Automatic dynamic template tracking of inner lips based on CLNF

Li Liu, Gang Feng, Denis Beautemps

AUTOMATIC DYNAMIC TEMPLATE TRACKING OF INNER LIPS BASED ON CLNF

Li Liu[1, 2] *Gang Feng*[1, 2] *Denis Beautemps*[1, 2]
(1) Univ. Grenoble Alpes, GIPSA-lab, F-38040 Grenoble
(2) CNRS, GIPSA-lab, F-38040 Grenoble

ABSTRACT

In this paper, a novel automatic approach to extract the inner lips contour of speakers without using artifices is proposed. This method is based on a recent facial contour extraction model developed in computer vision, called Constrained Local Neural Field (CLNF), which provides 8 characteristic points (landmarks) defining the inner lips contour. However, directly applied to our visual data including Cued Speech (CS) data, CLNF failed in about 50% of cases. We propose a Modified CLNF to estimate inner lips contour based on original CLNF landmarks. A dynamic template using the first derivative of smoothed luminance variation is explored in this new model. This method gives precise estimation of aperture for inner lips. It is evaluated on 4800 images of three French speakers. The proposed method corrects 95% CLNF errors and total RMSE of one pixel (i.e. 0.05cm in average) is reached, instead of four pixels using original CLNF.

Index Terms— CLNF, luminance variation, dynamic template, inner lips contour parameter, Cued Speech.

1. INTRODUCTION

Lips detection is an active research topic since lips (especially inner lips) hold significant information on speech production and play important parts in speech recognition based on visual features. Cued Speech (CS) is a complements for lips reading to enhance speech perception from visual input combining lips and hand. This work includes the CS case, which deals with hand occlusion for inner lips detection.

Several approaches for extracting lips contour in speech processing have been proposed in the literature. One of the most widely used robust technique is model-based lips detection. Active Shape Model (ASM) [1] and Active Appearance Model (AAM) [2] were proposed. Shape and appearance of lips are learned from training data with manually annotated faces and lips configurations are described by a set of model parameters. Bandyopadhyay [3] proposed a lip feature extraction technique combined with ASM and used contrast between the lips and the face to segment lips. Large training set and good initial condition are necessary for model-based technique. Stillitano *et al.* [4] proposed to use both active contours and parametric models for lip contour extraction. This method need prior knowledge of the lip shape. Another technique is based on segmentation in color spaces [5, 6]. Color-based clustering assumes that there are only two classes i.e. skin and lips, and this technique may not be true if facial hair or teeth exist.

CLNF was proposed by Baltusaitis *et al.* [7] in 2013, which is robust for facial landmark detection in the wild. It is a novel instance of Constrained Local Model (CLM) [8] that deals with the issues of feature detection in complex scene. CLNF learns the variation in appearance of a set of template region surrounding individual feature landmark and is more accurate than ASM, AAM and CLM. It replaces SVR patch expert of CLM by the Local Neural Fields (LNF) and uses Non-Uniform Regularised Landmarks Mean Shift as new optimization method. CLNF was trained on 4000 faces from independent databases HELEN, LFPW and Multi-PIE.

In this work, we deal with the extraction of inner lips from video recordings of the "natural" face (i.e. without help of artifices on lips). A new modification of CLNF with a dynamic template and Discrete Cosine Transform (DCT) filter to automatically extract inner lips in the case of word are presented. The dynamic template uses the first derivative of smoothed luminance variation to detect the outer lower lip and it provides more coherent result for adjacent image. To further improve the performance, we introduce an automatically lips region of interest (ROI) detector based on CLNF and a DCT filter indicator of closed lips. Finally, inner lower lip position is estimated by subtracting a statistical constant from the outer lower lip position. The proposed method is referred as Modified CLNF in this paper. Its performance is evaluated on a large set of images, and it has been shown that Modified CLNF provides effectively inner lips contour even in different context e.g. lips with hand occlusion, different lighting conditions.

2. RELATION WITH PRIOR WORKS

In prior works, lips were painted with blue color before the video recording to extract lips contours [9]. The inner lips were obtained by applying a single threshold in the "blue" color of RGB image [10]. The lips detection in [9] also contains CS case. To avoid using artifices on the lips, our previous work [11] on CLNF aims at detecting lip contour for annotated vowels which are simpler than annotated

words. Lip shape (closed lips, open lips, round lips and middle open one) changes more frequently and more complex in the case of words. Since the closed lips and slightly open lips can be similar and betweenness of two lips state is hard to recognize with time goes by. As a consequence, it is still challenging to automatically extract inner lips contour more efficiently. This work is devoted to automatically extract the inner lips contour without artifices in the case words.

3. DATABASE

The database contains recording videos of 50 French words which are numbers and daily words. Each corpus is uttered 10 times by 3 French subjects, one female CS speaker and two male speakers. The recording is made in a sound-proof booth in Gipsa-lab, France. Images are obtained every 20 ms. Words are annotated with Praat based on speech signal. For this work, we used the first repetition of each speaker. Thus, the corpus was made of 150 words and each word contains 32 images on average. In order to evaluate the performance of original CLNF model and the proposed solution, real inner lip contour was extracted manually by an expert using several landmarks placed on lips.

4. CLNF MODEL

CLNF contains three main parts: a point distribution model (PDM) of 68 points for face, LNF patch expert and Non-Uniform RLMS. From the 68 points, 12 landmarks are dedicated to describe the outer lips contour and 8 landmarks for inner lips contour. By applying PCA to the data, any lips can be estimated using sum of mean shape and variation part.

LNF can obtain relationships between neighbor and long distance pixels by learning both similarity and sparsity constraints. Neural network layer and convolution kernel are used to capture non-linear relationships between pixel values and the output responses. LNF gives a conditional probability distribution with probability density:

$$P(y \mid X) = \frac{\exp(U)}{\int_{-\infty}^{+\infty} \exp(U)dy}, X = \{x_1, ..., x_n\}, y = \{y_1, ..., y_n\} \quad (1)$$

where X is the observed input and y is the predict output. U is the potential function in linear combination of vertex features and edge features with different coefficients. All the parameters are estimated by maximizing condition log-likelihood of LNF. Considering different weights of patch expert for each landmark, Non-Uniform RLMS takes into account the patch expert reliability compared with original RLMS [12]. The aim is to minimize the following objective function:

$$\arg\min_{\Delta P} \left(\|P + \Delta P\|^2_{R^{-1}} + \|J\Delta P - v\|^2_W \right) \quad (2)$$

where P is a vector of output prediction parameter which is obtained from PDM and LNF. ΔP is the optimal parameter minus the initial estimation. W is a weight matrix which allows for weighting of mean-shift vector [7]. The vector v is the mean-shift vector over the patch response. The matrix J is the Jacobian of the landmark location for P . R is the matrix describing the prior on P .

5. PROBLEMS AND PROPOSED SOLUTIONS

CLNF is firstly straightforward applied to the database. In 68 facial points, 8 landmarks are used to describe inner lip contour. Spline smooth function is applied to generate a whole lip contour. The parameters A (width, expressed in cm), B (aperture, in cm) and S (surface in cm2) are calculated from the contour using the classic method for visual [10]. However, the main error of CLNF is the mistaken aperture. In this paper, we focus on correcting parameter B.

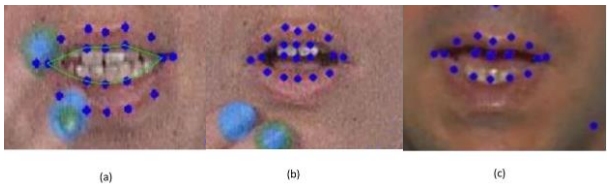

Fig. 1. Examples of 20 CLNF landmarks placed in the full lips region. Eight points describe inner lip contour. (a) Good inner lips contour of CLNF with hand occlusion. Green curve is lip contour obtained with linear interpolation for upper lip and spline interpolation for lower lip. (b) Mistaken CLNF landmarks for lower lip (tongue and teeth are visible). (c) Mistaken lower lip landmarks for another speaker with different lighting condition.

A comparison between original CLNF and expert values showed that CLNF obtained about 50% accuracy. But it remains robust even with part of lips occluded by hand. An example with hand occlusion is shown in Fig. 1a. However, the remaining cases have evident mistakes. In particular, the landmarks of the lower lip are badly placed with a much higher position (Fig. 1b) while no error is presented for upper lip. This phenomenon can be explained by the fact that the CLNF is based on a dictionary of training images. If the lips shape and appearance are not properly taken into account during the learning phase, it may lack the image during the optimization phase. The lower lip detection is challenging since lips area is often very complex (tongue and teeth may be visible) and lighting condition is variable (see Fig. 1b and 1c). The performance of CLNF model is shown in Fig. 6 (from top to bottom, each figure represent one subject) with one repetition uttered by each speaker. B parameter of expert is drawn by red curve and the B parameter obtained by original CLNF is blue. Their residual error is drawn in green. One can see that the residuals are extremely evident in some region, even over 10 pixels (0.5 cm). Generally negative errors means that B parameter is much smaller than true inner lips aperture. We found 30%

error for female CS speaker while errors are more than 50% for two other male speakers. RMSE shows an average value about 4 pixels (see Tab 1).

5.1. Dynamic template model

A post-processing procedure named Dynamic template model, which efficiently corrects CLNF errors is proposed. This method provides an estimation of inner position of lower lip by using the vertical luminance variation in a grid interval deduced by CLNF landmarks.

The luminance curve indicates lips edge where a relatively strong variation of the luminance can be found. It is calculated by the mean curve obtained from 22 luminance variations corresponding the landmarks of inner contour given by CLNF. The obtained curve is strongly noised in the original image and cannot be directly used. A suitable spline smoothing procedure is applied to give the first and second derivatives of the smoothed luminance variation curve. The smoothing degree should be carefully controlled so that the noise could be removed with minimum lost for useful information. Smoothing coefficient p = 0.01 is used since it gives a good compromise [13].

A dynamic template model corresponding to a typical variation around outer lower lip position is built. More precisely, this template is trained on a number of derivative curves reflecting different lips shape except closed lips. A very small template length makes results sensitive to noise while a very large length reduces detection precision. This length is set to 20 pixels experimentally. Another important factor is the interval in which the maximal correlation value is searched. Generally, the initial point of this interval is determined by the previous outer lower lip position so that the template can automatically learn the variation from image to image. The size of the searching interval is experimentally set to 16 pixels so that a rapid change of outer lower lip position could be taken into account. An example is illustrated in Fig. 2 by red curve with stars. In order to determine an optimal position of outer lower lip edge, a correlation value is calculated between current derivative curve and the template when the template scans through the searching interval. This method reduces the influence of noise, giving more coherent result for adjacent image.

In fact, inner lips detection is challenging since this area is fuzzy and several different cases have to be considered. For example, the luminance variation from teeth to lower lip is different with that from tongue to lower lip. The proposed method first detect outer lower lip position since the luminance variation in this region is less complicated than inner lower lip region. At this point, it remains to estimate the inner lower lip position based on outer lower lip position. We study the distance between the inner lower lip position determined by expert and the outer lower lip position estimated by the proposed method. A remarkable fact validated on 4800 images is found. This distance floats slightly around a constant for each speaker. The distance is 19.6 ± 1.1 pixels for the female speaker, and 19.2 ± 0.5 , 16.8 ± 0.5 pixels for other two male speakers respectively. This means the inner lower lip position can be estimated by subtracting a constant from the outer lower lip position. For any speaker, this constant can be obtained by training their data. The evaluation shows that this method gives satisfactory results. If the inner lower lip's vertical position obtained by Modified CLNF is less than that estimated by original CLNF, initial value is kept. Since CLNF landmarks properly describe lips shape, parallel translation with the same distance as the inner middle lip point is proposed to locate two other inner lower points.

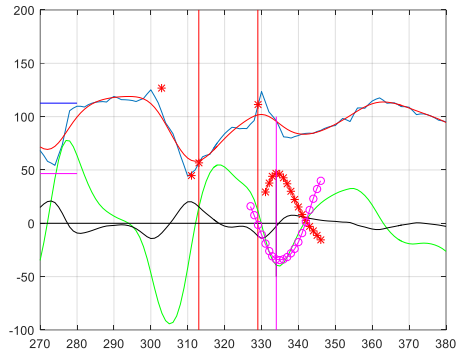

Fig. 2. Blue curve: original luminance variation. Red curve: smoothed luminance. Green curve: the smoothed first derivative. Black curve: the smoothed second derivative. Red curve with stars: correlation values between the template and the first derivative curve in function of the position of the template. Magenta curve with stars: Template.

5.2. Closed lips filter based on DCT analysis

CLNF error makes upper and lower inner lips points overlapped for open lips. It may be confused with closed lips (Fig. 3). CLNF performs perfect for closed lips while dynamic template is not suitable in this case. In order to eliminate closed lips and keep good result of original CLNF, closed lips filter based on DCT coefficients is presented.

Lips ROI is determined by the 20 landmarks of CLNF which efficiently delimit lips region and determine a precise center of this region. A suitable-sized ROI is determined from this center (Fig. 4). The ROI size is 140x95 pixels. We compared them with the ROI determined by a blue mark in speaker's front and it shows a precise ROI. Ten closed lips

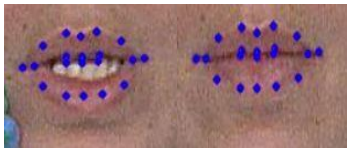

Fig. 3. Mistaken CLNF landmarks (left). Correct CLNF landmarks for closed lips (right). Note that CLNF landmarks are strongly similar in these two cases.

images are used for training. DCT coefficients are calculated from lips ROI and 10x10 coefficients in the low-frequency region are retained. Mean vector of these DCT coefficients is considered as a model for closed lips. By applying this model to all images, a threshold that distinguishes closed lips and open lips is obtained. For each new image, DCT coefficients of ROI are first calculated and then compared with the model. Image with distance less than the threshold is regarded to be closed lips, which will not be processed by dynamic template model.

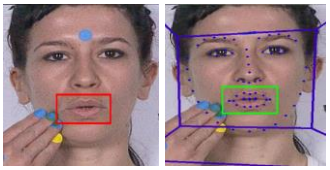

Fig. 4. Lips ROI determined by a blue mark in the front of the speaker (left). Lips ROI (same size) determined by a center point estimated from CLNF landmarks (right).

6. EVALUATION AND RESULTS

Modified CLNF method efficiently corrects main errors of CLNF. It can be visually shown in Fig. 5.

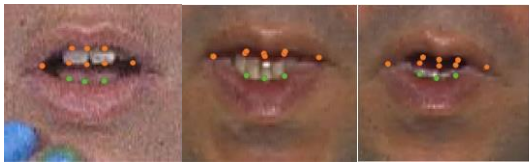

Fig. 5. Examples showing initial mistaken CLNF landmarks (red points) and corrected landmarks using the proposed method (green points) for inner lower lip.

Modified CLNF is evaluated by comparing the B parameter estimated from the inner lower lip position with that obtained by expert. In Fig. 6, the corrected B parameter is drawn by black curve, and their residual error is drawn magenta. One can see that errors are significantly reduced with respect to the original CLNF errors (green curves).

RMSE is calculated for these errors and results are shown in Tab. 1. Note that total RMSE of the final errors is reduced to 1 pixels (0.05 cm), instead of 4 pixels (0.20 cm) with original CLNF.

Tab. 1. RMSE values for original CLNF model and for Modified CLNF, expressed in pixels and in cm.

RMSE	Speaker 1	Speaker 2	Speaker 3	Total
CLNF	3.84 (0.20 cm)	4.02 (0.21 cm)	3.53 (0.18 cm)	3.81 (0.20 cm)
Modified CLNF	1.13 (0.06 cm)	0.90 (0.05 cm)	0.94 (0.05 cm)	0.99 (0.05 cm)

7. CONCLUSION

This paper presents a new automatic method for extracting

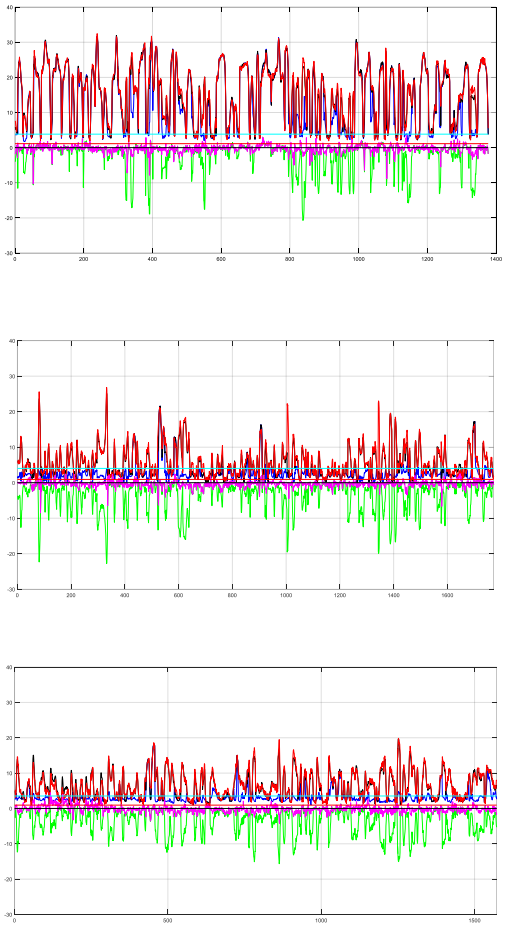

Fig. 6. X-axis is the number of images, y-axis means pixels unit. Red curve: Expert determined B parameter (expressed in pixels). Blue curve: B values by original CLNF landmarks. Black line: B values by proposed method. Green curve: errors of original CLNF. Magenta curve: errors of Modified CLNF. The female speaker (top), other two male speakers (middle and bottom).

inner lips contour. Experimental results demonstrate the efficiency of the proposed approach, even when part of lips are occluded by hand. This method can correct 95% CLNF errors and give precise estimation of B aperture values for inner lips contour. An efficient post-processing method named dynamic template searching based on the largest correlation is proposed in this paper. Another significant point of the proposed method consists of first estimating outer lower lip position in order to locate inner lower lip position. The evaluation of the proposed method on about 4800 images of three speakers confirms the performance. Modified CLNF outperforms original CLNF for inner lip detection in sense of CS processing. RMSE decreases from 4 pixels (0.2 cm) to 1 pixel (0.05cm). Future work involves to build a round lips shape model to correct lip width (parameter A) for small part of CLNF errors to realize full speech recognition based on visual.

8. REFERENCES

[1] Tim Cootes, ‘‘An Introduction to Active Shape Models,’’ *Model-Based Methods in Analysis of Biomedical Images in* ‘‘*Image Processing and Analysis*’’, Oxford University Press, pp. 223-248, 2000.

[2] T.F. Cootes, G.J. Edwards and G.J. Taylor, ‘‘Active Appearance Model,’’ *Proc. European on Computer Vision*, vol.2, pp. 484-498.1998.

[3] Samir K. Bandyopadhyay, ‘‘Lip Contour Detection Techniques Based on Front View of Face,’’ *Journal of Global Research in Computer Science*, vol. 2, No. 5, 2011.

[4] Stillitano S., Girondel V., and Caplier C., ‘‘Lip contour segmentation and tracking compliant with lip-reading application constraints,’’ *Machine Vision and Applications*, vol. 24, Issue 1, pp. 1-18, 2013.

[5] Evangelos Skodras and Nikolaos Fakotakis, ‘‘An unconstrained method for lip detection in color images’’, *ICASSP*, 2011.

[6] Jian-Ming Zhang, Liang-Min Wang, De-Jiao Niu, and Yong-Zhao Zhan, ‘‘Research and implementation of a real time approach to lip detection in video sequence,’’ *IEEE. Int. Conf. on Machine Learning and Cybernetics*, 2003.

[7] Baltrusaitis T., Morency L.-P., and Robinson P. ‘‘Constrained local neural fields for robust facial landmark detection in the wild,’’ IEEE, *Computer Vision Workshops (ICCV-W)*, Sydney, Australia, 2013.

[8] Cristinacce D. and Cootes T., ‘‘Feature detection and tracking with Constrained Local Models,’’ *Actes de British Machine Vision Conference*, vol. 3, pp. 929-938, 2006.

[9] Panikos Heracleous, Denis Beautemps and Noureddine Aboutabit, ‘‘Cued Speech Automatic Recognition in Normal Hearing and Deaf Subjects,’’ *Speech Communication*, vol.52, Issue 6, pp. 504-512, 2010.

[10] Lallouache T., ‘‘Un poste Visage-Parole. Acquisition et traitement des contours labiaux,’’ *Actes des Journées d’Etudes de la Parole*, Montréal, 1990.

[11] Li Liu, Gang Feng, and Beautemps D., ‘‘Extraction automatique de contour de lèvre à partir du modèle CLNF’’, *Actes des Journées d’Etudes de la Parole*, Paris, 2016.

[12] Saragih, J., Lucey, S. and Cohn, J., ‘‘Deformable Model fitting by Regularized Landmark Mean-Shift,’’ *International Journal of Computer Vision* (*IJCV)*, 2011.

[13] Feng G., ‘‘Data Smoothing by Cubic Spline Filters,’’ *IEEE Transactions on Signal Processing*, vol. 46, pp. 2790-2796, 1998.