Performance Evaluation of Correlation Energy Filter Based Lip Biometric

M Haider Mehraj, Ajaz Hussain Mir
[1](Department of Electronics and Communication Engineering, College of Engineering and Technology, Baba Ghulam Shah Badshah University,Rajouri,India)
[2](Department of Electronics and communication Engineering, National Institute Of Technology, Srinagar, India.)

Abstract

In this paper identification is carried out using correlation energy filter and then using relative operating characteristics (ROC) and recognition time as performance metric for lip biometric. Correlation energy filter evolves from matched filters which are optimal for detecting a known reference image in the presence of additive white Gaussian noise. Correlation filters are attractive because of their characteristics such as shift-invariant, close form expression and high discrimination ability. Therefore, executing correlation filters provides an ideal application for the recognition process. The evaluation of this approach is carried out using relative operating characteristics curve and recognition time as metric on a digit database. The relative operating characteristics is a good approach for determining optimal system operating point whilst minimizing the false acceptance rate.

Keywords: AUROC; Correlation Energy Filter; Lip Biometric; Recognition; ROC.

1. Introduction

Biometrics parameters like face, lips, iris and other various parameters are being used for individual identification and verification, recent works in field has enabled similar recognition for various states or expressions automatically. Lip is one of the most important benchmark and advance parameter used for individual recognition of its varying postures [1].

Lips can be represented as an image of size p x q pixels and further can be represented by a vector in p.q dimension space. In practical applications lip as parameters for identifying individuals outperform other biometrics [2]. Images may be acquired from the distance without the knowledge of examined person. Lips biometrics is anatomical i.e. better results are expected than other behavioral biometrics. They are usually visible, may be implemented in hybrid face or voice recognition system.

The use of visual information specifically a person’s mouth region as a feature to recognition system has been reported. T Wark et al.[3] proposed a novel technique for the tracking and extraction of features from lips for the purpose of speaker identification. In their system, syntactic information is derived from chromatic information in the lip region. M.I Faraj et al [4] proposed the scheme and evaluation of a robust audio-visual digit-and-speaker-recognition system using lip motion and speech biometrics. The acoustic and visual features are integrated at the feature level and evaluated first by a Support Vector Machine for digit and speaker identification and, then, by a Gaussian Mixture Model for speaker verification. S A Dzati et al.[5], proposed method based on lip motion sequence and the Unconstrained Minimum Average Correlation Energy (UMACE) filter as a classifier for person identification. H.ECetingul et al.[6], addressed the selection of robust lip-motion features for audio-visual open-set speaker identification problem. They considered two alternatives for initial lip motion representation. In the first alternative, the feature vector is composed of the 2D-DCT coefficients of the motion vectors estimated within the detected rectangular mouth region whereas in the second, lip boundaries are tracked over the video frames and only the motion vectors around the lip contour are taken into account along with the shape of the lip boundary. H.ECetingul et al.[7], proposed use of explicit lip motion information, instead of or in addition to lip intensity and/or geometry information, for speaker identification and speech-reading within a unified feature selection and discrimination analysis framework.

In this paper recognition is based on lip images which are frame sequences of speaker mouth region. The frame sequences have been varied to determine the optimal number of frames at which recognition system should operate. In this approach, the entire region containing the speaker’s mouth (region of interest – ROI) is considered as informative for lip reading. This works on the principle of differentiating skin color from non-skin color i.e. a face detection algorithm is used. Lip localization is carried as a next step which is implemented using color thresholding.

Then features necessary for discrimination among different speakers are obtained from localized lip region. After the lips have been localized, the data generated is divided into two sets-training and testing sequences. The training sequence is to be synthesized with correlation filter. During the testing stage, for each filter, cross correlations with all lip motion images (testing) from the person to be identified is performed. The correlation output will produce Peak to side lobe ratio(PSR) values corresponding to the number images in testing set. The PSR values are obtained for each filter per person. If the PSR values corresponding to filter representing person to be identified is higher than the rest, the person is positively identified. The Relative operating characteristics curve is then used to determine the recognition rate as well as area under curve which tells about the classification problems addressed by system. The recognition time is the average time required to identify the speaker.

2. Correlation Filter

Correlation filter[5] evolves from matched filters which are optimal for detecting a known reference image in the presence of additive white Gaussian noise. Correlation filters are attractive because of their characteristics such as shift-invariant, close form expression and high discrimination ability [8]. Therefore, executing correlation filters provides an ideal application for the recognition process.

Correlation Energy filter is synthesized in the Fourier domain using closed form equations. The aim is to yield sharp peaks when the input image is authentic. In order to achieve this target, the filter strives to minimize the average correlation energy while maximizing the correlation output in the origin [9]. This optimization leads to the following filter equation:

$$U = D^{-1}m \qquad (1)$$

where D is a diagonal matrix with the average power spectrum of the training images placed along the diagonal elements and m is a column vector containing the mean of the Fourier transforms of the training images.

Peak-to-Side lobe ratio (PSR) metric is used to measure the sharpness of the peak. The PSR is given by:

$$PSR = (Peak - Mean)/\sigma \qquad (2)$$

Here, the peak is the largest value of the test image yield from the correlation output. Mean and standard deviation are calculated from the 20x20 side lobe region by excluding a 5x5 central mask [9]. The correlation filter for each person is designed by using several training images of his/her own lip images. The number of training images used to synthesize the filter depends on the variations among the training images. The test image is then cross-correlated for each designed filter. Fig. 1 illustrates the process for one person using lip motion images.

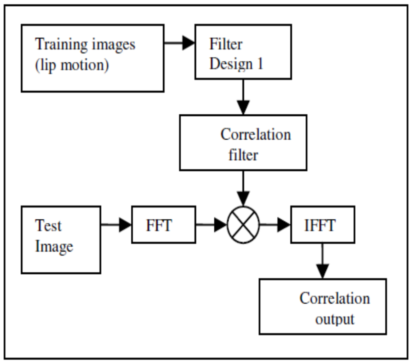


Fig. 1. Correlation Filter Process

3. Methodology

In this section, we shall describe in detail the lip identification process. The process can be divided into three main phases, face detection, lip localisation and feature extraction and identification process. The audio-visual Digit database developed by Sanderson was used for the experiments described in this paper [10]. The database consists of video and corresponding audio reciting digits zero to nine. The video of each person is stored as a sequence of JPEG images with a resolution of 512 x 384 pixels while the corresponding audio provided as a monophonic, 16 bit, 32 kHz WAV file.

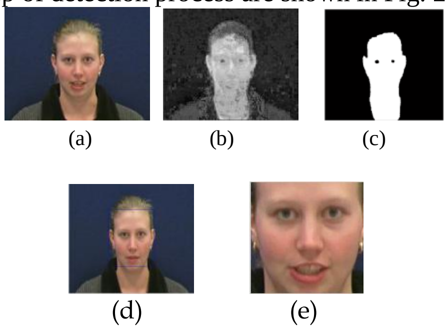
3.1 Face Detection

For the purpose of face detection color based and template matching technique is used. Summarily, the Gaussian model $N(\mu, c)$, with mean vector μ=E[x] and covariance matrix $C = E[(x-\mu)(x-\mu)^t]$ which are obtained from different skin color images are used to find out the skin likelihood for any pixel of an image. The skin likelihood is calculated as:

$$p(r,b) = \exp[-0.5(x-m)^t C^{-1}(x-m)] \qquad (3)$$

Where $x = (r,b)^t$ that is chromatic pair of red and blue.

The skin-likelihood image is transformed to the skin-segmented image (binary image). Since the skin regions are brighter than the other parts of the images, the skin regions can be segmented from the rest of the image through a thresholding process. To process different images of different people with different skin, a fixed threshold value is not possible to be found. Since people with different skins have different likelihood, an adaptive thresholding process is required to achieve the optimal threshold value. The adaptive thresholding is based on the observation that stepping the threshold value down may intuitively increase the segmented region. However, the increase in segmented region will gradually decrease (as percentage of skin regions detected approaches 100%), but will increase sharply when the threshold value is considerably too small that other non-skin regions get included. The threshold value at which the minimum increase in region size is observed while stepping down the threshold value will be the optimal threshold After optimal threshold has been set, all pixel values which have likelihood values higher than threshold are set to 1 and the rest of pixels are set to 0. Thus, resulting in binary image. A skin region is defined as a closed region in the image, which can have 0, 1 or more holes inside it. Since a real face region contains some unique features that differentiate the face from other skin regions, we exploit this characteristic by assuming that a real face, whether it is a frontal face or not, should contain at least one eye. Therefore, a region containing at least one region should be ruled out as a face region. The width and the height of the region is used to improve our decision process. The height to width ratio of the human faces is around 1. In order to have less misses however, we determined that a minimum good value is 0.8. Ratio values below 0.8 do not suggest a face since human faces are oriented vertically. The images obtained for person1 for each step of detection process are shown in Fig. 2.


(a) (b) (c)

.

(d) (e)

Fig. 2. (a) Original image. (b) Skin likelihood image. (c) Skin segmented image (d) Face region (e) Detected face region for person1.

3.2 Lip Localization

For the lip localization task, hue/saturation colour thresholding is implemented to differentiate the lips from the skin. The detection of the lip in hue/saturation colour is much easier owing to its robustness under wide range of lip colours and varying illumination conditions [11]. From the hue-saturation image, a binary image is then obtained by setting the threshold values, i.e. setting some specific threshold for hue and saturation. Lastly by employing morphological image processing, the lip region can be localized by finding the largest blob in the binary image. Lip Localisation process is shown below in Fig. 3.

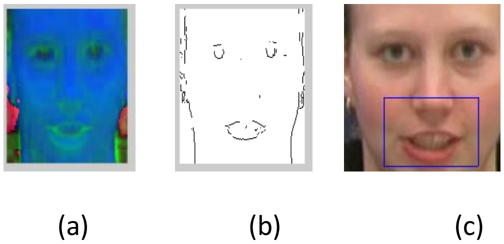

(a) (b) (c)

Fig. 3.(a) Hue Saturation Image. (b) Binary image (c) Localized lip region for person1.

3.3 Feature Vector and Identification.

In implementation we used twenty Correlation energy filter, one for each person. We have divided database into two sets-training and testing. We have two different sequences per person each consisting of 20 images for training and testing respectively. For testing cross correlation is performed with 10 or 20 or 30 or 40 lip images (depending on number of images taken during training).We have varied the number of training images to determine the minimum good number of frames at which identification would be successful. The correlation output will produce 40 PSR values for each filter in every case as number of testing images is 40.

Feature vector is the PSR values obtained for each person's case. It is calculated according to (2). The PSR values obtained for each sequence will be different depending on the person as well as the number of frames used during training. The PSR values are plotted against frame index (frame number) for various number of training images and then optimum number of frames to be taken for training can be easily decided from the curve. The PSR vs Frame index curve for 10,20,30 and 40 frames for the person1 used during training is shown in Fig. 4.

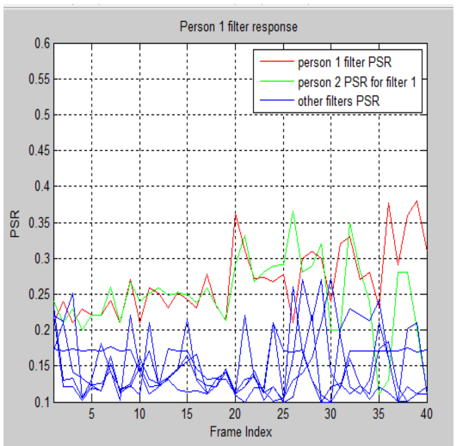


Fig. 4(a). PSR at 10 training images for person1.

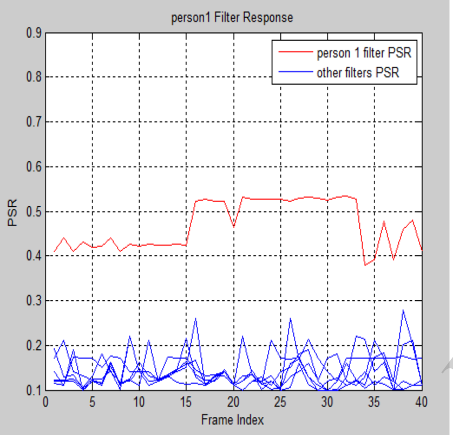


Fig. 4(b). PSR at 20 training images for person1

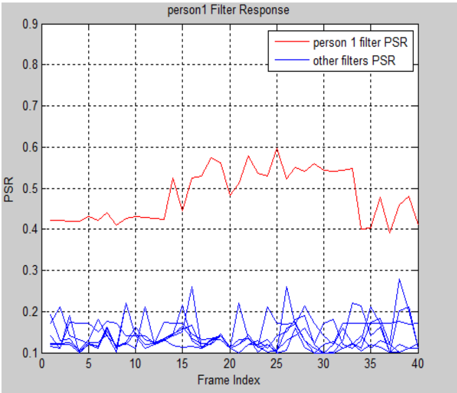


Fig. 4(c). PSR at 30 training images for person1

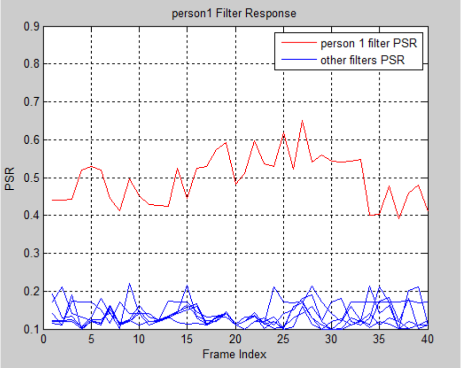


Fig. 4(d). PSR at 40 training images for person1

It is clear from the plot PSR values for person 1 for 10 frames(during training) are sometimes coinciding with other PSR values(which optimally should have been distinct). In curves for 20,30 and 40 frames respectively PSR values tend to be separated from rest of filters PSR values which means smooth filter response for these frame rates. Also the values do not show a steep change when we change number of frames from 20 to 30 to 40.

Person is positively identified if the filter corresponding to that person obtains the highest PSR value among all the filter responses. If PSR is lower than the person will be identified as imposter or wrongly as somebody else as in our implementation if we use only 10 training images, then in case of person 1, she will be wrongly identified as person 2.This is because Person 2's response (PSR value) for filter 1 (optimised for person 1) is quite close to person1. However if we choose 20 frames during training person1 will be positively identified.

4. Results

In a Biometric system for person recognition, the possible outputs are either positive, *p* (verified as the person in the system database) or negative, *n* (identified as someone who is not in the database). If the output is *p* and this person is the person in the database, the result is a true positive (TP); however, if this person is not really the person in the database, the outcome is a false positive (FP). Conversely, a true negative (TN) results when the output and the actual identity are *n*, and a false negative (TF) when the result is *n* but the actual identity is *p*.

The detection power (also called the true positive rate) is the fraction of all positives that are correctly classified as positive. The false rejection rate (FRR) is the fraction of all positives that are incorrectly classified negative. The false reject rate (FAR) is the proportion of all negatives that are incorrectly classified positive. The ROC is obtained by plotting True Positive rate against False Accept rate or against false reject rate. It can be also be obtained by plotting FAR against FRR. However it is not preferred because the Area under ROC (AUROC) which determines the accuracy can effectively be determined by plot of True Positive Rate against FAR. The AUROC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one For successful recognition AUROC should be greater than 50% and for good recognition it should be between (80-90)%. ROC can also be used determine the optimum threshold at which the system should operate by taking a point on true positive rate axis corresponding to ideal classifier.

Another Parameter that we have used is recognition time i.e is the time taken for the completion of recognition process. It may be important in certain application where time is an important factor such as access to schools by means of the biometrics.

The ROC curve is obtained when we use 10,20,30 and 40 training frames respectively. However recognition time is obtained for 10 and 20 frames only owing to the fact that we achieve good recognition rate at 20 frames. The ROC curve for 10,20,30 and 40 frames per person is shown in Fig 5.

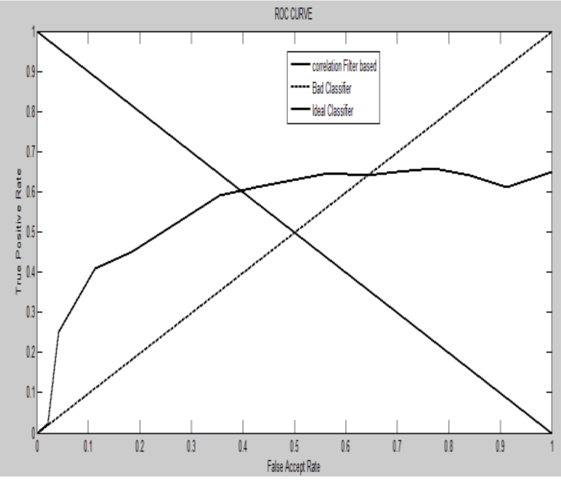


Fig. 5(a). ROC Curve for 10 frames per person.

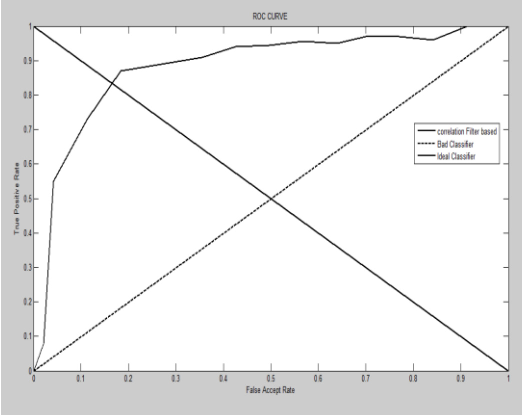


Fig. 5(b). ROC Curve for 20 frames per person.

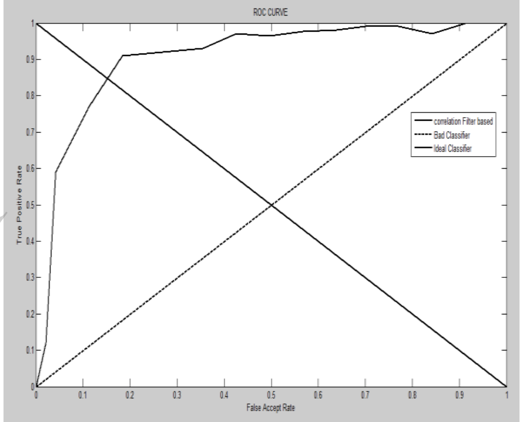


Fig. 5(c). ROC Curve for 30 frames per person.

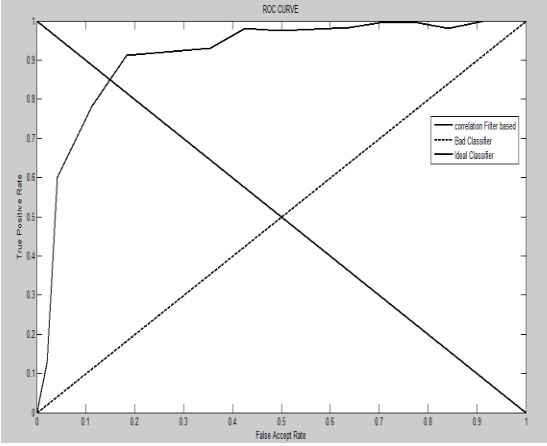


Fig. 5(d). ROC Curve for 40 frames per person.

When we used 10 frames per person (i.e 5 per sequence) for training, Area under ROC (AUROC) obtained is 51% while in case 0f 20, 30 and 40 frames per person, AUROC obtained is 81%,88% and 90% respectively.

Now if we select the cut off point i.e the point at which system will operate,it will be .59,.83,.84 and .84 for 10,20,30 and 40 frames per person. This is choosen because at this point ROC curve intersects the ideal classifier. Thus at 10 frames per person we will have recognition rate of around 60% but FAR will be around 40% which is too high. Such a low recognition rate is attributed to lack of the discrimination power of the feature vectors which is due to the fact change in pose and illumination is not duly incorporated in feature vector using 10 frames per person. In case of 20,30 and 40 frames per person recognition rate is around 84%. The recognition rate changes dramatically from 10 frames per person to 20,30 and 40 frames respectively because feature vector (PSR values) are discriminative for each person's case.The time in seconds is plotted against person number for 10 frames and 20 frames per person respectively as shown Fig. 6.

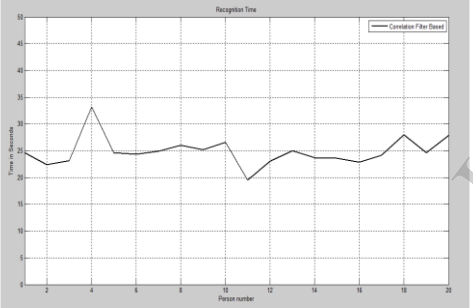


Fig. 6(a). Recognition time in seconds from person number 1 to person number 20 for 10 frames per person.

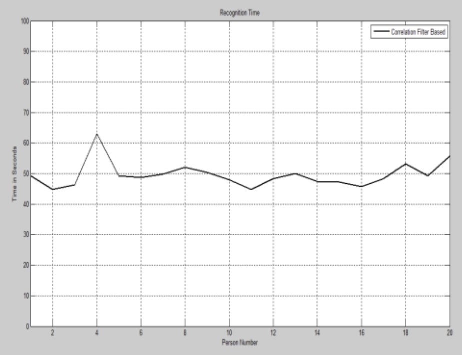


Fig. 6(b). Recognition time in seconds from person number 1 to person number 20 for 20 frames per person.

The recognition time almost doubles when number of frames per person is increased from 10 to 20 owing to the fact that number of frames in the training set also doubles.

5. CONCLUSION

This work carried out the performance evaluation of the correlation filter based approach using relative operating characteristics and recognition time as metric. The correlation filter approach performed worst when we used 10 frames per person for training because there was not enough discrimination among feature vectors obtained for different people. However it obtained recognition rate of 84% when 20 frames per person for training gave good discrimination because there was enough variations in pose, illumination and facial expressions at those frame rates. The recognition rate did not improve much as number of frames for training were increased from 20 to 30 or 40 owing to fact necessary discrimination among feature vectors was clearly visible using 20 frames per person.

The recognition time at 20 frames per person is between (50-60) seconds which can be highly useful in systems using fast access.

The recognition rate could be further increased by incorporating other related biometric techniques such as face.

REFERENCES

[1] Mok, L.L. Lau, W.H. Leung, S.H. Wang, S.L.Yan, H., "Person authentication using ASM based lip shape and intensity information," Image Processing, 2004. ICIP '04. 2004 International Conference on , vol.1, no., pp. 561- 564 Vol. 1, 24-27 Oct. 2004.

[2] Lievin, M. Luthon, "A hierarchical segmentation algorithm for face analysis. Application to lipreading," Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on , vol.2, no., pp.1085-1088 vol.2, 2000.

[3] T Wark and S Sridharan, " A Syntactic Approach to Automatic Lip Feature Extraction for Speaker Identification", IEEE International Conference on Acoustics Speech and Signal Processing 6, 3693–3696 (1998).

[4] Maycel-Isaac Faraj and Josef Bigun, " Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition", IEEE Transactions on Computers, Vol. 56, No.9, Sept. 2007.

[5] Salina Abdul Samad, Dzati Athiar Ramli, and Aini Hussain, "Person Identification Using Lip Motion

Sequence", Apolloni et al. (Eds.): KES 2007/WIRN 2007, Part I, LNAI 4692, pp. 839–846, 2007. © Springer-Verlag Berlin Heidelberg 2007.

[6] H.E.Cetingul, Y.Yemez, E.Erzin, and A.M.Tekalp, " Robust Lip-Motion Features For speaker Identification", IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP'05), Vol. 1, pp. 509-512, Philadelphia PA, USA, March 2005.

[7] H.E. Cetingul, Y. Yemez, E. Erzin, A.M. Tekalp "Discriminative Lip-Motion Features for Biometric Speaker Identification," IEEE Int. Conf. on Image Processing, Singapore, 2004.

[8] Savvides, M., Venkataramani, K., Vijaya Kumar, B.V.K,"Incremental Updating of Advanced Correlation Filters for Biometric Authentication System", Proceeding of ICME 3, 229–232 (2003).

[9] Savvides, M., Vijaya Kumar, "B.V.K, Efficient Design of Advanced Correlation Filters for Robust Distortion-Tolerant Face Recognition", Proceeding of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03), pp. 229–232 (2003).

[10] Sanderson, C., Paliwal, K.K., "Noise Compensation in a Multi-Modal Verification System", Proceedings of International Conference on Acoustics, Speech and Signal Processing, 157–160 (2001).

[11] Matthews, I., Cootes, J., Bangham, J., Cox, S., Harvey, R, "Extraction of visual features for Lip reading", IEEE Trans. on Pattern Analysis and Machine Intelligence 24(2), 198–213 (2002).