Threshold-based outer lip segmentation using support vector regression

Ashley Gritzman, Michiel Postema, David Milton Rubin, Vered Aharonson

Noname manuscript No.
(will be inserted by the editor)

Threshold-based outer lip segmentation using support vector regression

Ashley D. Gritzman · Michiel Postema · David M. Rubin · Vered Aharonson

Received: date / Accepted: date

Abstract Automated lip reading from videos requires lip segmentation. Threshold-based segmentation is straightforward, but it is rarely used. This study proposes a histogram threshold based on the feedback of shape information. Both good and bad lip segmentation examples were used to train an ϵ-support vector regression model to infer the segmentation accuracy from the region shape. The histogram threshold was optimised to minimise the segmentation error. The proposed method was tested on 895 images from 112 subjects using the AR Face Database. The proposed method, implemented in simple segmentation algorithms, reduced segmentation errors by 23.1%.

A. D. Gritzman[1,2]
E-mail: a.gritzman@eie.wits.ac.za
E-mail: ashley.gritzman@za.ibm.com

M. Postema[1,3]
E-mail: michiel.postema@wits.ac.za

D. M. Rubin[1]
E-mail: david.rubin@wits.ac.za

V. Aharonson[1,4,5]
E-mail: vered.aharonson@wits.ac.za

[1] Department of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

[2] IBM Research, Johannesburg, South Africa

[3] BioMediTech, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

[4] School of Sciences, University of Central Lancashire — Cyprus, Cyprus

[5] Afeka Tel Aviv Academic College of Engineering, Tel Aviv, Israel

Keywords Lip reading · support vector regression (SVR) · histogram threshold · shape-based adaptive thresholding (SAT)

1 Introduction

The shape and movement of human lips convey valuable visual information for automatic lip reading, emotion recognition, biometric speaker identification, and virtual face animation. Extraction of accurate visual information from the lips typically involves three stages: face detection; location of the region of interest; and lip segmentation.

Recent lip segmentation techniques employed machine learning and deep learning, demonstrating segmentation accuracies around 98% [1–4]. Deep learning methods are inherently computationally intensive, compared to colour-based segmentation. The low complexity of colour-based segmentation has resulted in a renewed interest [5].

Colour-based segmentation techniques alone are prone to false contour detection, due to low chromatic and luminance contrast between lips and skin [6].

Model-based segmentation techniques use previous knowledge of lip shape to construct a lip model by iteratively matching to an image and minimising a cost function. While usually invariant to translation, rotation, scale, and illumination, these techniques remain susceptible to variations in speaker appearance [7, 8]. Optimisation is also computationally expensive and it is affected by real-time performance [9].

Hybrid segmentation techniques combine elements from colour- and model-based categories [9]. They attempt to reduce the computational burden of model-based techniques by first obtaining a rough estimate

using colour-based techniques, and then fitting a lip shape model (LSM) [10].

This study proposes a a hybrid algorithm, referred to as shape-based adaptive thresholding (SAT). SAT acts as an extension to colour-based lip segmentation. The new algorithm constructs a model for the segmentation errors, rather than for the lip contours, and employs support vector regression (SVR) that uses this model to guide threshold selection.

2 Methods

The algorithm consists of two stages, both of which were implemented using MATLAB® (The MathWorks, Inc., Natick, MA, USA). The first stage is a one-time pre-training, displayed in Figure 1. The training is performed on a set of segmented-lips image pairs. One image in the pair contains the lips contours estimated by a colour-based lip-segmentation algorithm. The second image contains the ground-truth lips contours, which is usually a manually-marked outline. All images are binary, where the mouth region is white. The training produces a mapping of the image-pairs unto a segmentation errors graph. The graph includes a user-selected boundary line that separates the successful-segmentations region from the unsuccessful one. The second stage is an optimisation process, shown in Figure 2. Its inputs are the segmentation errors graph that was trained at stage 1 and a binary image after colour-based lip-segmentation. The algorithm infers the segmentation error of the input image using the segmentation-errors graph. If the segmentation error resides in the successful-segmentations region, the lips-contour estimation is accepted. If the segmentation error resides in the successful-segmentations region, the algorithm searches for a new threshold. The optimisation process iteratively employs the new threshold in the colour-based segmentation, infers the segmentation error, and searches for a minimum error.

2.1 Training of the segmentation error model

During stage 1, augmentation was employed by performing the threshold-based segmentation four times for each image in the training set. The four runs used four different thresholds and thus provided a larger training set and a wider range of performance. Each image was paired with a ground truth image for the lips contour. All images were coded into binary images containing the lip region in white and the skin region in black. A segmentation error was calculated for each input pair, between the threshold-estimated image and its ground truth image. The segmentation error values for the training set pairs were divided into three classes of perfect, good and bad segmentation regions. The threshold value between these regions could be arbitrarily set. The abels attached to the threshold-estimated images were “perfect”, “good” and “bad”.

Fourteen geometric features are extracted from the threshold-estimated images. Feature normalisation was performed to account for size and distance attribute which may be dependant on the physical characteristics of the subject, the proximity to the camera, and the zoom of the camera. These attributes were normalised using reference quantities. The features and their normalisation references are listed in Table 1. All features were centred and scaled by calculating their z-score. Linear discriminant analysis was used to reduce the feature set from the three classes into two eigenvectors.

Table 1 Features extracted from each IMAGE and their corresponding normalisation references.

Mouth shape feature	Reference
Area	IMAGE area
Perimeter	IMAGE perimeter
Centroid horizontal axis	IMAGE width
Centroid vertical axis	IMAGE height
Major axis length	IMAGE diagonal
Minor axis length	IMAGE diagonal
Bounding box horizontal axis	IMAGE width
Bounding box vertical axis	IMAGE height
Width	IMAGE width
Height	IMAGE height
Distance transform mean	IMAGE diagonal
Distance transform STD	IMAGE diagonal
Eccentricity	centre of IMAGE
Orientation	horizontal line

An ϵ-SVR model was trained on the datasets of 2-dimensional eigenvectors to estimate the segmentation error [11]. A radial basis function kernel was used, as well as the LIBSVM library [12]. To prevent overfitting of the model to the training data, a 10-fold cross-validation was performed, using the mean square error measure. The model parameters C and γ were optimised using a grid search approach with cross-validation [13].

2.2 Adaptive thresholding

The second stage was performed to optimise the segmentation thresholds. The input photos were subjected to colour-based thresholding. Feature extraction and reduction was performed on each image, after which

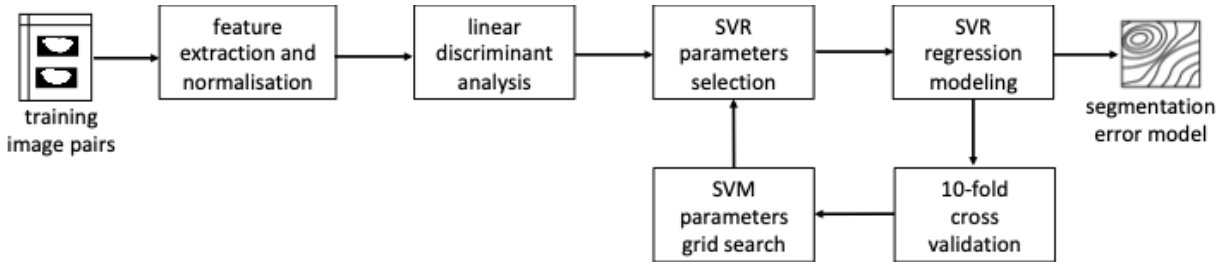

Fig. 1 Algorithm stage 1: Training of the segmentation error model.

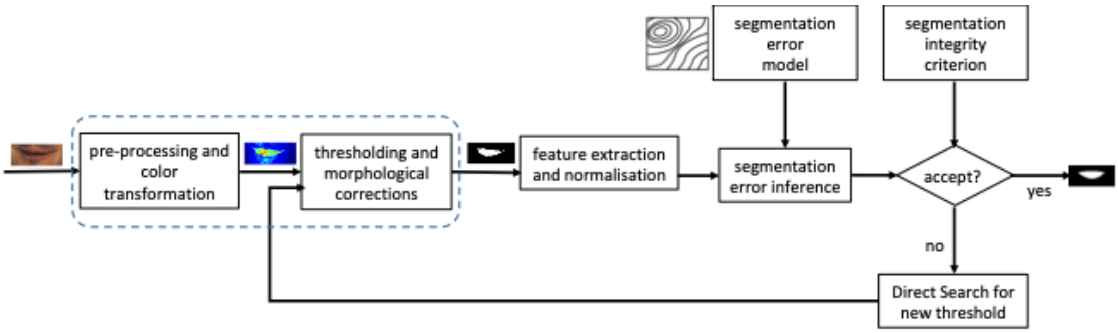

Fig. 2 Algorithm stage 2: Adaptive thresholding.

segmentation errors (SE) were inferred using the segmentation error model from stage 1. The inferred SE were compared to a segmentation criterion. If the inferred SE was lower than the SE criterion boundary, the candidate region was accepted and became the output segmentation. If the inferred SE was above the SE criterion boundary, the candidate segmentation was rejected and the thresholding was redone.

An optimisation objective function was defined by the thresholding and all subsequent image-processing steps performed during this stage. The objective function was minimised using Direct Search. A set of points was polled around a current threshold [14]. If and only if the poll was successful, *i.e.*, the objective function had a new value less than the previous step, the current threshold would replace the previous.

2.3 Evaluation

We used automatic histogram thresholding for the colour-based segmentation [10], which employed Otsu's discriminant, a nonparametric and unsupervised maximisation of the separability of the lip-pixels and non-lip pixels [15]. The dataset to train and test our algorithm included 895 images of 112 different subjects, from the AR Face Database [16] and their corresponding ground-truth images [17]. The ground truth had been obtained from manual markings of the lip contours by three independent human judges [17]. The markings of the outer lip contours included 20 points, while the markings of the inner lip contours included only 8. In this study, we concentrated on the outer lip markings. The markings of the outer lips were interpolated and coded into a binary image similar to the outputs of the thresholding algorithm.

The 112 database subjects were randomly divided into two groups: 60% in the training group, and 40% in the test group. The training dataset included 544 images and the test dataset included 351 images, from the subjects in the training group and test group, respectively.

The performance of the algorithm was evaluated using two analyses: SE interference performance and threshold optimisation performance. SE inference performance examined the reliability of the SE model, the output of stage 1. The accuracy of identifying poor segmentations for different criteria was quantified using receiver operating characteristic (ROC) curves. The ROC curves depicted the true positive (correctly accepted) segmentations rate vs. the false positive (incorrectly accepted) rate of the images in the test set. The true positive rate is defined as TP/(TP+FN), the false positive rate by FP/(TN+FP), in which FN is the number of lip pixels classified as skin pixels, FP is the number of skin pixels classified as lip pixels, TN is the number of skin pixels classified as skin pixels, TP is the number of lip pixels classified as lip pixels. Precision and recall were computed for each criterion.

Threshold optimisation performance examined the accuracy of the adaptive threshold. The criterion that produced the best trade-off between precision and recall was selected and employed. The algorithm performance was quantified for the 351 test images using two metrics: the segmentation error

$$\mathrm{SE} = \frac{\mathrm{FP} + \mathrm{FN}}{2 \times (\mathrm{TP} + \mathrm{FN})} \times 100\% \quad (1)$$

and the overlap

$$\mathrm{OL} = \frac{2 \times \mathrm{TP}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100\%. \quad (2)$$

SE yields the difference, in number of pixels, between the actual and ground-truth images, relative to the number of lip pixels in the ground-truth area [16]. OL is the percentage overlap between the segmented lip region and ground-truth lip contour. Perfect lip segmentation thus returns OL = 100% and SE = 0%.

3 Results and Discussion

Figure 3 illustrates the optimisation process. This example was chosen since the image is a challenging one, with facial hair, low contrast between lips and skin, and low illumination. The initial threshold segmentation at poll 1 had an SE of 44.8% and an inferred SE of 26.1%, qualifying it for optimisation according to the segmentation integrity criterion. At poll 2, the threshold was increased from 0.31 to 0.56, which resulted in the exclusion of part of the mouth after segmentation. This stringent threshold resulted in a high number of false negative lip pixels. The inferred SE improved decreased from 26.1% at poll 1 to 19.8%, corresponding to a successful poll. At poll 3, the threshold decreases back to 0.31, the inferred SE increased from 19.8% to 26.1%, indicating an unsuccessful poll. At poll 4, the threshold was set at 0.435, and the inferred SE dropped to 3.49%, corresponding to a successful poll. The segmentation image of poll (4) closely follows the lower lip contour and has minor deviations from the upper lip contour. From poll 5 on, the inferred SE of 3.49% from poll 4 did not decrease; thus the threshold of 0.435 was selected as final threshold value. The actual SE at poll (4) is 8.5%, which indicates a decrease of 81% compared to the initial SE of 44.8%

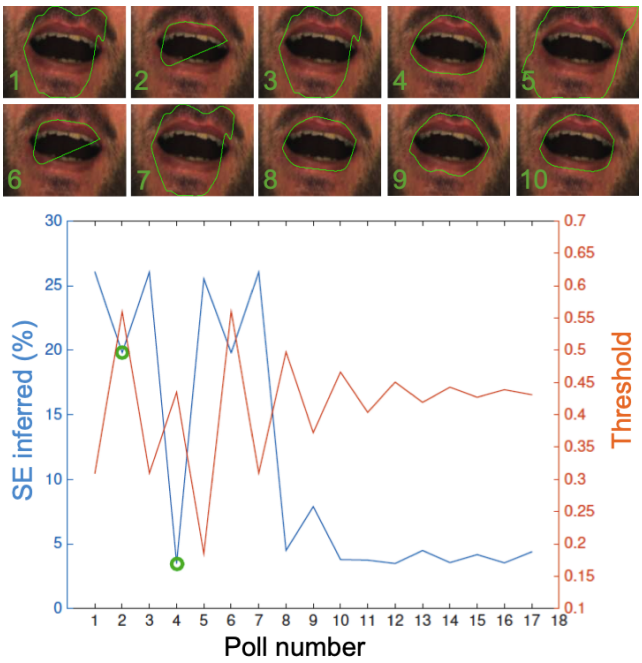

Fig. 3 An example of SE inferred and corresponding thresholds used, for 17 consecutive polls. Two successful polls are highlighted by a green ○. The contours computed after segmentation of polls 1–10 are represented by green lines, superimposed on the original lip photo.

Figure 4 shows four typical examples of photos that were selected for optimisation and decreased their SE after the optimised threshold selection. The optimised threshold visibly shows better matches of the outer lip contours than colour-based segmentation alone.

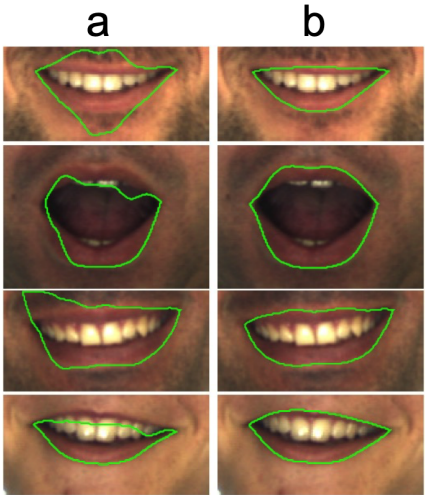

Fig. 4 Four typical examples of color-based segmentation (a) and optimised segmentation using the segmentation error model (b). The contours computed after segmentation are represented by green lines, superimposed on the original lip photos.

Figure 5 presents the ROC of the performance of all test images in stage 1 and three subsets thereof. The area under curve of 0.86 of all data demonstrates

good performance. True positive rates of 0.7 and less were correlated to low false positive rates of 0.2 and less. The subsets depict higher true positive rates correlating to these low positive rates. A tradeoff integrity

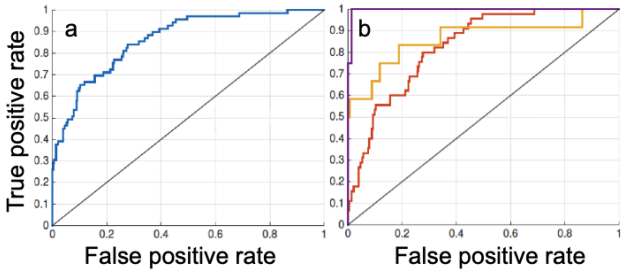

Fig. 5 ROC curves of all test images in stage 1 (a) and of three subsets of test images (b). The blue line represents all test images (AOC=0.86), the red line represents test images with SE between 10% and 15% (AOC=0.83), the yellow line represents test images with SE between 15% and 20% (AOC=0.87), and the purple line represents test images with SE greater than 20% (AOC=1.00).

criterion of 10% resulted in the performance depicted in the confusion matrix of Table 2. The matrix reveals that the SE inference failed to accept 28.4% of the images and accepted only 2.35% of the images that needed to be optimised. These properties are reflected in the performance metrics for this matrix, of 88.3% recall and 37.9% precision.

Table 2 SE inference confusion matrix of an SE decision boundary of 10%.

	Inferred pass	Inferred fail
Actual pass	51.9%	28.4%
Actual fail	2.3%	17.4%

The performance of the algorithm compared to colour-based thresholding alone is presented in Table 3. The table shows that without exception, the algorithm decreased SE and increased OL compared to colour-based segmentation alone. Especially the regimes of SE<5% and OL>95% demonstrate a drastic increase of the number images after running the algorithm.

The average SE of the baseline algorithm was reduced by 15.1% for the entire test set, and by 23.1% for the subset of images that were selected for optimisation by the segmentation integrity criterion. The difference stems from the fact that the full test set contains images that need no optimisation and their

Table 3 Cumulative frequency table containing SE cumulative distribution values and OL cumulative distribution values.

SE range	Thresholded (%)	Optimised (%)
$< 5\%$	11.3	27.4
$< 10\%$	61.0	73.8
$< 15\%$	86.3	90.5
$< 20\%$	92.9	96.4
OL range	**Thresholded (%)**	**Optimised (%)**
$> 95\%$	8.9	24.4
$> 90\%$	56.0	75.0
$> 85\%$	82.7	90.5
$> 80\%$	91.1	95.8

SE therefore did not change. A trade-off is involved in selecting the segmentation integrity criteria. The 10% criterion used in this study provided a recall of 88.3%, indicating a miss rate of 11.7%. This resulted in poor segmentations that did not continue to the optimisation stage, and hence did not improve by the optimisation. This miss-detection impacted the average accuracy of our algorithm. The partial ROC curves imply that the very poor segmentations, SE of 20% or more reidentified better than the moderately-poor segmentations of 15% and less. The precision of the acceptance criteria yielded that 37.9% of the images selected for optimisation actually need optimisation. This low precision does not necessarily reduce the accuracy of the algorithm, since unnecessary optimisations may even improve segmentation accuracy. The low precision does, however, incur a computational cost due to good segmentations undergoing unnecessary optimisation.

Our examples hint that our algorithm may be particularly efficient for images that present challenging conditions: facial hair, low contrast between lips and skin, and inconsistent illumination.

The proposed algorithm can be grouped under the hybrid colour-based and shape-based segmentation algorithms. The segmentation error model and the inferred segmentation error computation essentially use shape features. Comparison to previous hybrid segmentation and to machine learning lip segmentation algorithms is challenging. Different algorithms were employed on different image datasets, used different preprocessing and in some cases used different performance metrics. The average SE and OL of our algorithm for the subset of images selected for optimisation were 6.50% and 93.5%, respectively. These values are higher than the results of the hybrid approach published in [9] of 14.0% and 87.2%, respectively, and the results published in [6] of 8.40% and 90.80%, respectively.

The algorithm proposed in this study selected a new threshold for a colour-based discrimination of the lips from the skin area through minimisation of segmentation error. It trained a segmentation error model from a training set of labelled images and provides an inferred segmentation error for unlabelled images. A segmentation integrity criterion was imposed on the inferred SE and either accepted the colour-based segmentation or proceeded to new threshold selection. The current study evaluated the algorithm's performance in comparison to colour-based segmentation alone. Our algorithm might thus serve as an extension to any colour-based segmentation method.

4 Conclusions

An algorithm was designed to improve the segmentation accuracy of colour-based lip segmentation. The results portrayed a significant improvement in both segmentation error and overlap. It has been proven effective in challenging mouth image conditions including facial hair obscuring the lips, and low contrast between the lips and skin. The simplicity of the algorithm might enable fast computing for real-time applications.

Acknowledgements This study has been supported by the National Research Foundation of South Africa Grant Numbers 97742 and 127102. It has been based on a Doctoral project [18].

References

1. Zhang, X., Cheng, F. and Shilin, W., "Spatio-temporal fusion based convolutional sequence learning for lip reading," In: Proceedings IEEE/CVF International Conference on Computer Vision (ICCV), pp. 713–722 (2019)
2. Guan, C., Wang, S., Liu, G. and Liew, A. W., "Lip image segmentation in mobile devices based on alternative knowledge distillation," In: Proceedings 2019 IEEE International Conference on Image Processing (ICIP), pp. 1540–1544 (2019)
3. Guan, C., Wang, S., and Liew, A. W., "Lip image segmentation based on a fuzzy convolutional neural network," IEEE Transactions on Fuzzy Systems, Vol. 28, No. 7, pp. 1242–1251 (2020)
4. Nascimento, J.C. and Carneiro, G., "One shot segmentation: unifying rigid detection and non-rigid segmentation using elastic regularization," IEEE Transactions on Pattern Analysis and Machine Intelligence, in press
5. Spyridonos, P., Saint, A. F., Likas, A., Gaitanis, G., and Bassukas, I., "Multi-threshold lip contour detection," In: Proceedings 25th IEEE International Conference on Image Processing (ICIP), pp. 1912–1916 (2018)
6. Cheung, Y. M., Li, M., Peng, Q. and Chen, C. P., "A cooperative learning-based clustering approach to lip segmentation without knowing segment number," IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, No. 1, pp. 80–93 (2017)
7. Zheng, Z., Jiong, J., Chunjiang, D., Liu, X. and Yang, J., "Facial feature localization based on an improved active shape model," Information Sciences, Vol. 178, No. 9, pp. 2215–2223 (2008)
8. Wang, S. L., Liew, A. W. C., Lau, W. H. and Leung, S. H., "An automatic lipreading system for spoken digits with limited training data," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 18, No. 12, pp. 1760–1765 (2008)
9. Saeed, U. and Dugelay, J. L., "Combining edge detection and region segmentation for lip contour extraction" In: International Conference on Articulated Motion and Deformable Objects, pp. 11-20 (2010)
10. Gritzman, A. D., Aharonson, V., Rubin, D. M. and Pantanowitz, A., "Automatic computation of histogram threshold for lip segmentation using feedback of shape information," Signal, Image and Video Processing, Vol. 10, No. 5, pp. 869–876 (2016)
11. Smola, A. J. and Schölkopf, B., "A tutorial on support vector regression," Statistics and Computing, vol. 14, no. 3, pp. 199–222 (2004)
12. Chang, C. C. and Lin, C. J., "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), Vol. 2, No. 3, pp. 1–27 (2011)
13. Hsu, C. W., Chang, C. C. and Lin, C. J., "A practical guide to support vector classification," Technical report, Department of Computer Science, National Taiwan University (2003)
14. Lewis, R. M., Torczon, V. and Trosset, M. W., "Direct search methods: then and now," Journal of computational and Applied Mathematics, Vol. 124, No. 1–2, pp.191–207, (2000)
15. Otsu, N. , "A threshold selection method from gray-level histograms", Automatica, Vol. 11, No. 285–296, pp. 23–27 (1975)
16. Martinez, A. M., "The AR face database," CVC Technical Report, Vol. 24 (1998)
17. Ding, L. and Martinez, A. M., "Features versus context: an approach for precise and detailed detection and delineation of faces and facial features," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 11, pp. 2022–2038 (2010)

18. Gritzman, A. D., "Adaptive Threshold Optimisation for Colour-based Lip Segmentation in Automatic Lip-Reading Systems," PhD Thesis, University of the Witwatersrand, Johannesburg (2016)