Visual Localization of Non-stationary Sound Sources

Yuyu Liu
The University of Tokyo
4-6-1 Komaba, Meguro-ku
Tokyo, 153-8505 JAPAN
liuyuyu@iis.u-tokyo.ac.jp

Yoichi Sato
The University of Tokyo
4-6-1 Komaba, Meguro-ku
Tokyo, 153-8505 JAPAN
ysato@iis.u-tokyo.ac.jp

ABSTRACT

Sound source can be visually localized by analyzing the correlation between audio and visual data. To correctly analyze this correlation, the sound source is required to be stationary in a scene to date. We introduce a technique that localizes the non-stationary sound sources to overcome this limitation. The problem is formulated as finding the optimal visual trajectories that best represent the movement of the sound source over the pixels in a spatio-temporal volume. Using a beam search, we search these optimal visual trajectories by maximizing the correlation between the newly introduced audiovisual features of inconsistency. An incremental correlation evaluation with mutual information is developed here, which significantly reduces the computational cost. The correlations computed along the optimal trajectories are finally incorporated into a segmentation technique to localize a sound source region in the first visual frame of the current time window. Experimental results demonstrate the effectiveness of our method.

Categories and Subject Descriptors

I.2.10 [**Computing methodologies**]: Artificial intelligence—*Vision and scene understanding*

General Terms

Algorithms

1. INTRODUCTION

The ability to visually localize sound source captured by a camera is useful for various applications. For instance, the pan-tilt camera used with a video conferencing system can be controlled to follow a speaker. An interviewee could be automatically overlaid with mosaics to protect privacy. Other applications of sound source localization include surveillance, video analysis, and audio-to-video synchronization adjustment.

Microphone arrays are commonly used for sound source localization. However, the use of such special devices severely limits the applicability of techniques based on this approach. For this reason, much attention has been put on techniques that are based on audiovisual correlation analysis [1, 2, 4, 5, 6, 9], which can localize sound source with only one microphone.

Originated from the discovery that audiovisual correlation lies in synchrony [3], previous works in this field have concentrated on how to computationally analyze this audiovisual correlation, where different audiovisual features [1, 2, 6, 9] and correlation measures [1, 2, 4, 5, 6, 7, 9] have been developed or introduced.

However, all of the existing techniques share a common limitation in that they cannot be used for non-stationary sound sources. When a sound source moves, visual features computed at a fixed position in different frames no longer correspond to the same point of this sound source. The existing techniques choose to ignore this problem by assuming that sound source is stationary within a given time window.

In this work, we develop a method to correctly analyze the audiovisual correlation for non-stationary sound sources. Within each time window, optimal visual trajectories starting from the pixels in the first frame are independently searched by maximizing the audiovisual correlation between the features extracted from local patches. The visual trajectory found in this search is regarded as the best possible translation of that pixel following the movement of the non-stationary sound source. The correlations of the pixels analyzed following their optimal visual trajectories are incorporated into a segmentation technique as [6] to localize the sound source region in the first visual frame. By shifting the time window, the sound source region in other frames can also be localized.

Two aspects drive our technique. First, we developed a method to efficiently search for optimal visual trajectory by using a beam search with an incremental analysis of the audiovisual correlation. Second, we introduce the inconsistency as an audiovisual feature to robustly analyze the magnitude of acceleration in a local patch.

The rest of this paper is organized as follows. In Sec. 2, we give an overview of our method. In Sec. 3, we introduce the audiovisual feature of the inconsistency. In Sec. 4, we explain the incremental analysis of correlation. In Sec. 5, we demonstrate and discuss the experimental results, and we present our conclusions in Sec. 6.

MM'09, October 19–24, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-608-3/09/10 ...$10.00.

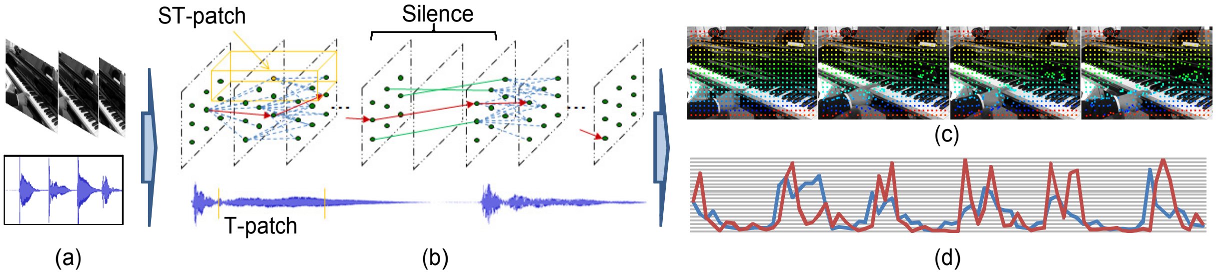

Figure 1: Audiovisual correlation maximization. (a) shows original audiovisual data, (b) demonstrates search of visual trajectory, (c) figures optimal visual trajectories starting from different pixels (differently colored), and (d) audio (red) and visual (blue) features following one of the optimal visual trajectories.

2. OUTLINE OF OUR METHOD

First, we compute the visual inconsistency feature in each local Spatio-Temporal patch (ST-patch), and the audio inconsistency feature in each local Temporal patch (T-patch). The audiovisual features are explained in detail in Sec. 3. Examples of an ST-patch and a T-patch are illustrated in Fig. 1 (b).

Second, we search for the visual trajectory that maximizes the correlation between the visual and audio features by using a beam search. Given the number of beams L and the search range of a pixel between two successive frames d, the beam search orders the correlation of the $L(2d+1)^2$ possible visual trajectories in every frame and retains only the L best ones to begin from in the next frame as illustrated in Fig. 1 (b). All the pixels in the first frame are regarded as starting points and independently searched till the last frame. An example of some search results is shown in Fig. 1 (c).

Special care needs to be taken for silent frames since the absence of auditory information makes it uninformative to analyze the audiovisual correlation. Silent intervals can be detected using the method in [6]. During the silent interval, the L visual trajectories are propagated with sub-pixel accuracy following optical flows computed by Lucas-Kanade method [8]. The beam search is then resumed at the ending frame of each silence interval as demonstrated in Fig. 1 (b). Note that the audiovisual feature is not extracted in the silent intervals. These frames are also not included in the audiovisual correlation analysis, either.

Finally, we detect the sound source region in the first visual frame using the technique developed in [6], which evaluates the likelihood of each pixel as the sound source based on the analyzed audiovisual correlation and segments the sound source region out by graph cut.

3. AUDIOVISUAL FEATURE

Differential features, like velocity and acceleration, have recently attracted a lot of interest [1, 6, 9]. We believe that audiovisual correlation should be analyzed between the acceleration of the visual motion and that of the audio energy change. For instance, when a human beats a drum, sound is generated when a hand hits the drum. There are sudden changes in both the velocity of the hand and the energy of the sound, which implies the existence of acceleration. Playing pianos, walking, and so on follow this pattern. Speaking is similar as well. Although the way that a voice is generated is far more complex than beating a drum, it can still be regarded as a sound jointly modulated by the throat, tongue, teeth, and lips [12]. When lip movements are accelerated, the energy of the modulated sound changes simultaneously.

Therefore, we base the audiovisual feature on the evaluation of the magnitude of acceleration, which is from the concept of motion inconsistency. As demonstrated in Fig. 1 (d), high synchrony can be observed with our feature for a hand playing a piano.

3.1 Visual inconsistency

The usual way to compute the acceleration of visual motion [1] is to use visual tracking [8] to estimate the visual translation first and then compute the acceleration. However, visual tracking is not stable in low textured areas and usually can be applied to several featured points only [1].

For this reason, our method base the visual feature on the concept of motion inconsistency computed in a local ST-patch. Motion inconsistency was first introduced by Shechtman and Irani in [14], which measures the degree of the moving direction change in a local ST-patch. The size of a ST-patch was recommended to be 7×7×3. The degree of motion inconsistency can be evaluated robustly [14].

It is important to note that motion inconsistency is closely related to the magnitude of acceleration of an object. If the direction of movement is altered in a ST-patch, an inconsistent motion will be detected. The more the direction is altered, the higher the degree of inconsistency. Accordingly, motion inconsistency can be used for measuring the magnitude of acceleration.

Therefore, we define the visual feature $v(x, y, t)$ as the degree of motion inconsistency computed in a ST-patch centered at (x, y, t), which is given by

$$v(x, y, t) = (\lambda_2 \lambda_3)/(\lambda_1^{\diamond} \lambda_2^{\diamond}), \tag{1}$$

where λ_2 and λ_3 are the second and the third eigenvalue of a 3×3 gradient matrix $\mathbf{M}$. $\lambda_1^{\diamond}$ and $\lambda_2^{\diamond}$ are the first and the second eigenvalue of its top left 2×2 submatrix $\mathbf{M}^{\diamond}$. It can be shown that the value of the visual feature is normalized [14]. The matrix $\mathbf{M}$ is defined as

$$\mathbf{M} = \begin{pmatrix} \sum I_x^2 & \sum I_x I_y & \sum I_x I_t \\ \sum I_y I_x & \sum I_y^2 & \sum I_y I_t \\ \sum I_t I_x & \sum I_t I_y & \sum I_t^2 \end{pmatrix}, \tag{2}$$

where I_x, I_y, and I_t respectively denote the partial derivative $\partial I/\partial x$, $\partial I/\partial y$, and $\partial I/\partial t$ of the intensity at each pixel of the ST-patch. In the implementation, it was recommended [14] that weighted gradient sums should be used instead of the regular ones in Eq. (2). A fast computation of Eq. (1) was also proposed.

Note that there is another possibility that leads to inconsistent motion, as also mentioned in [14]. When an ST-patch is located at the boundary of two different motion fields, it has inconsistent motion. It was claimed [14] that this is negligible considering the minor proportion which it holds of the total pixels. Furthermore, it casts fewer effects on our system as the boundaries always demonstrate a high inconsistency and are less correlated with the audio. For these reasons, this point is disregarded in this work, too.

3.2 Audio inconsistency

Similarly to the visual feature, the audio feature is defined based on the inconsistency of audio energy change in a local T-patch. First, we compute the audio energy $e(t)$ in frame t using the method proposed in [6]. Second, we evaluate the degree of inconsistency in the change of audio energy in a local T-patch. By applying the same consideration of [14] to audio energy, we deduce out a method to compute audio inconsistency in a T-patch. The audio feature is defined as this audio inconsistency, which is given by

$$a(t) = \lambda_2/\lambda_1^{\diamondsuit}, \tag{3}$$

where λ_2 is the second eigen-value of a gradient matrix $\mathbf{Q}$. $\lambda_1^{\diamondsuit}$ is the only element of its top left 1×1 sub-matrix $\mathbf{Q}^{\diamondsuit}$. The matrix $\mathbf{Q}$ is defined as

$$\mathbf{Q} = \begin{pmatrix} \sum 1 & \sum e_t \\ \sum e_t & \sum e_t^2 \end{pmatrix}, \tag{4}$$

where e_t denotes the derivative de/dt of audio energy at each frame in the T-patch. It can be shown that the value of the audio feature is also normalized. Additionally, weighted gradient sums were used in the implementation instead of the regular ones in $\mathbf{Q}$.

4. INCREMENTAL ANALYSIS OF AUDIO-VISUAL CORRELATION

Audiovisual correlation is analyzed between the quantized audiovisual features. Since both features are normalized, we uniformly quantize the interval of [0 1] into C discrete levels, where C is fixed to 20 in this work. We have tried other C values but observed only minor changes when analyzing audiovisual correlation.

The correlation is evaluated by using mutual information. Although it is possible to compute mutual information by its definition [13], this calls for a lot of computation time since the correlation is evaluated millions of times in the maximization process. We develop a method to compute mutual information incrementally in our beam search, which significantly speeds up the process. We would like to mention that this incremental computation is general and can be used in other situations where mutual information needs to be computed in multiple stages.

We explain the incremental computation of mutual information using entropy. Since mutual information can be divided into the sum of entropies following $MI(a;v) = H(a) + H(v) - H(av)$, conclusions here can be easily applied to mutual information. The entropy of a discrete random variable z can be computed based on its histogram $h(z)$ by definition [13]. Suppose that we have the entropy computed in frame k as $H^{(k)}(z)$ and the histogram $h_z^{(k)}$ that has accumulated k samples. In frame $k+1$, a new sample is added to histogram $h_z^{(k)}$, which results in a new entropy $H^{(k+1)}(z)$. Following the equation deductions, we can represent $H^{(k+1)}(z)$ based on the known $H^{(k)}(z)$ as

$$H^{(k+1)}(z) = \frac{1}{k+1}\left(kH^{(k)}(z) - \log\frac{(1+n)^{1+n}}{n^n}\right) + C(k), \tag{5}$$

where $n = h_z^{(k)}(i^{(k+1)})$ is the number of the samples in the bin $i^{(k+1)}$ of $h_z^{(k)}$, $i^{(k+1)}$ is the index of the sample added in frame $k+1$, and $C(k) = \frac{k}{k+1}\log k - \log\frac{k}{k+1}$ is a coefficient computed with k.

A problem of division by zero exists in Eq. (5), which causes the incremental quantity undefined when $n = 0$. Yet, if we take n as a continuous variable, the limitation exists when n approaches zero. It can be proved that $\lim_{n\to 0}\frac{(1+n)^{1+n}}{n^n} = 1$. We adopt this limitation to define the value when $n = 0$ as $\log 1 = 0$.

We can adopt Eq. (5) to incrementally compute entropy and then mutual information. As the incremental quantity $\log\frac{(1+n)^{1+n}}{n^n}$ is related to only one of the histogram bins, to compute $H^{(k+1)}(z)$ by using Eq. (5) is significantly faster than to compute it by definition. Since the histogram has 400 bins in our work, this speed-up is 400 times.

Furthermore, since $H(a)$ is invariant to the search of visual trajectories, minimizing $H(v|a) = H(av) - H(v)$ is equal to maximizing $MI(a;v)$. Substituting Eq. (5) into $H(av) - H(v)$, we can get the computation of $H(v|a)$ as

$$H^{(k+1)}(v|a) = \frac{1}{k+1}\left(kH^{(k)}(v|a) + \log\frac{\frac{(1+n)^{1+n}}{n^n}}{\frac{(1+m)^{1+m}}{m^m}}\right), \tag{6}$$

where $n = h_v^{(k)}(i^{(k+1)})$, and $m = h_{av}^{(k)}(i^{(k+1)}, j^{(k+1)})$. $j^{(k+1)}$ is the index of the quantized audio feature in stage $t+1$. The coefficient $C(k)$ in Eq. (5) is cancelled in the subtraction. As the computation of $H(v|a)$ is simpler than the one of $MI(a;v)$, we in fact minimize $H(v|a)$ in the beam search instead of maximizing $MI(a;v)$.

5. EXPERIMENTAL RESULTS

In this section we present the experimental results from using our method. We used the same parameters for all the experiments. The length of the time window was three seconds. The search range and beam number in the beam search were $d = 1$ and $L = 10$, respectively. The video was sampled at 30 fps and converted to monochrome at a resolution of 240×160, while the audio was sampled at 44.1 KHz.

The localization results of non-stationary sound sources were demonstrated in Fig. 2. The two clips in Fig. 2 captured both a sound source (a hand or a walking man) and an ambiguous moving object (a rotating cover or a man riding a bicycle). Both sound sources were successfully localized. Interestingly, the positions that corresponded to the obscure reflections of the hand also demonstrated fragmentally high audiovisual correlation in Fig. 2 (b). This is reasonable since both movements were synchronous with the audio change. Similar phenomenon was also discussed in [5].

Speaker is an important class of sound sources. We used a CUAVE database [11] to test the performance of our method for speaker localization.

The localization results of non-stationary speakers were demonstrated in Fig. 3. Our method successfully found the facial region of the speaker. As a comparison, we imple-

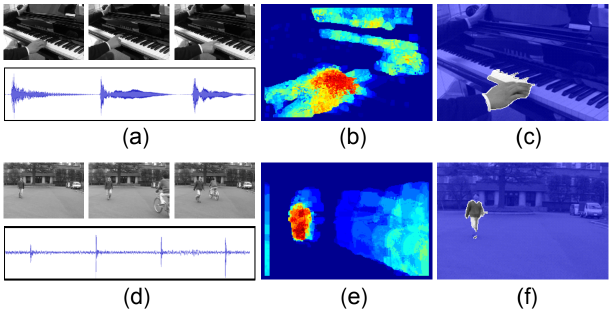

Figure 2: Localizations of non-stationary sound sources. Figures (a) and (d) show the original data. Figures (b) and (e) visualize the analyzed audiovisual correlation with jet color map. The redder a pixel, the higher its correlation. Figures (c) and (f) show the sound source region localized.

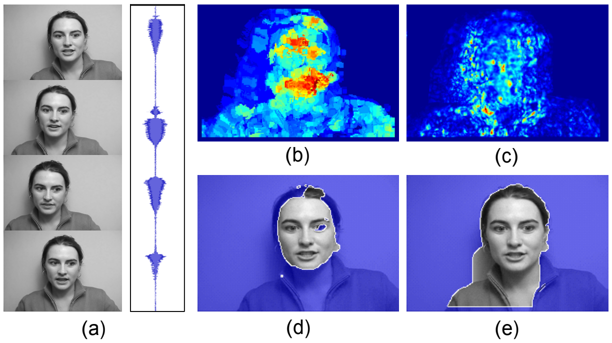

Figure 3: Localization of non-stationary speaker. Figure (a) shows the original audiovisual data. Figures (b) and (d) show the analyzed audiovisual correlation and localization results of our method, respectively. Figures (c) and (e) show the results when using the method in [6].

mented the method in [6], which was designed to localize stationary speakers, and applied it to this clip. The results are shown in Fig. 3 (c) and (e), where we can observe how the analysis of audiovisual correlation failed for a non-stationary speaker and resulted in a wrong localization.

We can detect the sound source at different times by applying our method to different time windows. The speaker localization results from multiple persons were demonstrated in Fig. 4. Our method localized the current speaker.

Our method has to make a tradeoff between the tolerable moving speed determined by the search range and the computation time. To accommodate for fast movement, we need to set a large d and L. Yet, the computation time rises when d and especially L are increased. If we have many sound frames that need to be searched by the beam search, the rising speed is fast. For example, Fig. 2 (a) has 69 sound frames in the 90-frame time window, which is the largest number in all the experimental data. The computation on our desktop, which has an Intel Core2 2.6 G CPU and a 3 G memory, took 87 seconds when $d = 1$ and $L = 10$, 195 seconds when $d = 2$ and $L = 10$, and 457 seconds when $d = 2$ and $L = 20$.

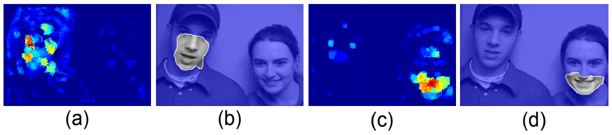

Figure 4: Speaker localization of different time windows. Figures (a) and (c) show the analyzed audiovisual correlation, and (b) and (d) show the localization results.

6. CONCLUSION AND FUTURE WORK

We have developed a method to visually localize non-stationary sound sources by searching for the movements that maximize the audiovisual correlation. The search is efficiently conducted by using a beam search with the incremental analysis of the audiovisual correlation. We have also introduced inconsistency as an audiovisual feature. Our method is capable of localizing different kinds of sound sources for different time windows.

There is a tradeoff in our current method between the computation time and the tolerable motion speed, as discussed in Sec. 5. We are considering using a coarse-to-fine approach to resolve this problem.

7. REFERENCES

[1] Z. Barzelay and Y. Schechner. “Harmony in Motion”. In *Proc. CVPR*, pp.1–8, 2007.

[2] T. Darrell and J. Fisher III. “Speaker Association with Signal-Level Audiovisual Fusion”. *IEEE Trans. on Multimedia*, 6(3):406–413, 2004.

[3] J. Driver. “Enhancement of Selective Listening by Illusory Mislocation of Speech Sounds due to Lip-Reading”, *Nature*, 381:66–68, 1996.

[4] J. Hershey and J. R. Movellan. “Audio Vision: Using Audiovisual Synchrony to Locate Sounds”. In *Proc. NIPS*, pp.813–819, 1999.

[5] E. Kidron, Y. Schechner, and M. Elad. “Pixels that Sound”. In *Proc. CVPR*, pp.88–95, 2005.

[6] Y. Liu and Y. Sato. “Finding Speaker Face Region by Audiovisual Correlation”. In *Proc. ECCV Workshop*, pp.1–12, 2008.

[7] Y. Liu and Y. Sato. “Recovering Audio-to-Video Synchronization by Audiovisual Correlation Analysis”. In *Proc. ICPR*, pp.1–4, 2008.

[8] B. Lucas and T. Kanade. “An Iterative Image Registration Technique with an Application to Stereo Vision”. In *Proc. Int’l Joint Conf. on Artificial Intelligence*, pp.674–679, 1981.

[9] G. Monaci and P. Vandergheynst. “Audiovisual Gestalts”. In *Proc. CVPR Workshop on Perceptual Organization in Computer Vision*, pp.1–8, 2006.

[10] A. O’Donovan, R. Duraiswami, and J. Neumann. “Microphone Arrays as Generalized Cameras for Integrated Audio Visual Processing”. In *Proc. CVPR*, 1–8, 2007.

[11] E. Patterson, S. Gurbuz, Z. Tufekci, and J. 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.

[12] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[13] C. Shannon. “Prediction and Entropy of Printed English”. *The Bell System Technical Journal*, 30:50-64, 1951.

[14] E. Shechtman and M. Irani. “Space-Time Behaviour-Based Correlation”. *Trans. on Pattern Analysis and Machine Intelligence*, 29(11):2045–2056, 2007.