Paper

Relating Audio-Visual Events Caused by Multiple Movements: In the Case of Entire Object Movement and Sound Location Change

Jinji Chen* Non-member
Toshiharu Mukai** Non-member
Yoshinori Takeuchi*,** Non-member
Tetsuya Matsumoto* Non-member
Hiroaki Kudo* Member
Tsuyoshi Yamamura*** Non-member
Noboru Ohnishi* Member

Relating audio-visual events is important for constructing an artificial intelligent system, which can acquire the audio-visual knowledge of moving objects through active observation without a supervisor. This paper proposes a method for relating multiple audio-visual events observed by a camera and a microphone according to general laws without object-specific knowledge, which copes with including entire object movement and sound location change. As corresponding cues, we use Gestalt's grouping law; simultaneity of sound onsets and changes in movement, similarity of repetition between sound and movement. Based on the correlation coefficient between auditory and visual sequences, the component of frequency at sound onset is related to the spatiotemporal invariant sequence (STI sequence) of movement. We experimented in the real environment and obtained satisfactory results showing the effectiveness of the proposed method.

Keywords: sensor fusion, event correspondence, spatiotemporal invariant sequence, motion invariant, repetition similarity, occurrence simultaneity

1. Introduction

Several-month-old infant can relate speech sound (audio information) to its mouth movement (visual information) [1]. It is necessary for an artificial intelligent system to correspond and integrate multi-modal sensor information in order to realize the function of sensor fusion and the automatical acquisition of knowledge without being taught like humans. In this paper, we'll focus on the senses of sight and hearing, which are typical and important senses for all us. We will also consider how to find the correspondence of events observed by visual and auditory senses. It can be used to realize a sensory substitution system for hearing or sight handicapped persons to understand their environment. The system presents the visual (audio) information in language about corresponded events.

The purpose of this paper is to relate multiple audio-visual events observed by a camera and a microphone according to general laws without object-specific knowledge. Namely this paper aims to obtain the knowledge of audio-visual information of movement automatically, and to understand the environment by observation. As corresponding cues, we use Gestalt's grouping laws: simultaneity of sound onsets and direction changes of movement, and similarity of repetition between sound and movement.

Over last few years, a few studies of audio-visual fusion in machine learning have been reported. Suyama et al. [2] developed a method for finding a face in images uttering speech based on the cue signal. The method uses multiple-microphones and relies on the relation between mouth movement and voice level. Mukai et al. [3] focused only on a periodic movement in the scene. Hayakawa et al. [4] extended Mukai's method to find the correspondence in the scene where one sound and multiple movements exist. But these two studies have a limitation of the number of movements and sounds.

In our recent research [5], we corresponded the multiple audio-visual events, but we haven't treated the case of audio-visual events caused by entire object movement or by sound location change. In this paper, we propose a solution of this problem by using the spatiotemporal invariant (STI) [6]. When we observe the movement of the same object at different positions, we observe the different movement. But there exists an invariant in movement independent of the viewpoint.

This paper is organized as follows. Audio-visual events

* CIAIR Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464-8603
** Bio-Mimetic Control Research Center, RIKEN,
2271-130 Anagahora, Moriyama-ku, Nagoya
*** Aichi Prefectural University,
1522-3 Aza Ibaragabasama, Oaza Kumabari, Nagakute-cho, Aichi-gun, Aichi, 480-1198

Table 1. Relations between movement and sound.

	Location of sound changes	Location of sound does not change
Object moves	Foot stepping	—
Object does not move	Clapping with rotation	Metronome

and corresponding cues are described in section 2. In section 3, we explain the concept of spatiotemporal invariant. In section 4, we show how to find audio-visual event correspondence. In section 5, we describe several experiments as well as its results and discussion. Finally, we conclude this paper by mentioning future subjects in section 6.

2. Audio-Visual Events and Corresponding Cues

2.1 Relations Between Movement and Sound

Sound generation requires some energy, which is mostly supplied by movement. Thus, the decrease of kinetic energy due to collision is changed into other energy forms such as sound and heat according to the law of energy conservation in physics. In Table 1, we summarize the relation between object movement and sound from the viewpoint of whether the entire object moves or not, and whether the location of sound source changes or not.

In our recent research [5], we can't treat the case in which entire object moves and the sound location changes. In this paper, we study all of the situations expressed in Table 1.

2.2 Clues of Audio-Visual Correspondence

In psychology, we tend to perceptually group elements in stimuli. This is known as the Gestaltist's perceptual organization (or perceptual grouping). Gestalists proposed the laws for perceptual grouping. The law of common fate and the law of similarity are ones of these laws. We tend to group stimuli with the same characteristics. This is the law of similarity. If the stimuli change or move together, we also tend to group these stimuli. This is the law of common fate. Though Gestalt's laws have been proposed only in one sensation, we extend this concept to correspond events among different types of sensation.

We summarized clues for relating visual and auditory information along with Gestalt's grouping factors corresponding to each cue [4]. In this paper, we only use two clues to determine the correspondence of audio-visual events. One is the simultaneity of change in movement and sound onset. This clue corresponds to the law of common fate. The instance of sound occurrence generally coincides with that of moving direction change. Another is similarity of repetition between sound and movement. This clue corresponds to the law of similarity. If a movement is repetitive, the corresponding sound has the same repetition as the movement.

3. Spatiotemporal Invariant (STI)

When we observe the moving object from different viewpoints, the observed movement are different. Sato [6] reported that there exits one kind of invariant, spatiotemporal invariant (STI), in the movement that does not depend on viewpoint. In this paper, we use this STI to discriminate movements. Here, we will explain STI and how to calculate it.

If a movement on a plane is projected by an affine camera, the point $M(X, Y, Z)$ which is on the moving object is projected as the point $m(x, y)$ on an image plane. There exists affine invariants between the spatiotemporal locus (X, Y, Z, t) of point M and the spatiotemporal locus (x, y, t) of point m. Using the locus (x, y, t), STI can be calculated by Eq. (1).

$$STI(m_i, m_j, m_k, m_l, m_n) = \frac{|(\tilde{m_i}, \tilde{m_j}, \tilde{m_k}, \tilde{m_l})|}{|(\tilde{m_i}, \tilde{m_j}, \tilde{m_k}, \tilde{m_n})|} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (1)$$

,where $\tilde{m_t}$ is the homogeneous coordinates $(x, y, t, 1)$ of spatiotemporal image point (x, y, t). $|(\tilde{m_i}, \tilde{m_j}, \tilde{m_k}, \tilde{m_l})|$ means the volume of a parallelepiped constructed by the four points $\tilde{m_i}, \tilde{m_j}, \tilde{m_k}, \tilde{m_l}$, and the order of $i \sim n$ is arbitrary. STI doesn't depend on camera inner and outer parameters. The proof is shown in Appendix 1.

There is difference between the same movements, because of movement fluctuations, measurement noise and selection of slightly different feature points. For improving the precision, we adopt a similarity measure for identifying silimar movements. The measure is the correlation coefficient between STI sequences (STI-Seq.). STI sequence consists of STI's calculated for short time interval. If the correlation coefficient is higher than a threshold, two movements are caused by the same object movement; otherwise they are caused by different ones.

4. Audio-Visual Event Correspondence

4.1 Outline of Processing This paper assumes that a scene is measured by one camera and one microphone. The processing is divided into three parts, audio information processing, visual information processing, and correspondence determination as shown in Fig. 1. The details of each part are described below.

4.2 Audio Information Processing

4.2.1 Onset Detection Onset (the start part of sound) has two characteristics. One is that onset is not influenced by echo. Another is that only onset contains a single sound, when multiple sounds are not emitted at the same time. Multiple sounds rarely begin at the same time, so that it is possible to separate mixed sound signals at the onsets. Huang et al. [7] proposed the multivalue thresholding for onset detection. The period of silence is from t_1 to t_2, and the onset of the sound signal is from t_2 to t_3. Parameters a and b are amplitudes of the sound signal (see Fig. 2).

An onset is detected if the following conditions are satisfied.

1) The increase rate b/a of the amplitude between t_2 and t_3 exceeds one.

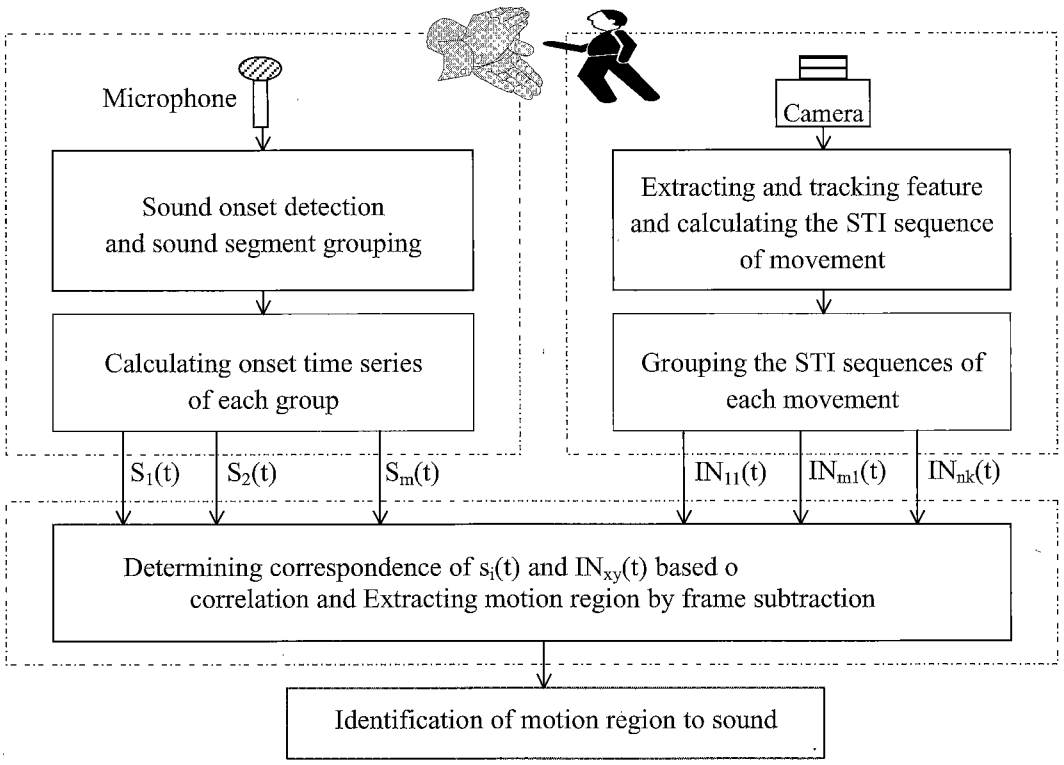


Fig. 1. Overview of signal processing.

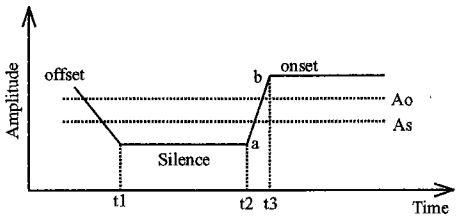


Fig. 2. Onset detection.

2) $t_2 - t_1 > T_s$ (T_s should be the minimum time, which is long enough to ignore former sound influence.)

3) $a < A_s$ (A_s is the maximum amplitude of silence, and exceeds the average level of the background noise.)

4) $b > A_o$ (A_o is the threshold value for detecting onset. It exceeds A_s and the maximum level of the background noise, and is smaller than the level of the signal sound.)

4.2.2 **Separation of Sound Sources** Because it is very unlikely that different sources emit sound simultaneously, one onset corresponds to one sound source. We transform audio signals at each onset into the Time-Frequency space (TF space) by a short-term Fast Fourier Transform (FFT). If frequency components in different onsets are similar to each other, these onsets are related to the same sound source. Whether sounds in different onsets are from the same sound source or not is evaluated by the similarity of frequency components. We adopted the correlation coefficient $Cor(E_i, E_j)$ in Eq. (2) as a measure of the similarity.

$$Cor(E_i, E_j) = \frac{Cov(E_i, E_j)}{\sqrt{Var(E_i)Var(E_j)}}, \quad \cdots\cdots\cdots (2)$$

where E_i and E_j are power spectra at onsets i and j, respectively, and E_i takes a value of power at each frequency. $Cor(E_i, E_j)$ is the correlation coefficient between power spectra E_i and E_j. If the value of $Cor(E_i, E_j)$ is near to 1, there is correlation between E_i and E_j, and the possibility that onset i and onset j are the same sound source is high. If the absolute value of $Cor(E_i, E_j)$ is near to 0, there is no correlation between E_i and E_j, and the possibility that onset i and onset j are the same sound source is low. Based on the values of the correlation coefficient, onsets are grouped by the max-min method [8].

4.2.3 **Detection of Sound Onset Series** For each separated sound source, we obtain time series $S_j(t)$ ($j = 1, 2, \cdots, m$, where m is the number of sound sources) of sound onsets. $S_j(t) = 1$ if an onset exists at time t; otherwise $S_j(t) = 0$. Thus $S_j(t)$ represents the onset time of the jth sound source and is called sound onset series in this paper.

4.3 **Visual Information Processing** In this paper, to simplify the problem, extraction and tracking of feature points are made manually. The same movement accompanies the same sound. STIs of movements are calculated using Eq. (1) for short time interval after each sound onset. Then, all obtained STI sequences at onset times are classified using the correlation coefficient as shown in Eq. (3).

$$Cor(IS_{mh}, IS_{mg}) = \frac{Cov(IS_{mh}, IS_{mg})}{\sqrt{Var(IS_{mh})Var(IS_{mg})}}, \quad \cdots\cdots\cdots (3)$$

where IS_{mh} and IS_{mg} are the STI sequences of object

m at different onset times h and g, $Cor(IS_{mh}, IS_{mg})$ is the correlation coefficient between IS_{mh} and IS_{mg}, respectively. Based on the value of correlation coefficient, the STI sequences are separated using the same method as that used in the separation of sound source. Then we obtain STI time series $IN_{mi}(t)(m = 1, 2, \cdots, n$, where n is the number of objects; $i = 1, 2, \cdots, k$, where k is the number of STI sequence patterns of object m). $IN_{mi}(t) = 1$, if object m's i-th STI sequence pattern exists at onset time t; otherwise $IN_{mi}(t) = 0$. Thus $IN_{mi}(t)$ represents the time series of the i-th STI sequence pattern of object m and is called as STI time series.

4.4 Determining the Correspondence Between Movement and Sound

4.4.1 Determining the Correspondence Between Two Series of STI and Sound Onset In most cases, the repetition of a sound correlates with the repetition of the corresponding movement. We therefore calculate the correlation coefficient of the two time series S_j and IN_{mi} by using Eq. (4).

$$Cor(S_j, IN_{mi}) = \frac{Cov(S_j, IN_{mi})}{\sqrt{Var(S_j)Var(IN_{mi})}}, \quad \cdots\cdots (4)$$

If the value of $Cor(S_j, IN_{mi})$ is the highest, IN_{mi} corresponds to S_j, audio-visual correspondences are thus determined; Otherwise, IN_{mi} and S_j are not related to the same object.

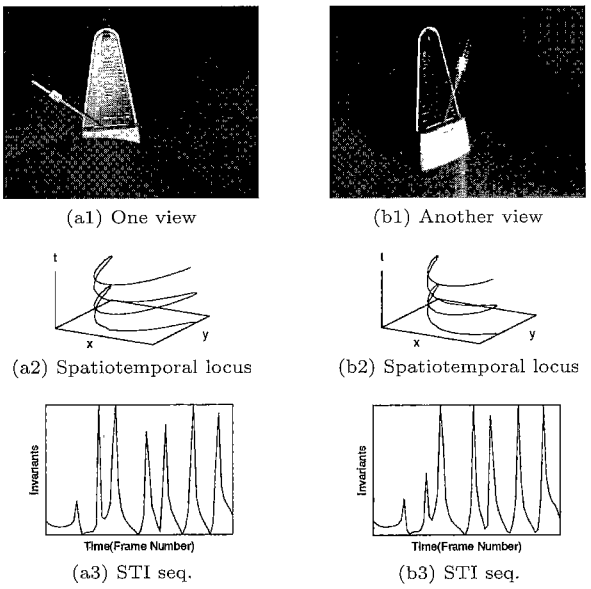
4.4.2 Motion Region Extraction Firstly, we find out motion region candidates by image subtraction between adjacent frames around an onset. Secondly, we cluster the movement regions by adjacency. Finally, we select the motion region which is the nearest to the feature point.

5. Experiments and Discussion

5.1 Experiment on the Spatiotemporal Invariant of Movement The movement of a metronome is observed from two viewpoints (see Fig. 3(a1) and Fig. 3(b1)). Fig. 3(a2) and Fig. 3(b2) show the spatiotemporal locus of one feature point of the object, respectively. Fig. 3(a3) and Fig. 3(b3) show the STI sequences calculated by Eq. (1). The correlation coefficient of these two STI sequences is 0.99. It means that there exists the invariance in a movement, even if the viewpoints are different.

5.2 The Experiments on Audio-Visual Information Correspondence

5.2.1 The Correspondence of one Movement and one Sound We observed a scene for 15 seconds by a cameara and a microphone. We set the sampling rate at 30 frames/second for image, 44.1 kHz for sound. In the middle time of measuring, the metronome was moved manually from a location (Fig. 4(a)) to another location (Fig. 4(b)). Object moving is considered as noise. And the feature points were extracted manually, through the experiments the same feature points are used. Onsets (Fig. 4(d)) are extracted from sound source shown in Fig. 4(c). All onsets are classified into

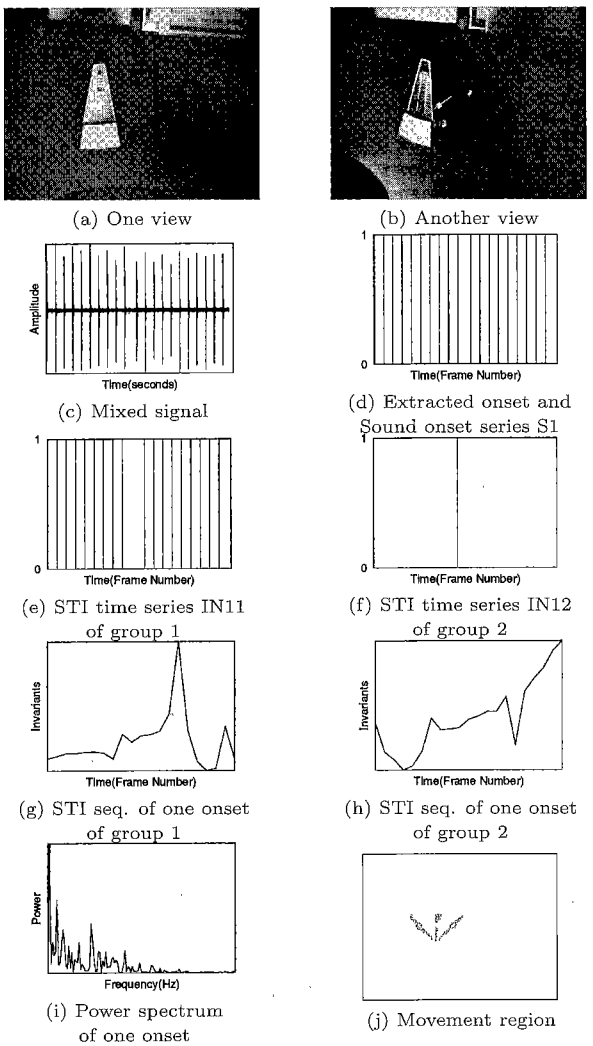


(a1) One view (b1) Another view

(a2) Spatiotemporal locus (b2) Spatiotemporal locus

(a3) STI seq. (b3) STI seq.

Fig. 3. The STI sequences from two different viewpoints.



(a) One view (b) Another view

(c) Mixed signal

(d) Extracted onset and Sound onset series S1

(e) STI time series IN11 of group 1

(f) STI time series IN12 of group 2

(g) STI seq. of one onset of group 1

(h) STI seq. of one onset of group 2

(i) Power spectrum of one onset

(j) Movement region

Fig. 4. The correspondence of one movement and one sound.

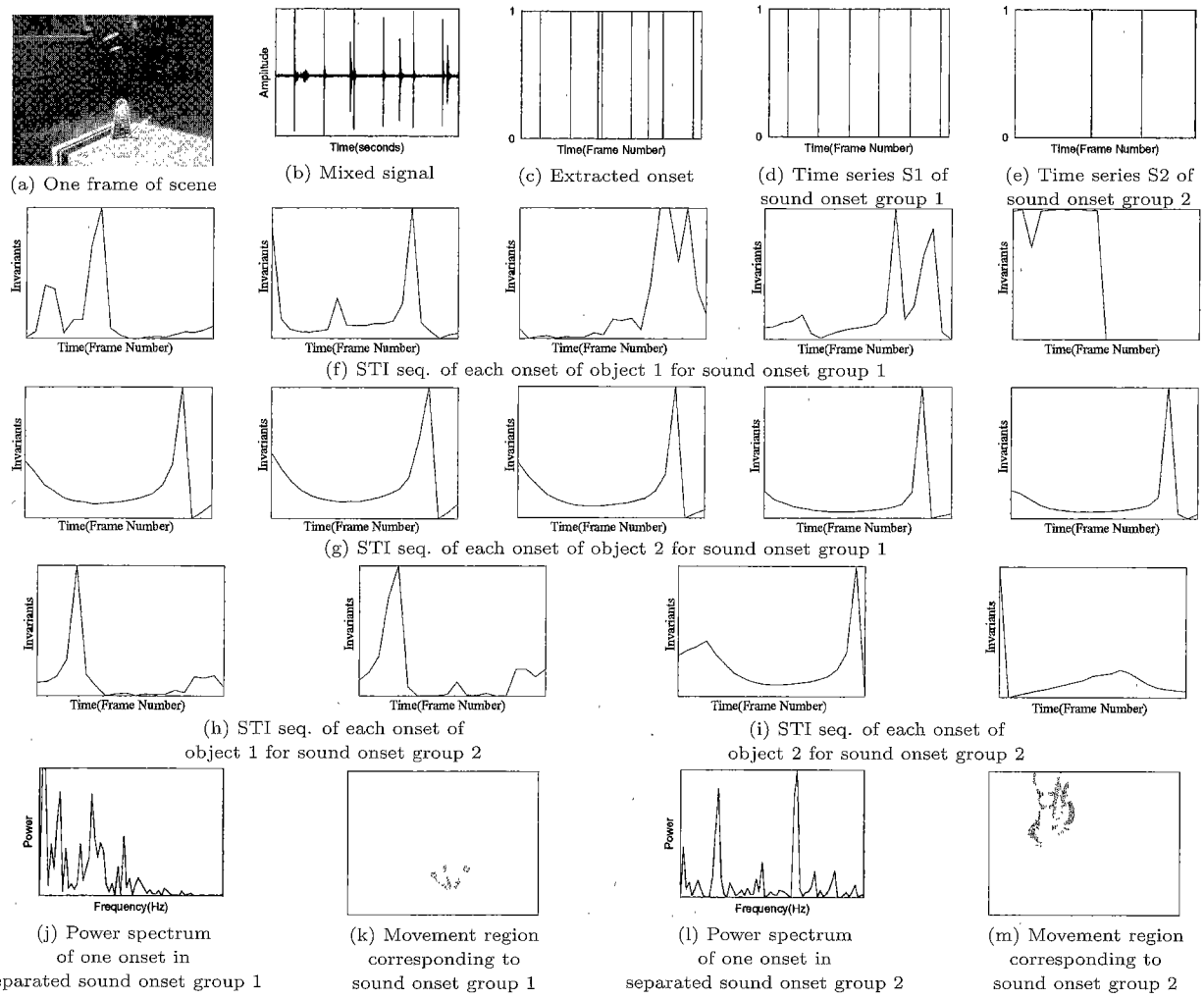


Fig. 5. The correspondence of two movements and two sounds.

the same group based on their spectral similarity. Then there is only one sound onset series S_1 shown in Fig. 4(d). Fig. 4(i) shows spectra at an onset of this sound group. The STI sequence is calculated at each onset, and these STI sequences are classified into two groups. Fig. 4(e) and 4(f) show the STI time series IN_{11} and IN_{12} of two groups, Fig. 4(g) and 4(h) show the STI sequences at one onset in each group. The STI sequence shown in Fig. 4(h) is caused by the combined moving of the metronome and the manual metornome location change. Based on the similarity of repetition between S_1 and IN_{11}, the STI sequence (Fig. 4(g)) is related to the sound (Fig. 4(i)). Fig. 4(j) shows the movement region that corresponds to it. This experiment shows that one movement can correspond to the sound correctly even if there exists noise (a movement manually) in the scene.

5.2.2 **The Correspondence of Two Movements and Two Sounds** As shown in Fig.5 (a), there are two moving sound sources in this experiment, a metronome and a footstep of one person. We observed the scene for 6 seconds. We set the sampling rate of image and sound at the same value as the above experiment. Feature points were extracted manually, too. Fig.5 (c) shows onsets extracted from the mixed signal (Fig. 5(b)). The onsets are classified into two groups based on their spectral similarity. Fig. 5 (d) and (e) show the time series S_1 and S_2 of sound onset groups.

Fig. 5 (f) shows the STI sequences of object 1 at each onset for sound onset group 1. In Fig. 5 (d), there exist 6 onsets. We excluded the last one, because its corresponding STI sequence is too short to calculate the similarity. The others are separated into three classes of STI sequence based on the similarity (The correlation coefficient between the first onset and 5th onset is 0.50, between the third and 4th 0.55, between the second and the others near to 0). Then there are three STI time series $IN_{11}, IN_{12}, IN_{13}$ of object 1 for sound onset group 1. Fig. 5 (g) shows the STI sequences of object 2 at each onset for sound onset group 1. The correlation coefficient among these onsets is over 0.98. It means that these STI sequences are similar to each other, and can be considered as the same movement. Then there is only one STI time series IN_{21} of object 2 for sound onset group 1. Because the two time series of S_1 and IN_{21} are similar, they are corresponded to one event. Fig. 5(j) shows the power spectrum of one onset of sound onset group 1. Fig. 5(k) shows the corresponding movement region of object 2.

Fig. 5 (h) shows the STI sequences of object 1 at each onset for sound onset group 2, Fig. 5 (i) the STI se-

quences of object 2 at each onset for sound onset group 2. Because the STI sequences among onsets are similar to each other in Fig. (h), STI sequences of object 1 are judged as the same class of STI sequence, so that sound group 2 corresponds to this object. Fig. 5 (l) shows the power spectrum at an onset of sound onset group 2, Fig. 5 (m) shows the extracted movement region of object 1. This experiment shows that moving objects are respectively identified with each sound even when there are two sound sources and one of objects does entire movement.

5.3 **STI Sequence Similarity of Different Feature Points on the Same Object** Here we compare the STI sequences $STI_M(t)$ and $STI_N(t)$ of the different feature points M and N, which are on the same object in three-dimensional space. $STI_M(t)$ and $STI_N(t)$ can be calculated by Eq. (1). We assume that the object moves according to matrix R,

$$R = \begin{bmatrix} \cos\theta & -\sin\theta & T_{x_n} \\ \sin\theta & \cos\theta & T_{y_n} \\ 0 & 0 & 1 \end{bmatrix}$$

where θ is rotation speed and T_x, T_y are translation speed in x and y axes. Then we obtained the following results.

(1) In the case that the rotation and translation speed are constant, the values of STI_M and STI_N are also constant value, and the value is only related to rotation speed.

$$STI_M(t) = STI_N(t) = constant \quad \cdots\cdots (5)$$

(2) In the case that only rotation speed changes with time, i.e. translation speed is zero, STI_N equals to the quantity STI_M.

$$STI_M(t) = STI_N(t) = f(t) \quad \cdots\cdots\cdots (6)$$

(3) In the case that only translation speed changes with time, i.e. rotation speed is zero, STI_N equals to the quantity STI_M.

$$STI_M(t) = STI_N(t) = f(t) \quad \cdots\cdots\cdots (7)$$

See the appendixes 2 and 3 for the proofs of Eqs. (5) to (7).

(4) In the case that both translation and rotation speed change with time, STI_N doesn't equal to STI_M. We conducted a simulation, in which the moving object is settled 4 m away from a camera. In the case that the distance of two feature points is 0.13 m, rotation speed 60 degrees/second and translation speed 2.5 km/hour, the correlation coefficient between two STI sequences is 0.4903. In the case that the distance of two feature points is 0.048 m, rotation speed 60 degrees/second and translation speed 1.2 km/hour, the correlation coefficient between two STI sequences is 0.7116. If the distance of two feature points is 0.048 m, rotation speed 30 degrees/second and translation speed 1.2 km/hour, the correlation coefficient between two STI sequences is 0.95.

We also compared the similarity between two different feature points on the same motion object in the real environment as well. The motion of a metronome is shown in Fig. 3 (a), the metronome is 3 m away from a camera. The distance of two feature points is 30 mm, rotation speed 180 degrees/second and without translation speed. We obtained high correlation coefficient of more than 0.95. It means that STI sequences of two different points are the same one when only rotation movement exits. In the motion of a foot shown in Fig. 4 (a), it is 4 m away from a camera. When the two point distance is 40 mm, rotation speed 30 degrees/second and translation speed 4 km/hour, the correlation coefficient is more than 0.75. It means that we can still use the STI sequence to discriminate a motion although it is not correct in principle, but in some range of rotation speed, translation speed, and the distance of points, STI sequence still can work well.

5.4 **Discussions** We have obtained satisfactory results in corresponding audio-visual information by using the onset of sound and the invariants of movement. Because the extraction and tracking of feature points were executed manually, we can use the same feature points through the sequence. We can't assure that the same feature point is obtained for each event when they are extracted and tracked automatically. Sometimes the correspondence of feature points doesn't work well in image series. In the case of only rotation or only translation movement, we can obtain the same space-invariants in movement even if we use different feature points for two events. But the case of only rotation or only translation or two constant speeds is very rare in the real world. So we introduced a similarity model to discriminate motion by using STI sequence. Because the correlation coefficient between two different points on the same motion object is high in simulation and in real experiments, we can say STI sequence is useful for motion recognition in some case (cosidered at viewpoints of rotation and translation speed, the displacement between feature points).

6. Conclusion

We have proposed the method for finding correspondence between visual information (including to a entire movement) and audio information based on general laws without object-specific knowledge. The effectiveness was shown by experiments. The knowledge that what kind of sound is emitted from what kind of movement and vice versa can be acquired. In other words, the knowledge and the understanding of the environment could be acquired actively by observation without teaching. Future subjects are to extract and track feature point of movement object, and to construct a similar invariance measure using different feature points in 3D movement.

Acknowledgment

This research was supported in part by a Grant-in-Aid for COE Research, and Science Research (A) (#11308012) of the Ministry of Education, Culture, Sports, Science and Technology Japan.

(Manuscript received March 17, 2003,
revised July 25, 2003)

References

(1) P.K.Kuhl and A.N.Meltozoff: "The Bimodal perception of Speech in Infancy", *Science*, Vol.218, pp.1138–1140(1982)

(2) K. Suyama, K. Takahashi and H. Iwakura: "Multiple Sound Source Location Using Two-Stage Data Selection", *IEICE A*, Vol.J79-A, No.6, pp.1127–1137 (1996-6)(in Japanese)

(3) T. Mukai and N. Ohnishi: "Grouping Corresponding Parts in Vision and Audition Using Perceptual Grouping among Different Sensations", *IEEE/RSJ Integration for Intelligent System*, pp.713–718 (1996)

(4) K. Hayakawa, R. Suzuki, T. Mukai, and N. Ohnishi: "Finding Correspondence Between Vision and Audition Based On Physical Law", *Technical Report of IEICE*, EID98–147, IE98-138, pp.13–18 (1999)(in Japanese)

(5) J. Chen, T. Mukai, Y. Takeuchi, T. Matsumoto, H. Kudo, T. Yamamura, and N. Ohnishi: "Finding the Correspondence of Audio-Visual Events Caused by Multiple Movements", *The Journal of The Institute of Image Information and Television Engineers*, Vol.55, No.11, pp.1450–1459(2001)

(6) J. Sato: "Space-Time Invariants and Recognition of Motions from Arbitrary Viewpoints", *IEICE D-II*, Vol.J84-D-II, No.8, pp.1790–1799(2001)(in Japanese)

(7) J. Huang, N.Ohnishi, and N.Sugie: "A System for Multiple Sound Source Localization", The 5th International Symposium on Robotics in Construction(1988)

(8) J. Toriwaki: "Pattern Recognition Engineering", *CORONA PUBLISHING CO., LTD*, Japan, pp.88–89 (1993)(in Japanese)

Appendix

1. Proof of STI Not Depending on Camera Parameters

STI can be calculated by Eq. (A1).

$$STI(m_i,m_j,m_k,m_l,m_n) = \frac{|(\tilde{m_i},\tilde{m_j},\tilde{m_k},\tilde{m_l})|}{|(\tilde{m_i},\tilde{m_j},\tilde{m_k},\tilde{m_n})|} = \frac{Num}{Den} \quad \cdots\cdots\cdots \text{(A1)}$$

In the case of projection with an affine camera P,

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

the points $M(X,Y,Z)$ and $m(x,y)$ can be related as

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = P \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{(A2)}$$

Let time i,j,k,l,n are $0,1,2,3,4$, respectively. Then we have $\tilde{m_0},\tilde{m_1},\tilde{m_2},\tilde{m_3},\tilde{m_4}$ as follows.

$$\tilde{m_0} = \begin{pmatrix} p_{11}X_0 + p_{12}Y_0 + p_{13}Z_0 + p_{14} \\ p_{21}X_0 + p_{22}Y_0 + p_{23}Z_0 + p_{24} \\ 0 \\ 1 \end{pmatrix}$$

$$\tilde{m_1} = \begin{pmatrix} p_{11}X_1 + p_{12}Y_1 + p_{13}Z_1 + p_{14} \\ p_{21}X_1 + p_{22}Y_1 + p_{23}Z_1 + p_{24} \\ 1 \\ 1 \end{pmatrix}$$

$$\tilde{m_2} = \begin{pmatrix} p_{11}X_2 + p_{12}Y_2 + p_{13}Z_2 + p_{14} \\ p_{21}X_2 + p_{22}Y_2 + p_{23}Z_2 + p_{24} \\ 2 \\ 1 \end{pmatrix}$$

$$\tilde{m_3} = \begin{pmatrix} p_{11}X_3 + p_{12}Y_3 + p_{13}Z_3 + p_{14} \\ p_{21}X_3 + p_{22}Y_3 + p_{23}Z_3 + p_{24} \\ 3 \\ 1 \end{pmatrix}$$

$$\tilde{m_4} = \begin{pmatrix} p_{11}X_4 + p_{12}Y_4 + p_{13}Z_4 + p_{14} \\ p_{21}X_4 + p_{22}Y_4 + p_{23}Z_4 + p_{24} \\ 4 \\ 1 \end{pmatrix}$$

Using P, Num and Den are represented by Eq.(A3) and Eq.(A4), respectively.

$$Num = \begin{vmatrix} p_{11}A+p_{12}B+p_{13}C & p_{21}A+p_{22}B+p_{23}C \\ p_{11}D+p_{12}E+p_{13}F & p_{21}D+p_{22}E+p_{23}F \end{vmatrix} \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{(A3)}$$

$$Den = \begin{vmatrix} p_{11}A+p_{12}B+p_{13}C & p_{21}A+p_{22}B+p_{23}C \\ p_{11}G+p_{12}H+p_{13}I & p_{21}G+p_{22}H+p_{23}I \end{vmatrix} \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{(A4)}$$

,where $A \sim I$ in Eq.(A3) and Eq.(A4) are expressed as follows.

$$\begin{aligned} A &= X_2 - 2X_1 + X_0, \\ B &= Y_2 - 2Y_1 + Y_0, \\ C &= Z_2 - 2Z_1 + Z_0, \\ D &= X_3 - 3X_1 + 2X_0, \\ E &= Y_3 - 3Y_1 + 2Y_0, \\ F &= Z_3 - 3Z_1 + 2Z_0, \\ G &= X_4 - 4X_1 + 3X_0, \\ H &= Y_4 - 4Y_1 + 3Y_0, \\ I &= Z_4 - 4Z_1 + 3Z_0, \end{aligned}$$

where X_t means X axis coordinate value of point $M(X,Y,Z)$ at time t, Y_t and Z_t are similar. Because STI doesn't depend on viewpoint, we assume that the projection plane of a camera parallels that of motion, i.e. $Z_0 = Z_1 = Z_2 = Z_3 = Z_4$. Then $C = F = I = 0$. STI becomes Eq. (A5). From Eq. (A5), we know the STI does not depend on camera inner and outer parameters.

$$\begin{aligned} STI &= \frac{\begin{vmatrix} p_{11}A + p_{12}B & p_{21}A + p_{22}B \\ p_{11}D + p_{12}E & p_{21}D + p_{22}E \end{vmatrix}}{\begin{vmatrix} p_{11}A + p_{12}B & p_{21}A + p_{22}B \\ p_{11}G + p_{12}H & p_{21}G + p_{22}H \end{vmatrix}} \\ &= \frac{\begin{vmatrix} A & B \\ D & E \end{vmatrix} \begin{vmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{vmatrix}}{\begin{vmatrix} A & B \\ G & H \end{vmatrix} \begin{vmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{vmatrix}} \end{aligned}$$

$$= \frac{\begin{vmatrix} A & B \\ D & E \end{vmatrix}}{\begin{vmatrix} A & B \\ G & H \end{vmatrix}} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{(A5)}$$

2. **Proof in the Case of Only Rotation**

There exist two points X_0 and Y_0 on the object plane in the three-dimensional space. Y_0 can be expressed with X_0 by using scale α and rotation R_0 as shown in Eq. (A6).

$$\begin{aligned} Y_0 &= X_0 + \alpha R_0 X_0, \\ &= (I + \alpha R_0) X_0. \end{aligned} \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{(A6)}$$

And

$$\begin{cases} X_n = R_n X_0 + T_n, \\ Y_n = R_n Y_0 + T_n, \end{cases} \quad \cdots\cdots\cdots\cdots\cdots \text{(A7)}$$

where R_n is rotation speed, T_n translation speed. Then,

$$\begin{aligned} Y_n &= R_n Y_0 + T_n \\ &= R_n (I + \alpha R_0) X_0 + T_n \\ &= (I + \alpha R_0) R_n X_0 + T_n \\ &= (I + \alpha R_0)(X_n - T_n) + T_n \\ &= (I + \alpha R_0) X_n - \alpha R_0 T_n \end{aligned} \quad \cdots\cdots\cdots\cdots \text{(A8)}$$

If the changes from X_n to Y_n ($n = 0 \sim 4$) are the same, we can say that the STIs of different points of one object are also the same. The requirement is $\alpha R_0 T_n = 0$ by comparing Eq. (A6) and Eq. (A8).

If $\alpha R_0 = 0$, $X_n = Y_n$, it is meaningless. Then $T_n = 0$. So we can say the STIs of different points are the same when there is only rotation movement.

3. **Proof in the Case of Only Translation**

There exist two points X_0 and Y_0 on the object plan in the three-dimensional space. Y_0 can be expressed with X_0 by using vector A shown as Eq. (A9).

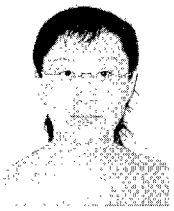
$$Y_0 = X_0 + A \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{(A9)}$$

Then,

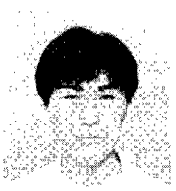
$$\begin{aligned} Y_n &= R_n Y_0 + T_n, \\ &= R_n (X_0 + A) + T_n, \\ &= R_n X_o + R_n A + T_n, \\ &= X_n - T_n + R_n A + T_n, \\ &= X_n + R_n A. \end{aligned} \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{(A10)}$$

For the same reason as the Appendix 2, by comparing Eq. (A9) and Eq. (A10), the requirement is $R_n A = A$, $R_n = I$. So we can say the STIs of different points are the same when there is only translation movement.

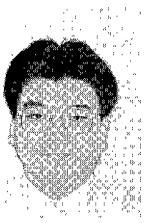
Jinji Chen (Non-member) She received the M. Eng. from Nagoya University, Japan in 2001. She is currently working for Ph. D. degree at the same place. Her research interests in sensor fusion.



Toshiharu Mukai (Non-member) He received the B. Eng., the M. Eng. and the Dr. Eng. degrees in mathematical engineering and information physics from the University of Tokyo in 1990, 1992 and 1995, respectively. He was a frontier research scientist at the Bio-Mimetic Control Research Center under the Institute of Physical and Chemical Research (RIKEN), Japan, from April 1995 to September 2000. From October 2000 to September 2001, he stayed at Laboratoire de Neurobiologie in Marseille, France, as a postdoctoral fellow. He has been the laboratory head of Biologically Integrative Sensors Laboratory of the Bio-Mimetic Control Research Center under RIKEN, from October 2001. His current research interests include sensor fusion, active sensing, shape recovery from an image sequence, and neural network. He got the 1999 Best Paper Award from Society of Instrument and Control Engineers (SCIE) and the 2000 Award for Research Achievement from the Virtual Reality Society of Japan.



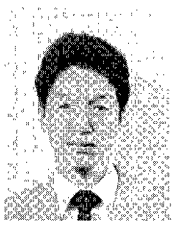
Yoshinori Takeuchi (Non-member) He received the degree of B. Eng., M. Eng., and Dr. Eng. at Nagoya University, Japan in 1994, 1996 and 1999. Since 1999, he has been a Research Fellow of the Japan Society for the Promotion of Science. Since 2000, he has been a member of the Graduate School of Engineering, Nagoya University. Since 2000, he has been a member of the Center for Information Media Studies, Nagoya University. Since 2003, he has been a member of Graduate School of Information Science, Nagoya University. His research interests include computer vision. He is a member of IEICE of Japan, RSJ and IEEE.



Tetsuya Matsumoto (Non-member) He received the B.E., M.E., and Dr. Eng. degrees from Nagoya University, Nagoya, Japan, in 1982, 1984 and 1996, respectively. From 1984 to 1989, he worked in Toshiba Corporation, Fuchu, Japan, where he engaged in research and development of the control system of nuclear power plant. In 1993, he joined Education Center for Information Processing, Nagoya University. In 1998, he moved Department of Information Engineering, Graduate School of Engineering, Nagoya University. Since 2003, he has been a member of Graduate School of Information Science, Nagoya University. His current interests include neural networks, image processing and machine learning. He is a member of IEICE of Japan, JNNS of Japan and JSAI of Japan.

Hiroaki Kudo (Member) He received the degrees of B. Eng., M. Eng. and Dr. Eng. at Nagoya University, Japan in 1991, 1993 and 1996, respectively. Since April 1996, he was a faculty member of the School of Engineering, Nagoya University as a Research Associate. Since April 1997, he was a faculty member of the Graduate School of Engineering, Nagoya University. Since April 1999, he was an Assistant Professor. Since August 2000, he has been an Associate Professor at Center for Information Media Studies. Since 2003, he has been a member of Graduate School of Information Science, Nagoya University. He was a Research Fellow of the Japan Society for the Promotion of Science in 1995.

Tsuyoshi Yamamura (Non-member) He received his B. Eng., M. Eng. and Ph. D. degrees from Nagoya University in 1987, 1989 and 1994, respectively. In 1992, he was a Research Associate in the Department of Electronic Information Engineering at Nagoya University. In 1995, he was an Assistant Professor in the Department of Information Engineering, Graduate School of Engineering, Nagoya University. Currently he is an Associated Professor in the Faculty of Information Science and Technology, Aichi Prefectural University. He is engaged in research on Natural Language Professing and Visual Information Processing. He is a member of IEEE, IEICE and IPS.



Noboru Ohnishi (Member) Noboru Ohnishi received the B. Eng., M. Eng. and D. Eng. degrees from Nagoya University, Nagoya, Japan, in 1973, 1975 and 1984, respectively. From 1975 to 1986 he was with the Rehabilitation Engineering Center under the Ministry of Labor. From 1986 to 1989 he was an Assistant Professor in the Department of Electrical Engineering, Nagoya University. From 1989 to 1994, he was an Associate Professor. Since 1994, he is a professor in Nagoya University. From 1993 to 2001, he concurrently holded a Head of Laboratory for Bio-mimetic Sensory System at the Bio-mimetic Control Research Center of RIKEN. He is now in the Graduate School of Information Science. His research interests include computer-vision and -audition, robotics, bio-cybernetics, and rehabilitation engineering. Dr. Ohnishi is a member of IEEE, IEEJ, IEICE, IPSJ, SICE, JNNS, IIITE and RSJ.