PAPER

Segmentation of the Speaker's Face Region with Audiovisual Correlation

Yuyu LIU[†a)], ***Nonmember and*** **Yoichi SATO**[†b)], ***Member***

SUMMARY The ability to find the speaker's face region in a video is useful for various applications. In this work, we develop a novel technique to find this region within different time windows, which is robust against the changes of view, scale, and background. The main thrust of our technique is to integrate audiovisual correlation analysis into a video segmentation framework. We analyze the audiovisual correlation locally by computing quadratic mutual information between our audiovisual features. The computation of quadratic mutual information is based on the probability density functions estimated by kernel density estimation with adaptive kernel bandwidth. The results of this audiovisual correlation analysis are incorporated into graph cut-based video segmentation to resolve a globally optimum extraction of the speaker's face region. The setting of any heuristic threshold in this segmentation is avoided by learning the correlation distributions of speaker and background by expectation maximization. Experimental results demonstrate that our method can detect the speaker's face region accurately and robustly for different views, scales, and backgrounds.

key words: *speaker detection, audiovisual analysis, segmentation, graph cut*

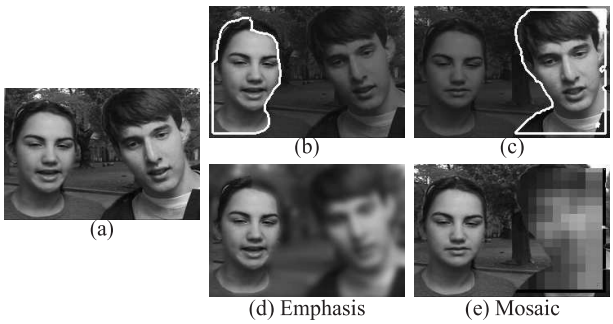

Fig. 1 Special effects given the speaker's face region. Figure (a) shows the original image, figures (b) and (c) show the speaker face region localized by our method, and figures (d) and (e) show the special effects imposed based on our estimation.

1. Introduction

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

However, regardless of the great progresses made in face and human detection (e.g., [26] and [5] respectively) in recent years, speaker detection is still under development. As the purpose is to distinguish a person from others, either face detection or human detection fails in this area. One solution [16] is to detect the face, locate the mouth, and check its movement. The weakness of this method is the requirement of a frontal view. View dependency is also a challenging problem for face and human detections [10]. Additionally, it is prone to be disturbed by unconscious movements from other persons.

Manuscript received August 20, 2009.
Manuscript revised February 21, 2010.
[†]The authors are with Institute of Industrial Science, The University of Tokyo, Tokyo, 153–8505 Japan.
a) E-mail: liuyuyu@iis.u-tokyo.ac.jp
b) E-mail: ysato@iis.u-tokyo.ac.jp
DOI: 10.1587/transinf.E93.D.1965

In this work, we develop a novel technique to find the speaker's face region for different time windows, which is robust against the changes of view, scale, and background. This technique is based on the recent developments in sound source localization by audiovisual correlation analysis and an segmentation technique of multiple video frames.

To localize sound source by analyzing audiovisual correlation is a relatively new research topic and has drawn much attention in recent years. Psychological research [7] discovered that human beings perceive the audiovisual correlation based on synchrony. Initiated from this work, Hershey and Movellan [9] introduced a method to localize the sound source by audiovisual correlation analysis. They assumed that audio energy and pixel intensity obey joint normal distribution and measured their mutual information to detect the sound source. Following this work, many methods were developed. Smaragdis and Casey [24] suggested combining pixel intensities and audio spectrum into a large vector and using first principal component analysis and then independent component analysis to find out visual independent components, which were regarded as the estimated sound sources. Fisher and Darrell [8] searched an optimum projection vector that can maximize the lower bound of the audiovisual correlation computed by mutual information. Kidron and Schechner [13] searched a projection vector that can maximize the audiovisual correlation by canonical correlation analysis between audiovisual signals and has minimum $L1$-norm. In 2005, Monaci *et al.* [18] claimed that the movements of the photographed objects convey better audiovisual correlation than the pixel values. They employed matching pursuit to extract local objects, tracked their move-

ments and computed feature-based audiovisual correlation to detect the sound source.

Unfortunately, all existing localization methods suffer a common problem: the estimated masks of sound source are highly fragmental. These locally detected fragments include both foreground and ambiguous background locations, which leads to difficulties in designating a correct speaker position, much less identifying a reliable speaker region. Therefore, most of them only detect pixels that are supposed to be sound source, except that Casanovas [4] clustered the detected pixels and adopted the cluster center to be the detected speaker position. One weakness of this method is that clustering may be vulnerable to the background outliers that appear often [8], [13], [18].

To solve this problem, we consider to detect a reliable face region of the speaker, rather than fragmental masks. Region center can be regarded as the speaker's position, which is more robust than the cluster center in [4]. Additionally, region information satisfies the needs of advanced applications. Our key idea is to integrate the techniques to analyze audiovisual correlation locally into a video segmentation framework. As an important topic, image segmentation has been researched for decades (for instance, [12], [29]). Recently, Boykov and Funka-Lea made an important progress step [3]. They developed a technique to efficiently achieve global optimum segmentation that balances pixel likelihood and image region information by using graph cut. The method works for not only a single image, but also for multiple video frames with inter-frame continuity considered [3]. A weakness of graph cut is the requirement of a manual operation to designate seeds of foreground and background. Fortunately, our incorporation of audiovisual correlation analysis not only takes advantage of the effective optimization of graph cut, but also removes the necessity of this manual operation.

Other works have also been developed to incorporate information into graph cut-based segmentation to enhance the performance and remove the manual operation. Kolmogorov *et al.* [14] adopted stereo depth information to segment foreground. Yu *et al.* [28] based their method on face detection to segment people. Schoenemann and Cremers [23] took advantage of motion information to divide motion layers. However, their incorporated information is still based on visual signal and cannot supply the cues beyond the visual signal. For example, if both people move their mouths, without audio, one can hardly tell who the real speaker is. Fusing audio not only resolves this ambiguity, but also improves the robustness compared to the usage of visual signal only. To the best knowledge of the authors, this is the first trial to fuse other modality information into the Graph Cut-based segmentation.

We analyze the audiovisual correlation by computing quadratic mutual information between our audio and visual features. We extract visual features locally, whose locality helps our method to be robust to the change of view, and compute quadratic mutual information to analyze the audiovisual correlation. Kernel bandwidth needed to compute this quadratic mutual information is estimated from data, which makes our method adaptive to the changes of visual scale and audio gain. It is demonstrated in this work that quadratic mutual information outperforms mutual information, by analyzing the correlation between two sets of randomly generated numbers that are linearly or non-linearly related.

We then incorporate local audiovisual correlations into a global optimization framework to extract the speaker's face region by using graph cut-based video segmentation technique. To avoid a heuristic decision of a segmentation threshold, we learn the distributions of the audiovisual correlation of speaker and background by using expectation maximization. The likelihoods of each pixel to these two distributions are combined with image smoothness constraints to form the energy function in the graph cut segmentation.

Our 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 [8], [9], [13]. The time window is generally within 2–4 seconds.

We detect the speaker's face region but not the mouth region because, when speaking, many unconscious movements happen also on face parts, which are as highly correlated with the audio as mouth movements. Consequently, in most cases our method can detect the whole face region. However, as discussed in Sect. 4, if the speaker intentionally restrains these unconscious movements, only the mouth region of this speaker can be detected. If the speaker position only is needed, this will not be a problem as the mouth region center can be adopted. If the whole face region is needed, a manual extension of the segmentation mask is necessary in such cases.

The main contribution of this work can be summarized as follows: 1) to find the speaker's face region by incorporating audiovisual correlation analysis into video segmentation, including the method to locally analyze the audiovisual correlation and the learning of correlation distributions, and 2) 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 Sect. 2, we introduce the audiovisual feature and the correlation computation using quadratic mutual information. In Sect. 3, we explain how we find the speaker's face region by performing video segmentation based on the audiovisual correlation. In Sect. 4, we present and discuss our experimental results. In Sect. 5 we present our conclusions.

2. Audiovisual Correlation Analysis

Within a time window, we extract the visual feature at each local position (x, y) and each time t, and the audio feature at each time t. The correlation between the temporal changes of the visual feature at (x, y) and the audio feature is analyzed by using quadratic mutual information. After the analysis, we can get a table $QMI(a; v(x, y))$ which shows

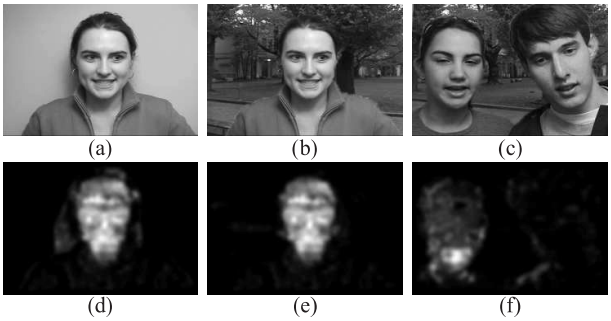

Fig. 2 Analyzed audiovisual correlation for different video sequence. Correlation in (d–f) are normalized independently. The whiter a pixel, the higher its correlation.

the audiovisual correlation of each image position (x, y) in the current time window. An example of the audiovisual correlation table is shown in Fig. 2.

Below we first introduce our audiovisual feature and then explain the correlation analysis by using quadratic mutual information. Note that our audiovisual correlation analysis is based on the prerequisite that the audio and visual signals are recorded synchronously. Asynchronous data may degrade the accuracy of this analysis.

2.1 Our Audiovisual Feature

Both our audio and visual features describe the differential between two continuous frames. However, because of the substantial difference between the audio and visual signals, their extraction methods differ significantly.

2.1.1 Audio Feature

Since audio is usually sampled at a much higher frequency than video, we first divide audio samples into frames to compute the audio feature. The frame duration, T_a, is set to be the same as the visual frame duration, T_v. In order to keep a temporal continuity, 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 [21], whose coefficients are computed by

$$w(i) = 0.54 - 0.46\cos\left(\frac{\pi i}{M}\right), \quad i = 1, \cdots, M \tag{1}$$

where M is the number of the audio samples in a $2T_a$ duration. The audio energy $e(t)$ of frame t is computed by

$$e(t) = \log\left(\frac{1}{M}\sum_{i=1}^{M}(w(i)s(t,i))^2\right), \tag{2}$$

where $s(t, i)$ refers to the processed audio sample i in frame t and the two surrounding overlaps of frame t. This process is demonstrated in Fig. 3.

The audio feature is defined as the differential energy between the current and next frames, which is given by

$$a_t = e(t+1) - e(t). \tag{3}$$

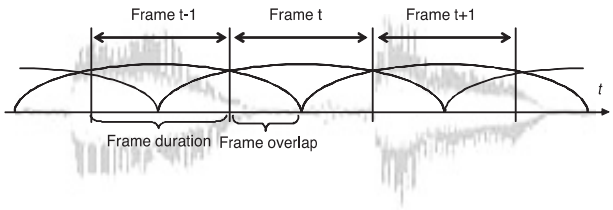

Fig. 3 Division of audio frames.

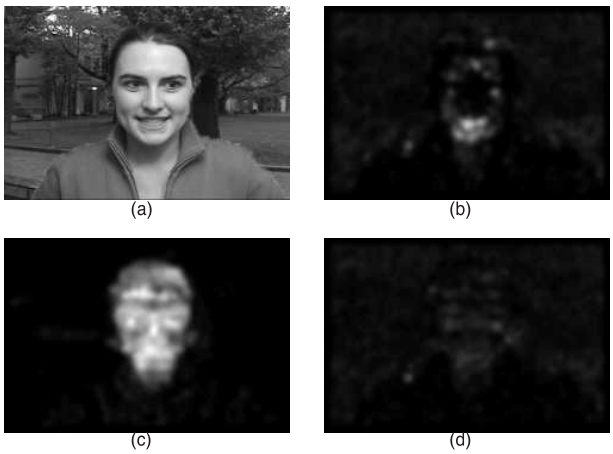

Fig. 4 Audiovisual correlation using different optical flow elements. Figure (a) shows an original video frame, and figures (b), (c) and (d) show the analyzed audiovisual correlation by using horizontal element, vertical element, and amplitude, respectively.

Since in silence durations the absence of audio information makes it impossible to analyze the audiovisual correlation, we ensure there is speech in all frames. This is done by checking whether or not audio energy $e(t)$ is larger than a pre-defined threshold. Frames failing this test are regarded as silent and dropped, together with their corresponding visual frames. Only frames passing this test are buffered till the frame number reached a pre-defined value. If we discuss N audiovisual frames in this chapter, it refers to frames that are buffered.

2.1.2 Visual Feature

The same as [18], we believe that there is an audiovisual correlation mainly between the movements of visual objects and the change of audio. Therefore, optical flow is used as the visual feature in our method. In particular, we only take the vertical element of optical flow considering that most speaking actions move vertically. We have compared three methods to get a scalar visual feature from the 2D optical flow vector: horizontal element, vertical element, and amplitude. The results of analyzed audiovisual correlation are shown in Fig. 4. The vertical element obviously has much higher correlation with the audio feature.

Visual feature $v_t(x, y)$ is defined as the vertical optical flow extracted at (x, y) between frames t and $t + 1$, which is computed by the Lucas-Kanade method [15]. Since optical

flow cannot be estimated stably in areas with less texture, we verify the variation of pixel intensities inside each window where we compute optical flow. If these are below a threshold, we set the flow value to be zero.

Adopting optical flow as the visual feature has three advantages for our system. First, it helps our system to be background robust. For a static background, movement is independent to its complexity. For a moving background, as its movements usually correlate marginally with audio, the influence can be suppressed in the subsequent correlation analysis also. Second, it helps our system to be view robust. No matter how different a face looks in different views, the local movements resulted from speaking action are similar. Optical flow captures this local movement and is thus view robust. Third, the locality of the optical flow also makes it possible for our method to achieve good segmentation boundary. Since optical flow describes the movement of each pixel, our segmentation can achieve an accuracy of every pixel.

2.2 Audiovisual Correlation by Quadratic Mutual Information

Many works have used mutual information to measure the audiovisual correlation [8], [9]. However, to analytically compute mutual information for continuous random variables, they either computed the second-order Taylor extension of mutual information [8], or assumed that audio and visual features obey normal distribution [9]. The former one can approach only an approximation of mutual information and requires an iterative computation process. The latter one is arguable since obeying normal distribution is a strong assumption. We have applied a normality test [6] to our audio and visual features extracted inside a speaker mouth region. The results showed that 77.8% of the tests fall into the refusal area $[9.49, +\infty)$ with ρ-value=0.05. That is to say, this assumption should be wrong in a confidence of 95%.

Instead, we use quadratic mutual information as a measure to analyze the audiovisual correlation, which can be computed analytically directly from the data without the necessity of any approximation and assumption. The computation of quadratic mutual information is based on the probability density functions (pdf) estimated by kernel density estimation [19]. The bandwidth of the kernel density estimation is estimated from the variance of the data, which makes our method robust to the changes of visual scale and audio gain.

Quadratic mutual information is computed based on the temporal samples of the audio and visual features. This analysis is independently performed for different image positions by using the visual feature extracted at each image position (x, y). After quadratic mutual information at all positions (x, y) are computed, we adopt the analyzed audiovisual correlation to segment out the speaker's face region in the next stage.

2.2.1 Pdf Estimation by Kernel Density Estimation

Kernel density estimation (also known as Parzen window estimation) is a method of estimating the arbitrary pdf of a random variable [19]. Given N data points $\{\mathbf{z}_i, i = 1, \cdots, N\}$, in n-dimensional space R^n, the multivariate kernel density estimation with kernel $K_{\mathbf{H}}(\mathbf{z})$ and a symmetric positive definite $n \times n$ bandwidth matrix $\mathbf{H}$, computed in point $\mathbf{z}$ is given by

$$p(\mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{z} - \mathbf{z}_i), \tag{4}$$

where $K_{\mathbf{H}}(\cdot)$ is the specified kernel function.

We adopt a Gaussian kernel with a diagonal bandwidth matrix, $\mathbf{H} = diag(\sigma_1^2, \cdots, \sigma_d^2)$. Compared to the other kernels, such as a triangle kernel, a Gaussian kernel results in an efficient computation of quadratic mutual information.

The selection of an appropriate bandwidth is important for kernel density estimation [25]. Small bandwidth values make the estimate look "wiggly" and show spurious features, whereas too big values will lead to an estimate which is too smooth in the sense that it is too biased and may not reveal structural features. Therefore, comparing an empirical decision of this kernel bandwidth, we estimate the bandwidth from data. For Gaussian kernel, a *rule of thumb* was proposed to estimate a proper bandwidth from the data [25]. We adopt this *rule of thumb* to compute the bandwidth as

$$\sigma = 1.06 \hat{\sigma} n^{-\frac{1}{5}}, \tag{5}$$

where $\hat{\sigma}^2$ is the sample variance.

2.2.2 Correlation Analysis by Quadratic Mutual Information

Quadratic mutual information was proposed by Xu *et al.* [27] in 1998, which was used as an objective function for the independent component analysis.

We use this quadratic mutual information to compute the audiovisual correlation between the audio and visual features. The same as mutual information, quadratic mutual information indicates the amount of information that one random variable conveys about another. At an image position (x, y), quadratic mutual information is computed between the audio feature a and visual feature v by definition as

$$QMI(a; v) = \log \frac{\iint p^2(a, v) da dv \iint p^2(a) p^2(v) da dv}{(\iint p(a, v) p(a) p(v) da dv)^2}. \tag{6}$$

It can be shown that $QMI(a; v) \geq 0$ and the equality hold true if and only if $p(a) = p(v)$ using Cauchy-Schwartz inequality [27].

As mentioned before, quadratic mutual information can be computed analytically directly from the data. Given the temporal samples of the audio and visual features to be $\{(a_t, v_t), t = 1, \cdots, N\}$, it has been shown [27] that quadratic

mutual information can be computed as

$$QMI(a; v|\{a_t, v_t\}) = \log \frac{V_c(\{a_t, v_t\})V_m(\{a_t\})V_m(\{v_t\})}{V_{nc}^2(\{a_t, v_t\})} \quad (7)$$

where $V_c(\{a_t, v_t\})$, $V_m(\{a_t\})$, $V_m(\{v_t\})$, and $V_{nc}(\{a_t, v_t\})$ are the terms computed from the data samples, which are given by

$$V_c(\{a_t, v_t\}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} K_{2\sigma_a^2}(a_i - a_j)K_{2\sigma_v^2}(v_i - v_j), \quad (8)$$

$$V_m(\{a_i\}) = \frac{1}{N}\sum_{j=1}^{N} V_s(a_j, \{a_i\}), \quad (9)$$

$$V_m(\{v_i\}) = \frac{1}{N}\sum_{j=1}^{N} V_s(v_j, \{v_i\}), \quad (10)$$

$$V_{nc}(\{a_t, v_t\}) = \frac{1}{N}\sum_{j=1}^{N} V_s(a_j, \{a_i\})V_s(v_j, \{v_i\}). \quad (11)$$

Inside them, $V_s(a_j, \{a_i\})$ and $V_s(v_j, \{v_i\})$ are computed as

$$V_s(a_j, \{a_i\}) = \frac{1}{N}\sum_{i=1}^{N} K_{2\sigma_a^2}(a_j - a_i), \quad (12)$$

$$V_s(v_j, \{v_i\}) = \frac{1}{N}\sum_{i=1}^{N} K_{2\sigma_v^2}(v_j - v_i). \quad (13)$$

Here $K_\sigma(\cdot)$ is a one-dimensional Gaussian kernel. σ_a and σ_v are the estimated bandwidths of the audio and visual features obtained by using Eq. (5).

Using Eq. (7), we can compute quadratic mutual information directly from the samples without even the necessity to explicitly formulate pdf. Additionally, although the complexity of quadratic mutual information computation by using the definition in Eq. (6) is $O(N^4)$, by using Eq. (7) we can compute it at a complexity of $O(N^2)$ by removing duplicated computation.

As shown in appendix, with our bandwidth estimation, this correlation analysis is invariant to the scale changes of both audio and visual features. This invariance makes our method robust against the changes of both visual image scale and audio signal gain.

2.3 Mutual Information and Quadratic Mutual Information

Many previous works have proposed to adopt mutual information to analyze the audiovisual correlation [8], [9]. In this section, we perform a quantitative evaluation on their performance of analyzing the correlation.

One thing need to be mentioned first is that the computation of mutual information is slightly different to the one of quadratic mutual information. As shown in Eq. (7), quadratic mutual information can be computed analytically from the samples of continuous random variables. However, to compute mutual information, these samples have to be quantized first since it can be computed analytically for discrete pdf's only. Certainly the way to quantize samples casts effects on the performance of mutual information. We perform a uniform quantization, whose quantization level is also computed from data. As proposed in [25], the computation of this quantization level follows Eq. (5), which is the same as how we compute the kernel bandwidth to estimate quadratic mutual information.

Previous works [8], [9] adopted mutual information because it can discover the hidden correlation between two random variables. The reason lies in that mutual information is invariant to one-to-one maps, i.e., we have

$$MI(X; f(Y)) = MI(X; Y), \quad (14)$$

if $f()$ is a one-to-one map. Being an extended version of mutual information, quadratic mutual information holds the same property also, i.e.,

$$QMI(X; f(Y)) = QMI(X; Y). \quad (15)$$

However, under real situations, this invariance cannot be perfectly kept because the statistics are performed with a limited number of samples. To test how much this invariance is obeyed, we conduct a comparison experiment by using random numbers that are uniformly distributed between 0 and 1. The random numbers are generated by Mersenne twister algorithm [17]. Mersenne twister provides for fast generation of very high-quality pseudorandom numbers, having been designed specifically to rectify many of the flaws found in older algorithms.

The comparison was done by evaluating the correlation between two sets of random numbers. We first generated 100 uniformly distributed numbers, and the computed the numbers of another set with some deterministic mapping function. Mapping functions include linear and nonlinear ones, which are listed in Table 1. The variances of mutual information and quadratic mutual information for all the different maps are also listed in Table 1. Since both quadratic mutual information and mutual information are theoretically invariant to one-to-one maps, the variance should be zero. However, compared to the variance of quadratic mutual information that was close to zero, the variance of mutual information was much larger. This invariance was not well maintained for mutual information. The reason mainly lies in the requirement of the quantization process to compute mutual information. In contrast, quadratic mutual information can be computed analytically from data.

Based on the comparison results, we believe that

Table 1 Evaluation with different maps.

	MI	QMI
$y = x$ $y = 2x + 3$ $y = x^2$ $y = 1/(1 + x)$	σ^2=7 × 10^{-2}	σ^2=2 × 10^{-8}

quadratic mutual information can evaluate the correlation more robustly than mutual information does. We thus adopt it to evaluate the audiovisual correlation.

3. Speaker Region Segmentation

We incorporate the analyzed audiovisual correlation into graph cut-based video segmentation to segment a speaker region.

Again using the retrieved N video frames, we build a N-D image as defined in [3] and perform the video segmentation. To avoid a heuristic threshold for this segmentation, we learn the correlation distributions of speaker and background. The segmentation is performed based on the likelihood of each pixel to these two distributions. Note that, since there is only one scalar correlation value $QMI(a; v(x, y))$ at each image position (x, y) computed by Eq. (6), distances are same for all the pixels at (x, y), 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 (x, y) 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 [3]. The global energy function is composed of two important terms: the sum of data costs of all the pixels, and the sum of the smoothness penalties between every two neighboring pixels in both temporal and spatial domains, whose definition is given by

$$E(l) = \sum_{p} D_p(l_p) + \lambda \cdot \sum_{\{p,q\} \in Ne} S_{pq}(l_p, l_q), \tag{16}$$

where l represents the segmentation labels of all the pixels in the N-D image. $l_p = 1$ means pixel p is labeled as speaker, while $l_p = 0$ means background. Ne defines the neighborhood relationship between two pixels, which is discussed in detail in Sect. 3.3. λ is a constant that adjusts the balance between the data costs and the smoothness penalties.

It has been shown that the energy function defined in Eq. (16) can be efficiently optimized by calculating the minimum cut of a graph using a maximum flow algorithm [2]. Moreover, the optimization result is guaranteed to be the global minimum solution of the energy function [1].

3.2 Data Cost by Audiovisual Correlation

To compute the data costs, we first learn the correlation distributions of speaker and background by using expectation maximization algorithm. These two distributions are assumed to be one-dimensional Gaussian, whose parameters are learnt by the process below. First, the highest and lowest audiovisual correlation values are selected as two seeds.

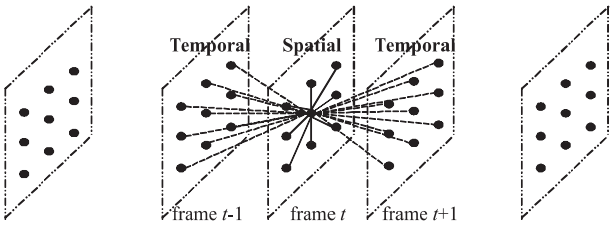

Fig. 5 A demonstration of the N-D image and the neighborhood.

Then, by iteratively applying expectation maximization algorithm to all correlation values, we can compute an optimum estimation of the parameters of the two Gaussian distributions. The one trained from the seed of the lowest correlation is regarded as the distribution of background, denoted as $G(\mu_0, \sigma_0^2)$. The other is regarded as the distribution of speaker, denoted as $G(\mu_1, \sigma_1^2)$.

We find that the number of iterations performs the role of controlling the degree of how high an audiovisual correlation should have a higher likelihood to be speaker than the one to be background. If this number is too high, the background pixels that have relatively high audiovisual correlation may be wrongly segmented as speaker. We empirically find that three iterations are enough for this learning process.

The data cost of each pixel in Eq. (16) is determined by the Mahalanobis distance to the correlation distributions of speaker and background, which is computed as

$$D_p(l_p) = \begin{cases} \frac{(QMI(a;v(x,y))-\mu_1)^2}{\sigma_1^2} & l_p = 1 \\ \frac{(QMI(a;v(x,y))-\mu_0)^2}{\sigma_0^2} & l_p = 0 \end{cases}. \tag{17}$$

3.3 Smoothness Penalties by Image Information

Smoothness penalties are 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 Fig. 5. Each pixel can maximally have 26 neighbors.

The value of smoothness penalty is computed by

$$S_{pq}(l_p, l_q) = \exp\left(-\beta(I_p - I_q)^2\right) \cdot (d(p, q))^{-1} \cdot T[l_p \neq l_q], \tag{18}$$

where p and q are two neighboring pixels. I_p and I_q are their intensity values. The constant β is chosen as in [3] to be

$$\beta = \left(2\left\langle (I_p - I_q)^2 \right\rangle\right)^{-1}, \tag{19}$$

where $\langle \cdot \rangle$ denotes the expectation over the N-D image sample. This choice of β ensures that the exponential term in Eq. (18) switches appropriately between high and low contrast. $d(p, q)$ 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 adopted both simulation and real data to test the performance of our method. All videos of the data were or

were supposed to be filmed at 30 fps, while the audios were sampled at 44.1 kHz, since most recent off-the-shelf video cameras supply such audiovisual data. As for the algorithm parameters, in all our experiments, the balance constant in Graph Cut is set as $\lambda = 20$. The window for optical flow computation is of the size 9×9. The threshold for the texture verification in a window is set as 3. Except the experiments in Fig. 7, we adopt 40 audiovisual frames to compute the results.

As for the computation time, it takes about 31 seconds to do segmentation for 40 frames at a resolution of 240×160 on our laptop, which is equipped with an Intel Core2 1.83 GHz CPU and a 1 GB RAM.

4.1 Simulation

To test the performance of our method when visual and audio signals change following an ideal pattern of a speaking action, we simulated a video clip and applied our method to it. Visual data photographed the movements of a random dot pattern, which was at a resolution of QVGA, 320×240. To include both mouth and background movements, we divided the dot pattern into two parts: a central rectangle face region and a background region. The central rectangle region was set to be slightly lighter than the background, as shown in Fig. 6 (a). Both parts shook vertically. The central region moved synchronously with the audio change to simulate speaking movements. It was realized by computing the vertical shift at each time t by the function $c(t) = \max\{\sin(2\pi f_f t), 0\}$, which was also adopted to modulate the magnitude of the audio. The audio, a 2 kHz modulated sine wave, was shown in Fig. 6 (b). Alternatively, the vertical shift of the remained region was computed by another sine function $c(t) = \max\{\sin(2\pi f_b t), 0\}$.

As the central region simulates mouth movement, f_f was set as $1/0.7$ Hz. For the remained region, we first chose a low frequency as $f_b = 1/2.3$ Hz to simulate a slow change background. The computed audiovisual correlation and the segmentation results are shown in Fig. 6 (c) and (d), respectively. Furthermore, we chose a high frequency as $f_b = 1/0.4$ Hz to simulate a fast change background. The experimental results are shown in Fig. 6 (e) and (f). In both cases, our method detected much higher audiovisual correlation in the central region and successfully segmented the region out.

The quantitative evaluation results were shown in Table 2. The accuracy was computed by an often used measure in image segmentation, which is defined by

$$accuracy = \frac{\sum_{i=1}^{N} C_i}{T}, \tag{20}$$

where C_i is the number of correct segmented pixels in the ith object. T is the total number of pixels. N is equal to 2 in this work. The two objects correspond to foreground and background, respectively. This measure can evaluate the performance of segmentation for both foreground and background. For both video clips, our method achieved high segmentation accuracy.

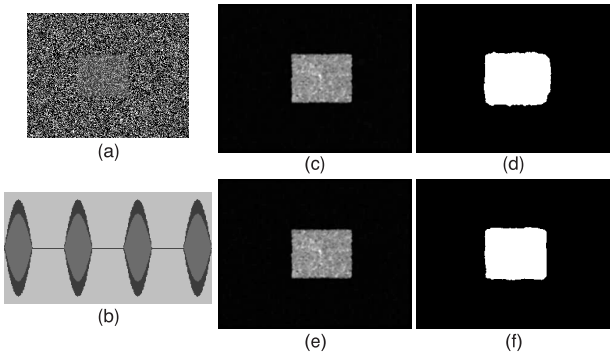

Fig. 6 Segmentation results of simulation data. Figures (a) and (b) respectively show the visual random dot pattern and the audio, figures (c) and (e) show the analyzed audiovisual correlation, and figures (d) and (f) show the mask of the segmented face region.

Table 2 Segmentation performance on simulation data.

	Accuracy (%)
Slow bg	95.9
Fast bg	96.5

4.2 Real Data

For real data, we adopted CUAVE audiovisual database [20], where 17 females and 19 males uttered English numbers in front of a green background with frontal, lateral, and moving views. An advantage of CUAVE database was that we could remove the green background by chroma-key and place 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 to make it possible to perform all the experiments in the 1 GB memory of our laptop.

We first investigated the relationship between the length of the time window and the analyzed audiovisual correlation. The experimental results are shown in Fig. 7. The left person was speaker. While the right person remained quiet. Yet, some ambiguous actions were still posed by him, which can be observed in Fig. 7 (a–c). Analyzed audiovisual correlation were shown in Fig. 7 (d–f). The ambiguous actions posed by silent person may happen to be synchronous with audio also if observed in some short time slices. As we analyze the audiovisual correlation based on synchrony, these places inevitably demonstrate high correlation values also. This problem has been addressed as chorus ambiguity in [13].

However, it can be observed in Fig. 7 (d–f) also that a longer length of the time window helps to remove the ambiguity to determine the current speaker from Fig. 7 (e) to Fig. 7 (f). As a defect, a long time window causes our assumption more possible to be broken since we assume that the speaker remains stationary in the processing time window. To make a good tradeoff of these two aspects, we empirically adopted 40 frames.

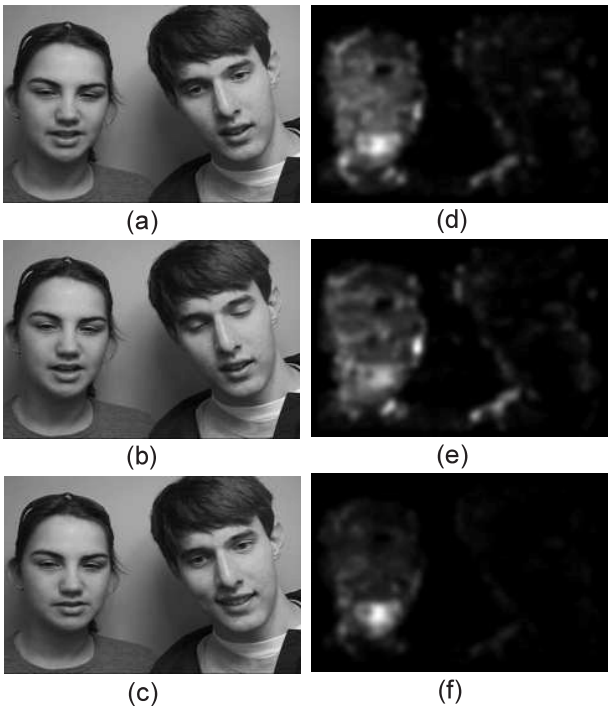

Fig. 7 Statistical audiovisual correlation using different numbers of frames. Figures (a), (b) and (c) show three video frames, and figures (d), (e), and (f) demonstrate the analyzed correlation using 20, 40 and 80 frames. The woman at the left was the actual speaker. While the man at the right side remained quiet, but still posed some actions unconsciously. Correlation values are normalized independently for a better visualization.

We tested the segmentation performance with different backgrounds. The results are shown in Fig. 8 (a–c), inside which only three of the 40 frames are shown. The segmented face region changed marginally under different backgrounds. To perform a comparison, we implemented the method in [3]. The manual seed we designated and the segmentation results over the same 40 frames are shown in Fig. 8 (d–g). The segmented face region changed largely for different backgrounds. Note that this does not mean a comparison of performance, as the results by [3] can be improved iteratively by adding seeds. However, the results here demonstrate that, through fusing audio information, our method can achieve better robustness to the change of background.

We tried to detect the speaker's face region with both frontal and lateral views. The results are shown in Fig. 9. 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 tested our method when visual scale and audio gain were changed. The results are shown in Fig. 10. As our method is adaptive to the scale or gain change, uniform segmentation results were achieved.

We also applied our method to other video clips in CUAVE database, where single or multiple persons were

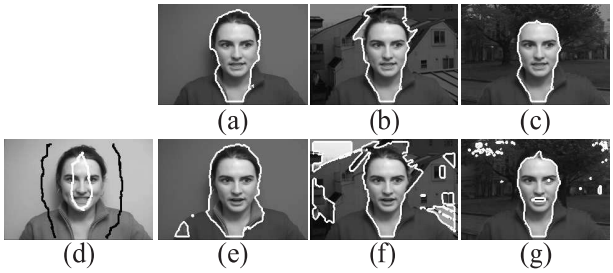

Fig. 8 Estimated results by the method in [3] and ours. The darkened areas represent the region of estimated background. The pixels located at the boundary between speaker and background are colored as white. Figures (a), (b) and (c) show the segmentation results of our method with different non-stationary backgrounds, figure (d) shows our designated segmentation seeds, and figures (e), (f) and (g) show the segmentation results by the method in [3].

Fig. 9 Segmentations for different views. Figures (a) and (c) show the segmented face region for a frontal view, and figures (b) and (d) show the results for a lateral view.

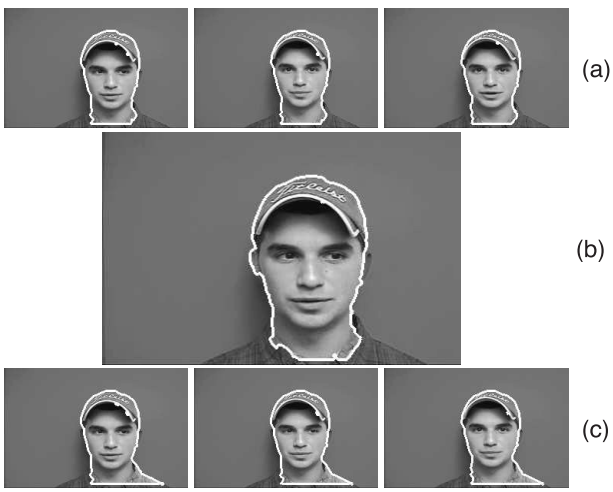

Fig. 10 Segmentation for different visual scales and audio gains. Figure (a) shows the results with visual resolution 240×160 and original audio data, and figure (b) shows the results when the visual resolution was increased to 360×240, i.e., visual scale was changed to 1.5 times. Figure (c) shows the results when original audio was gained by 3.5 dB, i.e., audio magnitude was increased by 1.5 times.

photographed. The backgrounds of some clips were intentionally replaced into complex ones to increase the difficulty of face region detection. Additionally, in the case of multiple persons, we applied to our method to different time windows within which different person was talking. The experimental results are shown in Fig. 11. In most situations, our method successfully found out the speaker's face region within the time window when it was applied,

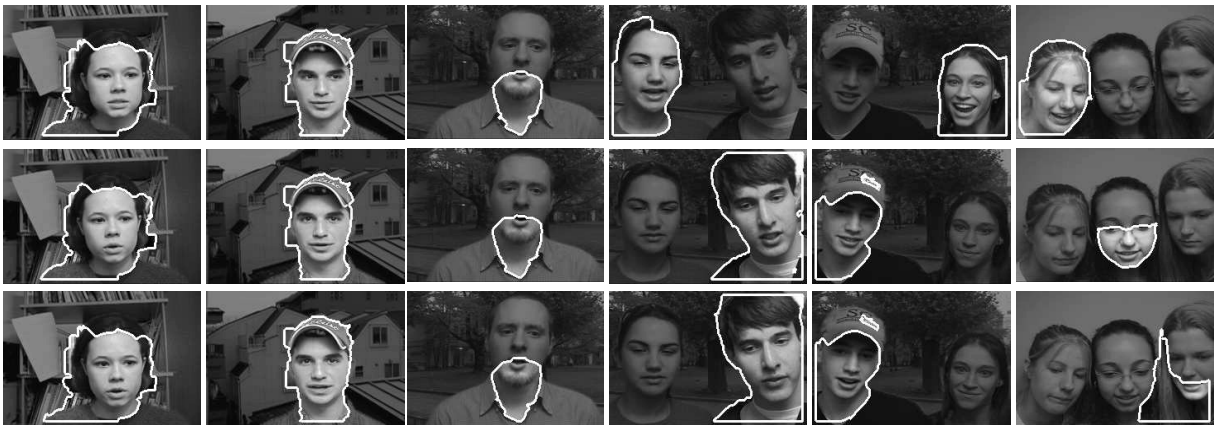

Fig. 11 The experimental results for other persons. Figures (a), (b) and (c) show the results for a single person, and figures (d), (e) and (f) show the results for multiple persons within different time windows.

Fig. 12 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.

except Fig. 11 (c), where our method only segmented out the speaker's mouth region. The reason lies in that the man in (c) intentionally restrained the movements of all his face parts other than his mouth when he was speaking. His speaking manner thus looked a little unnatural, which may come from the tension of being before a camera. As discussed in Sect. 1, our method can localize the speaker's mouth region in such cases.

To give a quantitative evaluation of our detection result, we have manually labelled the face regions for the first frame of four video sequences. Considering that additional error will be introduced by replacing the green background with other complex ones, we adopted the original CUAVE data only to do this test. Labelled ground truth and the detection accuracy of our method were shown in Fig. 12. Segmentation accuracy was computed by Eq. (20). The results of this quantitative evaluation in Fig. 12 demonstrated that, in most cases, our method can extract the speaker's face region with high accuracy.

5. Conclusions and Future Works

In this work, we have developed a method to find out the speaker's face region within time windows, which is robust against the changes of view, scale, and background. The main thrust of our idea was to integrate audiovisual correlation analysis into graph cut-based video segmentation. We have shown that our method is capable of finding less fragmented face regions than previous methods for both single and multiple persons under different conditions.

Our current evaluation of audiovisual correlation is sensitive to the noise. Visual noise may yield incorrect optical flow in untextured regions. While Audio noise may disturb the frame energy estimation and make the audiovisual correlation inaccurate. We plan to try other reliable methods to compute optical flow and test 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, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," IEEE Trans. Pattern Anal. Mach. Intell., vol.23, no.11, pp.1222–1239, 2001.
[2] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," IEEE Trans. Pattern Anal. Mach. Intell., vol.26, no.9, pp.1124–1137, 2004.
[3] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," Int. J. Comput. Vis., vol.70, no.2, pp.109–131, 2006.
[4] A.L. Casanovas, Blind audiovisual source separation using sparse redundant representations, Master thesis, Signal Processing Institute, EPFL, 2006.
[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Conf. on Computer Vision and Pattern Recognition, vol.2, pp.886–893, 2005.
[6] J. Doornik and H. Hansen, "An omnibus test for univariate and multivariate normality (Working paper)," NuOEeld College, Oxford, 1994.
[7] J. Driver, "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading," Nature, vol.381, pp.66–68, 1996.
[8] J.W. Fisher, III and T. Darrell, "Speaker association with signal-level audiovisual fusion," IEEE Trans. Multimed., vol.6, no.3, pp.406–413, 2004.
[9] J. Hershey and J.R. Movellan, "Audio vision: Using audiovisual synchrony to locate sounds," NIPS, pp.813–819, 1999.
[10] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," Proc. IEEE Int'l Conf. on Computer Vision, vol.1, pp.446–453, 2005.
[11] K. Kanatani, "Motion segmentation by subspace separation and model selection," IEEE Int'l Conf. on Computer Vision, vol.2, pp.586–591, 2001.
[12] M. Kass, A. Witkin, and D. Terzolpoulos, "Snakes: Active contour models," Int. J. Comput. Vis., vol.1, no.4, pp.321–331, 1988.
[13] E. Kidron, Y.Y. Schechner, and M. Elad, "Pixels that sound," IEEE Conf. on Computer Vision and Pattern Recognition, pp.88–95, 2005.
[14] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother, "Bi-layer segmentation of binocular stereo video," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol.2, pp.1186–1194, 2005.
[15] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," Int. Joint Conf. on Artificial Intelligence, pp.674–679, 1981.
[16] J. Luettin, N. Thacker, and S. Beet, "Speaker identification by lipreading," Proc. Int. Conf. on Spoken Language, vol.1, pp.62–65, 1996.
[17] M. Matsumoto and T. Nishimura, "Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator," ACM Trans. Model. Comput. Simul., vol.8, no.1, pp.3–30, 1998.
[18] G. Monaci, O. Escoda, and P. Vandergheynst, "Analysis of multimodal signals using redundant representations," Int. Conf. on Image Processing, pp.145–148, 2005.
[19] E. Parzen, "On the estimation of probability density function and the mode," The Annals of Mathematical Statistics, vol.33, no.3, pp.1065–1076, 1962.
[20] 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. Applied Signal Processing, vol.2002, no.11, pp.1189–1201, 2002.
[21] L. Rabiner and B. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.
[22] A. Renyi, "On measures of entropy and information," Fourth Berkeley Symp. Math. Stat. and Probability, vol.1, pp.547–561, 1961.
[23] T. Schoenemann and D. Cremers, "High resolution motion layer decomposition using dual-space graph cuts," IEEE Conf. on Computer Vision and Pattern Recognition, pp.1–7, 2008.
[24] P. Smaragdis and M. Casey, "Audio/visual independent components," Proc. Int. Symposium on Independent Component Analysis and Blind Source Separation, pp.709–714, 2003.
[25] B.A. Turlach, "Bandwidth selection in Kernel density estimation: A review," CORE and Institut de Statistique, pp.23–493, 1993.
[26] P. Viola and M. Jones, "Robust real-time face detection," Int. J. Comput. Vis., vol.57, no.2, pp.137–154, 2004.
[27] D. Xu, J. Principe, and J. Fisher, "A novel measure for independent component analysis (ICA)," Int. Conf. on Acoustics, Speech and Signal Processing, vol.2, pp.1161–1164, 1998.
[28] T. Yu, C. Zhang, M. Cohen, Y. Rui, and Y. Wu, "Monocular video foreground/background segmentation by tracking spatial-color Gaussian mixture models," Proc. IEEE Workshop on Motion and Video Computing, pp.55–63, 2007.
[29] S.C. Zhu and A. Yuille, "Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol.18, no.9, pp.884–900, 1996.

Appendix: Scale Invariance

In this appendix, we show that our method to analyze the audiovisual correlation is invariant to the change of scale.

First, we show that the change of the scale of a one-dimensional random variable z leads to the multiplication of a coefficient to the original pdf only. Suppose that z is multiplied by a scale coefficient of s, i.e., $z_s = sz$. The new bandwidth σ_s estimated by Eq. (5) becomes

$$\sigma_s = 1.06 s\hat{\sigma} n^{-\frac{1}{5}} = s\sigma. \quad (\text{A}\cdot 1)$$

Substituting Eq. (A· 1) into Eq. (4), $p(sz)$ can be represented by $p(z)$ as

$$\begin{aligned} p(sz) &= \frac{1}{N}\sum_{i=1}^{N} K_{\sigma_s}((sz - sz_i)) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}s\sigma} \exp\left(-\frac{s^2(z-z_i)^2}{2s^2\sigma^2}\right) \\ &= \frac{1}{s}p(\mathbf{z}). \end{aligned} \quad (\text{A}\cdot 2)$$

This conclusion can be easily extended to the n-dimensional case. Since the bandwidth matrix $\mathbf{H}$ is supposed to be diagonal, i.e., $\mathbf{H} = diag(\sigma_1, \cdots, \sigma_n)$, we have

$$\begin{aligned} p(\mathbf{sz}) &= \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{n} K_{\sigma_{s_j}}(s_j z - s_j z_{ij}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\prod_{j=1}^{n}\frac{1}{s_j}\right) K_{\mathbf{H}}(\mathbf{z} - \mathbf{z}_i) \\ &= \left(\prod_{j=1}^{n}\frac{1}{s_j}\right) p(\mathbf{z}), \end{aligned} \quad (\text{A}\cdot 3)$$

where $\mathbf{s}$ is a diagonal scale matrix, $\mathbf{s} = diag(s_1, \cdots, s_n)$.

Then, we show that quadratic mutual information is invariant to the change of the scale of audiovisual features.

Suppose that the scale of the audio feature a is s_a, and the scale of the visual feature v is s_v. Substituting Eq. (A· 2) into Eq. (6), we have

$$\begin{aligned} & QMI(s_a a; s_v v) \\ = & \log \frac{\iint \frac{1}{s_a^2 s_v^2} p^2(a,v) da dv \iint \frac{1}{s_a^2 s_v^2} p^2(a) p^2(v) da dv}{(\iint \frac{1}{s_a^2 s_v^2} p(a,v) p(a) p(v) da dv)^2} \\ = & QMI(a; v). \end{aligned} \tag{A· 4}$$

Consequently, our method to analyze the audiovisual correlation is invariant to the change of scale.

Yuyu Liu received the BS degree in communication engineering from Beijing University of Posts & Telecommunications in 2000, the MS degree in electronic engineering from Tsinghua University in 2003, and the Ph.D degree in computer vision from The University of Tokyo in 2009. He received the Best Industry Related Paper Award (BIRPA) of the International Conference on Pattern Recognition in 2008. He is currently a researcher in Sony Research China. His research interests include audiovisual analysis, image segmentation, pattern recognition, and machine learning.

Yoichi Sato is an associate professor jointly affiliated with the Graduate School of Interdisciplinary Information Studies, and the Institute of Industrial Science, at the University of Tokyo, Japan. He received the BSE degree from the University of Tokyo in 1990, and the M.S. and Ph.D. degrees in robotics from the School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, in 1993 and 1997 respectively. His research interests include physicsbased vision, reflectance analysis, imagebased modeling and rendering, tracking and gesture analysis, and computer vision for HCI.