with a single Gaussian density with a diagonal covariance matrix. All HMMs were implemented using the HMM Toolkit HTK V1.4.

The results for feature extraction methods 1 to 3 were obtained using visual features at 40 ms frame intervals. Method 4 used an interpolated interval of 20 ms. Methods 1 and 2 used a five state HMM. Methods 3 and 4 used a ten state HMM.

Method	Best talker	Worst talker	All talkers
1	27	4	7
2	42	4	9
3	50	12	17
4	62	23	34

Table 1. Percentage of correct classifications.

The feature vectors used for the recognition experiments for each method were:

1. The area tracker, a scalar feature.

2. The two element feature vector formed from the height and width of the mouth estimate located in method 1.

3. A reduced area-histogram. The area-histogram shown in Figure 5 has fifty channels linearly spaced from area 1 to area 4800. This 50 element feature vector was reduced to 20 elements by computing the eigenvectors of the covariance matrix of the histogram channels and selecting the 20 principal components.

4. The one-dimensional recursive m-sieve scale histogram with 60 channels that were also reduced to 20 via principal component analysis.

Table 1 summarises the recognition performance of the four methods. The 'Best talker' column indicates the recognition results after training on two, and testing on one, utterance per letter from the single best talker for that method. All of the results in this column are considerably better than chance (4%) which, given the small training set, is encouraging. When a classifier is trained and tested on all talkers the performances drop (column four of Table 1). This indicates that some features work well with some talkers but not with others. Column three, the worst talker results, confirms this.

It is clear from Table 1 that as the number of features increases so does recognition accuracy. The best, method 4, may be combined with audio information to form an audio-visual speech recogniser.

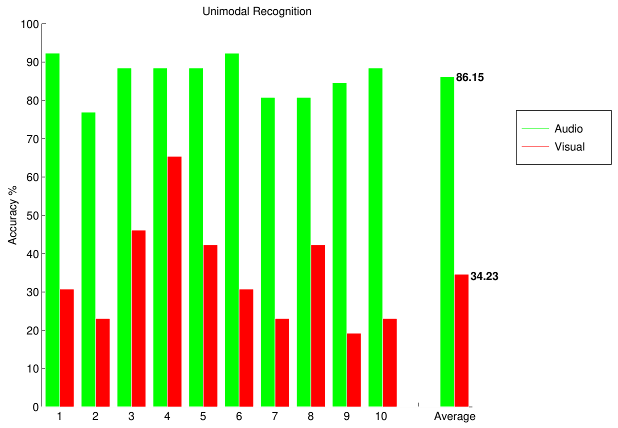

Figure 6. Unimodal results for each talker.

Figure 6 shows the results for method 4 (20 PCA coefficients calculated from a 1D scale histogram), the best visual case, on a per-talker basis with audio-only results for comparison. The audio features consisted of 12 Mel frequency cepstral coefficients (MFCCs) plus an energy term and delta coefficients, that is 26 coefficients, calculated at a frame rate of 20ms. The same HMM topology and training/test data was used for audio only as visual only tasks.

The variation in visual only performance (20–65% correct) is clearly much greater than that for audio only (77–92% correct). There is little correlation between poor audio and poor visual performance, further evidence that visual cues could improve audio recognition. However, combining the audio and visual information successfully is tricky and we describe this elsewhere [21].

6. Discussions

These results indicate that the scale histogram visual speech feature vector can be used successfully in lip reading. A recogniser using only visual information attained an average performance of 34% across ten talkers in a multi-talker, isolated letter recognition task (compared to 86% for audio only) and 62% when trained and tested for a single talker (compared to roughly 92% using audio only data). An informal test using a panel of three untrained humans on this talker using the same training and test data scored an average 57% correct.

A significant omission from this system is a method for normalising the scale. In Figure 4 the talker is at a constant distance from the camera. We are addressing this in a number of ways. Firstly, a head tracker allows normalisation by head size. Secondly we have found that horizontal granularity histograms contain little useful information apart from scale and thirdly by performing scale normalisation when the mouth is closed (minimum area in the histograms) with the advantage that it might control against variations in relative mouth size and shape.

The database used in these experiments is particularly challenging as it contains multiple talkers repeating 26 utterances under normal conditions. There has been no attempt to restrain the head motion of the talkers, the talkers did not wear lipstick and the sequences were collected at domestic frame rates and in greyscale. However it is difficult to know how well our method compares to others because all the research groups are using different data sets. We are tackling this problem in three ways. Firstly, by implementing alternative methods such as active shape models [11, 19]. Secondly, we are making our database publicly available and thirdly we are collecting more single-talker data so that we can assess our methods over this simpler task and compare them with others.

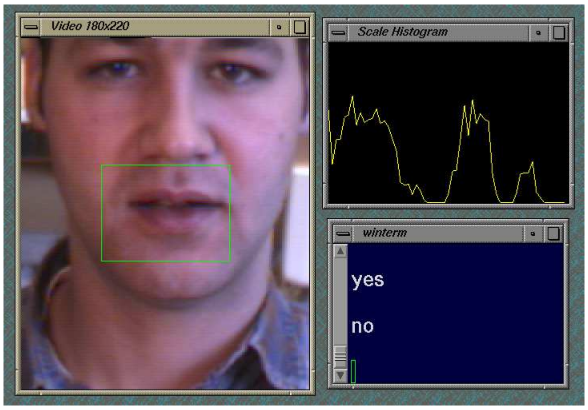

Figure 7. Screen from a prototype real-time lip reader

Current work is concentrating on how the area-histogram, scale histogram and active shape models may be combined to generate more robust features that will generalise better across talkers. This requires a larger database to permit reliable training of the HMMs or a real-time implementation. Figure 7 shows a frame from a real-time system implemented on a Silicon Graphics O2 workstation. The box delineates the region (60 by 80 pixels) that is sieved every full frame, 30 frames per second.

Future work will focus on finding effective ways of integrating the audio and visual information with the aim of ensuring that the combined performance is always at least as good as the performance using either modality [1,23,24,27].

References

[1] A. Adjoudani and C. Benoît. *On the Integration of Auditory and Visual Parameters in an HMM-based ASR*, pages 461–471. In Stork and Hennecke [25], 1996.

[2] J. A. Bangham, P. Chardaire, P. Ling, and C. J. Pye. Multiscale nonlinear decomposition: the sieve decomposition theorem. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18:518–527, 1996.

[3] J. A. Bangham, P. C. Chardaire, C. J. Pye, and P. D. Ling. Multiscale nonlinear decomposition: the sieve decomposition theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(5):529–539, May 1996.

[4] J. A. Bangham, R. Harvey, and P. D. Ling. Morphological scale-space preserving transforms in many dimensions. *Journal of Electronic Imaging*, 5(3):283–299, July 1996.

[5] J. A. Bangham, P. Ling, and R. Harvey. Scale-space from nonlinear filters. In *Proceedings of the Fifth International Conference on Computer Vision*, pages 163–168, 1995.

[6] J. A. Bangham, P. Ling, and R. Young. Multiscale recursive medians, scale-space and sieve transforms with an inverse. *IEEE Transactions on Image Processing*, 5:1043–1047, 1996.

[7] J. A. Bangham, P. W. Ling, and R. Harvey. Nonlinear scale-space causality preserving filters. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 18:520–528, 1996.

[8] C. Bregler, S. M. Omohundro, and Y. Konig. A hybrid approach to bimodal speech recognition. In *28th Annual Asilomar Conference on Signals, Systems, and Computers*, volume 1, pages 556–560, Pacific Grove, CA, Nov. 1994.

[9] N. M. Brooke. Using the visual component in automatic speech recognition. In *Proc. International Conference on Spoken Language Processing*, volume 1, pages 1656–1659, Philadelphia, PA, Oct. 1996.

[10] D. Chandramohan and P. L. Silsbee. A multiple deformable template approach for visual speech recognition. In *Proc. International Conference on Spoken Language Processing*, volume 1, pages 50–53, Philadelphia, PA, Oct. 1996.

[11] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. The use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12(6):355–366, 1994.

[12] R. Harvey, A. Bosson, and J. A. Bangham. Scale-space filters and their robustness. In *Proceedings of the First Internation Scale-space conference*, June 1997.

[13] H. J. A. M. Heijmans, P. Nacken, A. Toet, and L.Vincent. Graph morphology. *Journal of Visual Computing and Image Representation*, 3(1):24–38, March 1992.

[14] M. E. Hennecke, D. G. Stork, and K. V. Prasad. *Visionary Speech: Looking Ahead to Practical Speechreading Systems*, pages 331–349. In Stork and Hennecke [25], 1996.

[15] R. Kaucic, B. Dalton, and A. Blake. Real-time lip tracking for audio-visual speech recognition applications. In B. Buxton and R. Cipolla, editors, *Proc. European Conference on Computer Vision*, volume II of *Lecture Notes in Computer Science*, pages 376–387, Cambridge, Apr. 1996. Springer-Verlag.

[16] J. J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.

[17] A. Lanitis, C. J. Taylor, and T. F. Cootes. A unified approach to coding and interpreting face images. In *Proceedings of the Fifth International Conference on Computer Vision*, pages 368–373, 1995.

[18] T. Lindeberg. *Scale-space theory in computer vision*. Kluwer, 1994. ISBN 9-7923-9418-6.

[19] J. Luettin, N. A. Thacker, and S. W. Beet. Visual speech recognition using active shape models and hidden markov models. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 817–820, Atlanta, GA, May 1996. IEEE.

[20] I. A. Matthews, J. A. Bangham, and S. J. Cox. Audiovisual speech recognition using multiscale nonlinear image decomposition. In *Proc. International Conference on Spoken Language Processing*, volume 1, pages 38–41, Philadelphia, PA, Oct. 1996.

[21] I. A. Matthews, J. A. Bangham, and S. J. Cox. Scale based features for audiovisual speech recognition. In *IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication*, number 1996/213, pages 8/1–8/7, Savoy Place, London, Nov. 1996.

[22] L. Rabiner and B. Juang. An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, January 1986.

[23] J. Robert-Ribes, M. Piquemal, J.-L. Schwartz, and P. Escudier. Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition. In Stork and Hennecke [25], pages 193–210.

[24] P. L. Silsbee. Audiovisual sensory integration using hidden markov models. In Stork and Hennecke [25], pages 489–496.

[25] D. G. Stork and M. E. Hennecke, editors. *Speechreading by Humans and Machines: Models, Systems and Applications*. NATO ASI Series F: Computer and Systems Sciences. Springer-Verlag, Berlin, 1996.

[26] B. M. ter Harr Romeny, editor. *Geometry-driven diffusion in Computer vision*. Kluwer Academic, Dordrecht, Netherlands, 1994. ISBN 0-7923-3087-0.

[27] M. J. Tomlinson, M. J. Russell, and N. M. Brooke. Integrating audio and visual information to provide highly robust speech recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 821–824, Atlanta, GA, May 1996. IEEE.

[28] A. P. Witkin. Scale-space filtering. In *8th Int. Joint Conf. Artificial Intelligence*, pages 1019–1022. IEEE, 1983.