Finding Speaker Face Region by Audiovisual Correlation

Yuyu Liu [†,‡] Yoichi Sato [†]

†Institute of Industrial Science, The University of Tokyo
‡System Technologies Laboratories, Sony Corporation

Abstract. The ability to find the speaker face region in a video is important in various application areas. In this work, we develop a novel technique to find this region robustly against different views and complex backgrounds using gray images only. The main thrust of this technique is to integrate audiovisual correlation analysis into an image segmentation framework to extract the speaker face region. We first analyze the video in a time window and evaluate the audiovisual correlation locally at each pixel position using a novel statistical measure based on Quadratic Mutual Information. As only local visual information is adopted in this stage, the analysis is robust against the view change of the human face. Analyzed correlation is then incorporated into Graph Cut-based image segmentation, which optimizes an energy function defined over multiple video frames. As this process can find the global optimum segmentation with image information balanced, we thus can extract a reliable region aligned to real visual boundaries. Experimental results demonstrate the effectiveness and robustness of our method.

1 Introduction

The ability to detect a speaker accurately from multiple persons is important for various applications in video processing and content analysis systems. For example, a video-teleconferencing system may need to focus on a speaker, or a video analysis system may have to associate uttered words with the right speaker. Being able to identify the speaker face region is furthermore preferred because this makes various effects possible, such as to automatically emphasize the speaker by blurring all other persons and background, or, on the contrary, to impose mosaic over the speaker in an interview to protect privacy. An example of them is shown in Figure. 1 based on the results of our method.

However, regardless of the great progresses in face and human detection (e.g., [16] and [3] respectively) made in recent years, speaker detection is still under development. As the purpose is to distinguish a person from others, face or human detection fails in this area. One solution [10] is to detect the face, locate the mouth, and check its movement. The weakness of this method is the requirement of a frontal view. In fact, the problem of view dependency is also a challenge even in face and human detection systems [16] [3]. Additionally, as the

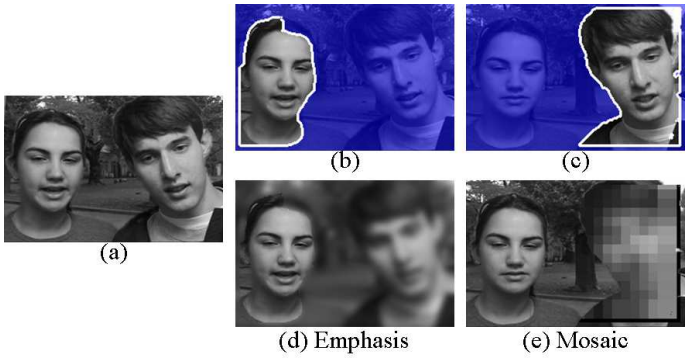


Fig. 1. Localized speaker face regions and application demonstrations. (a) is the original image. (b-c) show the speaker face region localized by our method. (d-e) show the imposed special effects based on our results.

correlation between the mouth movement and audio signal is not well analyzed, it is prone to be disturbed by unconscious movements from other persons.

In this work, we develop a novel technique to find the speaker face region from gray images of a video in a time window, which is robust against different views and complex backgrounds. This technique is based on the recent developments in audio source localization by audiovisual correlation analysis and an image segmentation technique over multiple video frames.

To localize an audio source by audiovisual correlation analysis is a relatively new research topic and has drawn much attention in recent years. Psychological research [5] discovered that audiovisual correlation mainly lies in the synchrony. Initiated from it, Hershey and Movellan [6] first introduced a method to localize the audio source by audiovisual correlation analysis. They assumed that audio energy and pixel intensity obey joint normal distribution and measured their *Mutual Information* (MI) to decide the audio source. Yet, this assumption is too strong and contradicts with the real observation. To avoid this problem, following researches tried another way, where they optimize an objective function to find the audio source. Darrell and Fisher [4] searched the optimum projection vector that can maximize the lower bound of MI between the projected audiovisual signals. Kidron and Schechner [8] searched a projection vector which can maximize the *Canonical Correlation Analysis* (CCA) between audiovisual signals and has minimum $L1$-norm. In 2005, Monaci *et al.* [11] claimed that the movements of the photographed objects convey better audiovisual correlation than the pixel values. They employed *Matching Pursuit* (MP) to extract local objects, tracked their translations and rotations in the video and computed the correlation with audio to decide which one should be the audio source.

Unfortunately, all methods mentioned above suffer a common problem: the estimated audio source is highly fragmental. Therefore, they in fact experience difficulties in designating a correct speaker position, much less identifying a

reliable speaker region. Most of them can detect only pixels that are supposed to be the audio source. Only Casanovas [2] clustered the estimated pixels and adopted the cluster center as the speaker position. Yet clustering is vulnerable to the outliers that appear often.

To be able to detect a reliable speaker face region, we consider a novel technique, whose key idea is to integrate audiovisual correlation analysis into image segmentation framework to extract out the region. Our current system requires that the speaker must stay nearly at the same position in the estimation time window for the sake of audiovisual correlation analysis, as was assumed in previous methods [4] [6] [8]. The time window is generally within $2-4$ seconds. We extract the audiovisual features inside this time window and analyze the statistical audiovisual correlation at each pixel position using a novel measure — *Quadratic Mutual Information* (QMI). We then do a global optimization to extract a reliable speaker region from the pixel-based correlation values, which is based on the Graph Cut-based image segmentation technique. Note that the speaker region here does not necessarily mean his/her face region, although our method can detect the whole face region in most cases. Sometimes only the mouth region is detected, as discussed in Sec. 4. This will not be a problem if only the speaker position is required. Yet, for the applications requiring the face region, we can use the face detection rectangle to supply a coarse mask because we now know who the speaker is.

In contrast to audio source localization, image segmentation has been studied for decades. Recently, Boykov and Funka-Lea made an important progress step [1], in which a globally optimum segmentation, which balances pixel likelihood and image region information, can be found efficiently using *Graph Cut*. The method works for not only a single image, but also for multiple video frames with inter-frame continuity considered [1]. The global optimization framework and the ability to process multiple video frames well fit our requirements. Additionally, although the method in [1] requires the user to designate seeds of foreground and background manually, which limits its usefulness, our method successfully eliminates this requirement by integrating audio information. Audio information also gives our method better robustness against backgrounds than [1].

The main contribution of this work can be summarized as follows: 1) to find the speaker face region robustly against different views and complex backgrounds by incorporating audiovisual correlation analysis into image segmentation, 2) to firstly use the QMI and kernel density estimation to analyze audiovisual correlation, and 3) to adopt audio information to eliminate the manual operations in Graph Cut-based segmentation and improve its robustness against complex backgrounds.

The rest of this work is organized as follows. In Section 2, we propose our audiovisual features and the computation of pixel-based audiovisual correlation using QMI. In Section 3, we explain how we find the speaker face region by performing image segmentation based on the results of audiovisual correlation analysis. In Section 4, we present and discuss our experimental results. In Section 5 we present our conclusions.

2 Audiovisual correlation analysis

If audio and visual signals are originated from one source, they should have strong correlation. To measure such correlation, some methods have been proposed; as reviewed in Sec. 1. However, none of the previous measures satisfies our requirements: to analyze the pixel-based audiovisual correlation under arbitrary distributions. Consequently, we introduce a novel measure — QMI, which is based on *kernel density estimation*. For continuous random variables, QMI can be computed much more efficiently than MI and is adopted to analyze the pixel-based audiovisual correlation. Moreover, kernel density estimation can estimate arbitrary probability density function (p.d.f.) [12]. Below we first introduce our audiovisual feature. Then we introduce the computation of audiovisual correlation by QMI.

2.1 Our audiovisual feature

As synchrony is the key to compute audiovisual correlation, we require the audio and visual signals to be recorded synchronously in the video. Based on this prerequisite, we extract the audio and visual features.

Audio feature. Input audio data are first divided into frames. The duration T_a of each audio frame is set to be the same as that of each visual frame. In order to keep the continuity between frames, it is set such that each pair of two successive frames have an overlap of the duration of $T_a/2$. Additionally, to reduce the boundary effect, a Hamming window is multiplied [14]. Coefficients of a standard Hamming window can be calculated through $h(m) = 0.54 - 0.46\cos(\frac{\pi m}{M}), \quad m = 1, \cdots, M$ where M is the number of the samples in an audio frame. The logarithm energy of each frame is then computed by $a(t) = \log\left(\frac{1}{M}\sum_{m=1}^{M} s^2(t,m)\right)$, where $s(t,m)$ refers to the processed audio sample m in frame t. The audio feature is defined to be the differential energy between the current and next frames as $fa_t = a(t+1) - a(t)$. Adopting differential energy is closely related to our adopted visual feature: the pixel-based intra-frame movement. Differential audio energy obviously demonstrates much higher correlation with this feature. The reason for this fact can be explained as follows: when the uttering movement is fast, audio energy should change quickly also.

We perform a verification of the existence of speech, although none of the previous methods has mentioned this process. Such process is important not only in the meaning that it checks whether audio conveys information, but also for its function to act as trigger to tell our system when it should work. The verification by now is to check whether or not the audio energy $a(t)$ is beyond a pre-defined threshold. It can be extended to check whether or not speech exists in the audio by using more complex algorithm. Frames failing this test are regarded as silence and dropped, together with their corresponding image frames. Note that, if $a(t+1)$ fails the test, fa_t will not be extracted either. From here on, if we talk about N video frames with speech, this refers to the frames that pass this test, and not to the actual number of video frames, which may exceed N.

Visual feature. Similar to the considerations in [11], we believe that better audiovisual correlation exists not in the pixel values but between the movements of the photographed objects and the audio. However, to accurately extract all objects and track their movements in a video sequence is difficult and not robust. Instead, we adopt the intra-frame pixel-based movement, which is described by optical flow and extracted between every two consecutive video frames.

To adopt optical flow as the visual feature has three advantages for our system. First, it helps our system to be background robust. Movement is invariant to the complexity of a static background. For a moving background, as its movements usually correlate marginally with audio, the influence can be suppressed in the subsequent correlation analysis. Second, it helps our system to be view robust. Since optical flow is extracted locally at each pixel position, view difference of the human face has only minor influence on it. No matter how different a face looks in different view, its local parts move similarly when speaking. Third, the locality of the optical flow also makes it possible for our method to achieve good segmentation boundary.

Note that our method is different from motion segmentation technique. Although both adopt visual movements as feature, the inside assumptions differ largely. Motion segmentation assumes that regions of pixels move with coherent motion and evaluate the coherence to do segmentation [7]. Yet the speaking-related actions are generally not coherent, e.g., the two lips of a people usually move in opposite directions. Motion segmentation cannot distinguish which motion correlates with audio either.

We compute the optical flow on gray images using the Lucas-Kanade method [9]. The window size we used is 9×9. Since the optical flow cannot be estimated stably for windows with low texture, we verify the pixel intensity variation inside the window. If it is below a threshold, we set its flow value to zero.

Optical flow has two elements: horizontal and vertical movement. We need to transfer it into a scalar value to compute the correlation with the audio feature. We tested three derived values: amplitude, horizontal element, and vertical element. Results are shown in Figure. 2. The vertical element obviously has much higher correlation with the audio feature. This matches our intuition since speaking movements generally happen vertically. Consequently, we adopt its vertical element to be our visual feature, i.e., the visual feature $fv_t(u, v)$ at pixel position (u, v) in frame t is defined as the vertical part of the pixel's optical flow extracted between frame t and $t + 1$.

2.2 Audiovisual correlation by QMI

We regard the audio and visual features as two random variables and compute the correlation statistic using their temporal samples extracted from each video frame. Note that the correlation will be computed independently at each pixel position (u, v).

Firstly we adopt kernel density estimation to estimate their joint probability density function (p.d.f.). Kernel density estimation [12] (also known as Parzen window estimation) is a method to estimate arbitrary p.d.f. of a random variable.

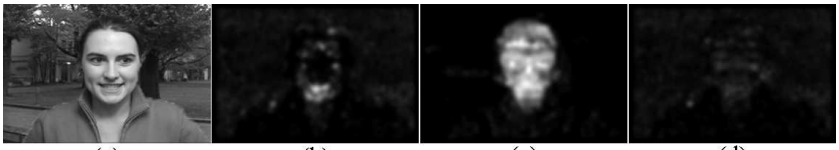


Fig. 2. Audiovisual correlation using different optical flow elements. The whiter a pixel, the higher its correlation. (a-d) show the original image, correlation by horizontal element, vertical element and mode, respectively.

Given N data points $\mathbf{x}_i$, $i = 1, \cdots, N$ in the d-dimensional space R^d, the multivariate kernel density estimation with kernel $K_\Sigma(\mathbf{x})$ and a symmetric positive definite $d \times d$ bandwidth matrix H, computed in the point $\mathbf{x}$ is given by

$$p(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} K_\Sigma(\mathbf{x} - \mathbf{x}_i), \tag{1}$$

where $K_\Sigma(\cdot)$ is the specified kernel function. In practice, Gaussian function with the diagonal bandwidth matrix $\Sigma = diag(\sigma_1^2, \cdots, \sigma_d^2)$ is often adopted.

Based on the estimated p.d.f., we compute the statistical correlation between the audio and visual feature. However, the computation of the Shannon MI by Eq. (1) is difficult since it does not yield a closed-form solution. Thus we adopt QMI proposed by Xu *et al.* [17], which is based on the quadratic form of Renyi entropy.

If one of the four constraints defined by Shannon is weakened, Renyi [15] showed that the entropy of a random variable can be evaluated by a group of functions defined as $H_\alpha(x) = \frac{1}{1-\alpha} \log \left(\int p^\alpha(x) dx \right)$, where $\alpha > 0, \alpha \neq 1$. As $\alpha \to 1$, $H_\alpha(x)$ approaches the Shannon entropy $H(x)$. In practice, the one most often used is its quadratic form $\alpha = 2$, which can be easily computed based on the p.d.f. estimated by kernel density estimation using a diagonal Gaussian kernel because we have [17]

$$\int K_\Sigma(\mathbf{x} - \mathbf{x}_i) K_\Sigma(\mathbf{x} - \mathbf{x}_j) d\mathbf{x} = K_{2\Sigma}(\mathbf{x}_i - \mathbf{x}_j). \tag{2}$$

Unfortunately, Renyi MI is left undefined. Although Shannon MI between two random variables x_1 and x_2 can be directly computed from their entropies as $MI(x_1; x_2) = H(x_1) + H(x_2) - H(x_1 x_2)$, this is not true for the Renyi entropy. In 1998, Xu *et al.* [17] proposed a form to compute QMI based on Renyi entropy, which is defined as

$$C(x_1, x_2) = \log \frac{\iint p^2(x_1, x_2) dx_1 dx_2 \iint p^2(x_1) p^2(x_2) dx_1 dx_2}{(\iint p(x_1, x_2) p(x_1) p(x_2) dx_1 dx_2)^2}. \tag{3}$$

It is easy to show that $C(x_1, x_2) \geq 0$ and the equality holds true if and only if $p(x_1) = p(x_2)$ using Cauchy-Schwartz inequality.

In our case, using the temporal samples of the audiovisual feature $\{(x_{1t}, x_{2t}) = (fa_t, fv_t(u,v)), t = 1, \cdots, N\}$, we can analyze their QMI at each pixel position (u, v). Following Eq. (2), QMI can be directly computed from the sample set in closed-form as [17]

$$\begin{cases} C(x_1, x_2|\{x_t\}) = \log \frac{V_c(\{x_t\})V_m(\{x_{1t}\})V_m(\{x_{2t}\})}{V_{nc}^2(\{x_t\})} \\ V_c(\{x_t\}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\prod_{k=1}^{2} G(x_{ki} - x_{kj}, 2\sigma_k^2) \\ V_s(x_{kj}, \{x_{ki}\}) = \frac{1}{N}\sum_{i=1}^{N} G(x_{kj} - x_{ki}, 2\sigma_k^2) \quad k = 1, 2 \\ V_m(\{x_{ki}\}) = \frac{1}{N}\sum_{j=1}^{N} V_s(x_{kj}, \{x_{ki}\}) \quad k = 1, 2 \\ V_{nc}(\{x_t\}) = \frac{1}{N}\sum_{j=1}^{N}\prod_{k=1}^{2} V_s(x_{kj}, \{x_{ki}\}) \end{cases} \tag{4}$$

where $G(x, \sigma^2)$ represents a 1D Gaussian p.d.f. and $K_\Sigma(\mathbf{x}) = \prod_{k=1}^{2} G(x_k, \sigma_k^2)$. An example of the analyzed correlation can be seen in Figure. 4.

3 Speaker region segmentation

Based on the analyzed pixel-based audiovisual correlation, we take advantage of the image segmentation technique to extract the speaker region.

Again using the retrieved N video frames, we build the N-D image as defined in [1] and perform the video segmentation. The distance of each pixel to the foreground (speaker) and the background is computed from the analyzed audiovisual correlation. Since there is only one scalar correlation value at each image position (u, v), the computed likelihood is same for all pixels at (u, v), regardless of in which frame t they are. On the other hand, as image information, like edge, pixel similarity and intra-frame continuity, is related to both (u, v) and t, segmentation results can still be different in each frame and capture the face deformation when speaking.

3.1 Graph Cut-based segmentation

Segmentation of video frames by optimizing a global energy function was proposed in [1]. The global energy function is composed of two important terms: the sum of the data cost to assign a pixel to be speaker, and the sum of the smoothness penalty between every two neighboring pixels in both temporal and spatial domains. Its definition is as Eq. (5),

$$E(l) = \sum_{p} D_p(l_p) + \lambda \cdot \sum_{\{p,q\} \in Ne} S_{pq}(l_p, l_q), \tag{5}$$

where l represents the segmentation labels of all pixels in the N-D image. $l_p = 1$ means pixel p is labeled as the speaker; while $l_p = 0$ means the background. Ne defines the neighborhood relationship between two pixels, which is discussed in detail in Sec. 3.3. λ is a constant adjusting the balance between the two terms.

It was shown that the energy function defined in Eq. (5) can be efficiently optimized by calculating the minimum cut of a graph using a maximum flow algorithm [1]. Moreover, the optimization result is guaranteed to be the global minimum solution of the energy function [1].

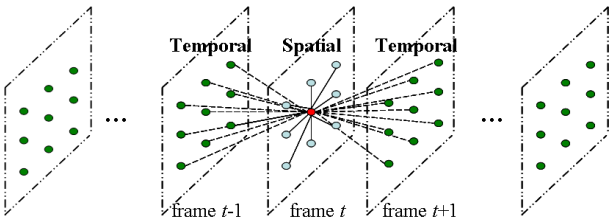


Fig. 3. A demonstration of the N-D image and the neighborhood.

3.2 Data cost by audiovisual correlation

Since QMI is not a normalized correlation measure, its dynamic range changes largely for different retrieved N video frames. Furthermore, as shown in Figure. 4, although speaker face region obviously has higher audiovisual correlation values, their variation is large. Thus, it is difficult to directly adopt the correlation values as the data cost.

In order to solve these difficulties, we apply a clustering process to the correlation values from all pixels and take the Mahalanobis distance to each cluster center as the data cost. First, the highest and lowest analyzed audiovisual correlation values are selected as two seeds. Then, applying the *Expectation-Maximization* (EM) algorithm to all correlation values, we can train two one-dimensional Gaussian distributions. The EM training generally converges after $6-10$ iterations. The one trained from the seed of the lowest correlation is regarded as the background correlation distribution, denoted as $G(\mu_0, \sigma_0^2)$. The other is regarded as the foreground (speaker) distribution, denoted as $G(\mu_1, \sigma_1^2)$. Finally, the pixel's data cost is computed by the Mahalanobis distance to each center, as defined in Eq. (6).

$$D_p(l_p) = \begin{cases} (C(fa, fv(u,v)) - \mu_1)^2/\sigma_1^2 & l_p = 1 \\ (C(fa, fv(u,v)) - \mu_0)^2/\sigma_0^2 & l_p = 0 \end{cases} \tag{6}$$

3.3 Smoothness penalties by image information

Smoothness penalty is forced between every two neighboring pixels in both spatial and temporal domains. In the N-D image, the spatial and temporal neighborhoods are defined as shown in Figure. 3. Each pixel can maximally have 26 neighbors.

The value of smoothness penalty is computed by

$$S_{pq}(l_p, l_q) = \exp\left(-\beta(I_p - I_q)^2\right) \cdot (d(p,q))^{-1} \cdot T[l_p \neq l_q], \tag{7}$$

where p and q are two neighboring pixels. I_p and I_q are their intensity values. The constant β is chosen as in [1] to be $\beta = \left(2\left\langle (I_p - I_q)^2 \right\rangle\right)^{-1}$, where $\langle\cdot\rangle$ denotes the expectation over the N-D image sample. This choice of β ensures that the exponential term in Eq. (7) switches appropriately between high and low contrast.

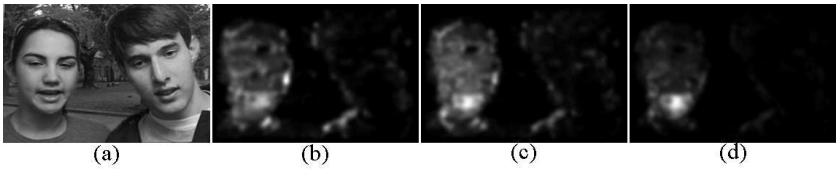


Fig. 4. Statistical audiovisual correlation using different frame numbers. (b-d) demonstrate the analyzed correlation values using 20, 40 and 80 video frames with speech. Correlation values are normalized independently in (b-d).

$d(p, q)$ calculates Euclidean distance between p and q in the three-dimensional grid, which may be 1, $\sqrt{2}$ or $\sqrt{3}$ in our neighborhood model. $T[\cdot]$ is a boolean function returning 1 when the condition inside is true and 0 otherwise.

4 Experimental results

We used the CUAVE audiovisual database [13] in all our experiments, where 17 females and 19 males uttered English numbers in front of a green background with frontal, lateral, and moving views. Video sequences were filmed at 29.97 fps, while the audio signals were sampled at 44 kHz stereo. An advantage of CUAVE database was that we could remove the green background by color and replace it with other complex backgrounds to test the performance of our algorithm. Color images were then converted into gray images and down-sampled from 720×480 to 240×160 for memory saving. For all our experiments, the kernel bandwidths are set as $(\sigma_1, \sigma_2) = (0.4, 0.3)$. The balance constant in Graph Cut is set as $\lambda = 20$.

Since we computed the audiovisual correlation by temporal statistics, we first tested the computed correlation values using different numbers of the video frames. The experimental results are shown in Figure. 4. We can see that using more frames helps to remove the ambiguity of another person besides the speaker. However, using more frames tends to break our assumption that the speaker stays at the same position. Thus, we generally adopt 40 frames to compute our experimental results. On our laptop whose CPU is Intel Core2 1.83G with 1G RAM, it takes about 31 seconds.

We tested the performance of our algorithm for different backgrounds, with results shown in Figure. 5 (a-c). 40 frames were used for computation, although only one of them is shown. The results of our algorithm show minor changes for different backgrounds. We also implemented the method in [1], whose manually designated seeds and the segmentation results over the same 40 frames are shown in Figure. 5 (d-g). Note that this does not mean a comparison of performance, as the results by [1] can be improved iteratively by adding more seeds. However, the results demonstrated that, through using audio information, our method achieved better robustness to the different complex stationary or non-stationary backgrounds.

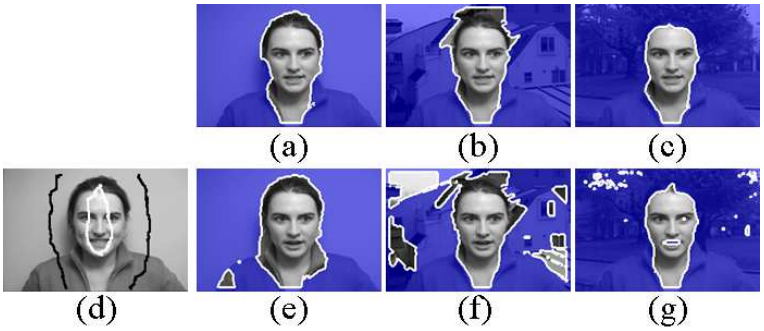


Fig. 5. Estimated results by the method in [1] and ours. Areas blended with blue mean estimated background. Boundary pixels are shown as white. (a-c) show our segmentation results of original video, video merged with a static background and video merged with a non-stationary background, respectively. (d) shows the user designated seeds. (e-f) are the segmentation results by the method in [1].

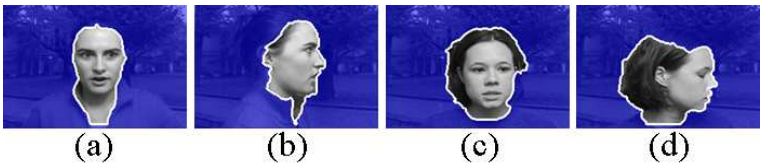


Fig. 6. Estimation results for different views. (a) and (c) show the detected speaker face region for a frontal view. (b) and (d) show the results for a lateral view.

As both frontal and lateral views of the speaker face were photographed, we applied our method to the data using completely same parameters. The results are shown in Figure. 6. Since frontal and lateral views are two extreme cases of the view change, the success of our method to process them elegantly in the same framework demonstrated its robustness against different views.

We also applied our method to several other sequences with same parameters, where single or multiple persons are photographed. The experimental results are shown in Figure. 7. Sequences (a-c) demonstrated the results of other single speakers under different backgrounds. For (c), our method could segment out the mouth region of the speaker only. As the man in (c) moved his mouth only when speaking and demonstrated none of other speaking-related actions, our method can segment his mouth region only. Sequences (d-f) demonstrate the results for multiple persons. Our method correctly located the current speaker and detected their face regions in most cases.

To give a quantitative evaluation of our detection result, we have manually labeled the face regions for the first frame of four video sequences. The ground

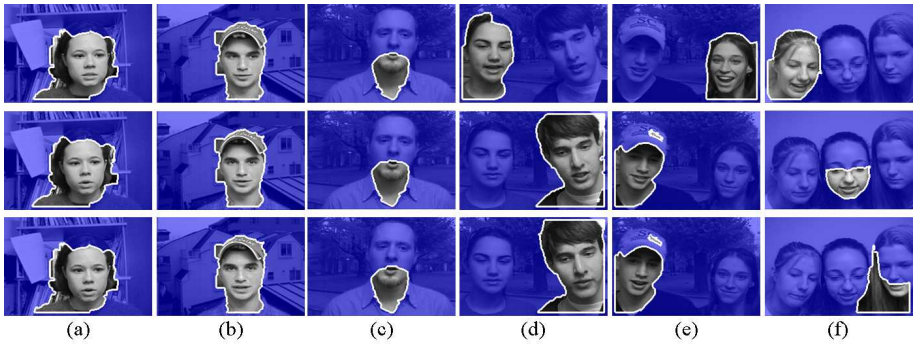


Fig. 7. The experimental results for other persons. Sequences (a-c) are the results when only a single person exists. Sequences (d-f) are results when multiple persons exist.



Fig. 8. Ground truth and the detection rate of our method. The ground truth in the first row shows the manually labeled face region superimposed over the original image.

truth and the detection rate of our method were shown in Figure. 8. In most cases, our method can extract the face region with high detection rate.

5 Conclusions and future works

In this work, we have developed a method to find speaker face region in gray images robustly against views and backgrounds by integrating audiovisual correlation analysis into Graph Cut-based image segmentation framework. We have shown that our method is capable of finding less fragmented speaker face regions than previous methods for both single and multiple persons under different views and different complex backgrounds.

Our current evaluation of audiovisual correlation is sensitive to the noise. Visual noise may yield incorrect optical flow in untextured regions. Audio noise may disturb the frame energy estimation and decrease the audiovisual correlation. We plan to try other reliable methods to compute optical flow and other

robust audio features, especially the audio features in frequency domain. Additionally, our method requires the speaker to stay at relatively the same position in the statistical time span. We will consider methods to eliminate this constraint.

References

1. Y. Boykov, G. Funka-Lea. "Graph Cuts and Efficient N-D Image Segmentation". *Int'l J. of Computer Vision*, 70(2):109-131, 2006.
2. A. L. Casanovas. "Blind audiovisual source separation using sparse redundant representations". *Master thesis*, Signal Processing Institute, EPFL, 2006.
3. N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2:886-893, 2005.
4. T. Darrell and J. W. Fisher III. "Speaker association with signal-level audiovisual fusion". *IEEE Trans. on Multimedia*, 6(3):406-413, 2004.
5. J. Driver. "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading", *Nature*, 381:66-68, 1996.
6. J. Hershey and J. R. Movellan. "Audio vision: Using audiovisual synchrony to locate sounds". In *NIPS*, 813-819, 1999.
7. K. Kanatani. "Motion segmentation by subspace separation and model selection". In *IEEE Int'l Conf. on Computer Vision*, 2:586-591, 2001.
8. E. Kidron, Y. Y. Schechner, and M. Elad. "Pixels that sound". In *IEEE Conf. on Computer Vision and Pattern Recognition*, 88-95, 2005.
9. B. Lucas, and T. Kanade. "An Iterative Image Registration Technique with an Application to Stereo Vision". In *Int'l Joint Conf. on Artificial Intelligence*, 674-679, 1981.
10. J. Luettin, N. Thacker and S. Beet. "Speaker identification by lipreading". In *Proc. Int'l Conf. on Spoken Language*, 1:62-65, 1996.
11. G. Monaci, O. Escoda and P. Vandergheynst. "Analysis of multimodal signals using redundant representations". In *Int'l Conf. on Image Processing*, 145-148, 2005.
12. E. Parzen. "On the estimation of probability density function and the mode". *The Annals of Mathematical Statistics*, 33(3):1065-1076, 1962.
13. E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. "Moving-talker, speaker-independent feature study and baseline results using the cuave multimodal speech corpus". *EURASIP J. on Applied Signal Processing*, 2002(11):1189-1201, 2002.
14. L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.
15. A. Renyi. "On measures of Entropy and Information". In *Fourth Berkeley Symp. Math. Stat. and Probability*, 1:547-561, 1961.
16. P. Viola and M. Jones. "Robust Real-Time Face Detection". *Int'l J. of Computer Vision*, 57(2):137-154, 2004.
17. D. Xu, J. Principe and J. Fisher. "A Novel Measure for Independent Component Analysis (ICA)". In *Int'l Conf. on Acoustics, Speech and Signal Processing*, 2:1161-164, 1998.