Theory and Design of Music Image System

Mikio Sasaki

DRAFT DOCUMENT FOR PRIVATE RESEARCH WHICH DOES NOT BELONG TO ANY COMPANIES OR ANY OTHER INDIVIDUALS

This is a draft document which is under consideration, originally uploaded on June 18, 2008.
Modified on June 20, Sept.22 (01:20 JPT), Sept.28 (04:20 JPT), Oct.03 (01:10 JPT), Oct.13 (06:20 JPT), Nov.03 (02:10 JPT), Nov.03 (14:10 JPT), Dec.07 (22:55 JPT), Dec.14 (01:07 JPT), Dec.21 (10:40 JPT), 2008. Jun.12 (1:00 JPT), Jun.28 (15:00 JPT), 2009. Jun.29 (01:15 JPT), 2009.
The Latest update was done on Jun.22 (02:00 JPT), 2010.

***Abstract*—In this document, I will introduce my idea and related system concept which are completely oriented for my private interests. When a user is listening to his/her favorite music, my proposed system concept will enable the machine to create the most appropriate images along a user's imagination dynamically which is learned and estimated automatically by the machine.**

***Index Terms*—sensibility, learning, description, mind map, brain player, image synthesis, concept acquisition**

I. Introduction

IN the last two decades, drastic changes have been widely propagated through the multimedia technologies in cyberspace all over the world. The author has a feeling that these movements are the symptoms that will no doubt bring new philosophic waves for every people. One of the practical possibilities is that we can handle all kinds of information in a single personal computer.

From the viewpoints of leading-edge information technology, the followings will be the key elements.
- Modeling of human sensibility
- Time series model
- Statistical inference
- Learning mechanism
- Stochastic tracking
- Knowledge database
- Confidence
- Profile
- Preference
- Particle filtering
- State tracking
- Knowledge database
- Image descriptions
- MPEG related media technologies
- Image indexing and retrieval
- Image synthesis
- Image segmentation
- Automatic concept formalization
- Estimation and prediction of unknown situation

In the aspects of new ideas, followings are the keywords:
- Mind map
- Brain map
- Brain player

This paper is still a draft version and the research is on going. All of the original properties written here do not belong to any specific companies or any other individuals. This document is being written under completely private research activities of the author.

M. Sasaki is with a private research institute in his home in Japan. Although he is working for an automotive parts company, it has nothing related to this research. If you are interested in this research send a e-mail to: (e-mail: ---------).

II. Related Work

From the view points of system, some related trials could be found although those are rather music-oriented systems.

To be filled later

III. System Concept

In 1986, firstly I proposed the basic concept of music image system which is expected to create images according to the listener's emotion when listening to the music [1]. However it is not aimed at creating any kinds of images. Basically it should let the people become comfortable. On the other hand, more than forty years ago, since the computers began to be well known, there were already some trials for generating images according to the music. But the trials were mainly devoted to control some elementary graphics in color, magnitude, and shape simply by volume, tone, and rhythm. After the decade, various arts which try to synchronize music and CG, and I sometimes come across the news that several splendid arts won the prizes. However those are not the ones I aimed at.

A. Mode

Fig.1 shows one of the simplest architecture.

1) Listener Mode

This mode defines basic behaviors of this system.

2) Game Mode

3) Player Mode

This mode means it works as a musical instrument.

B. Rulebase

Rule base to optimize comfort is to be discussed here.

1) Comfort vector

Comfort vector is rather similar to demand vector I have ever proposed. However, demand vector expresses an active demand. On the other hand, the demand which I discuss in this article such as desire to see images induced by listening to the music, is considered to be a passive demand. In this case, user demand is fairly controlled and guided by the machine.

When the problem is regarding estimation of personal emotional status and making it comfortable, it is not always effective to adopt complicated numerical calculation. In some cases it would be sufficient to use just a point scoring techniques as you often see in video games. I guess the comfort acquired through listening to the music would not be so simple numerical structure such as supervise of every factor's evaluated value for each music part. Moreover, it is considered that temporarily changing factor will dominate fairly a large part. I'm afraid it is not too much to say that the machine I am to design would not be able to create arts like the most sensitive brand-new artist.

Although it might be an eternal goal …

In order to increase the comfort degree, the following strategies are considered:

a) Macro strategy

(M1) stress control
regarding stress and laxity

(M2) sometimes unexpected
There are various levels such as image presentation, play music, composition, etc.

(M3) simultaneous occurrence of plural peaks

(M4) successive occurrence of a series of peaks

2) Saturation vector

The increase of comfort degree is not always unconstrained. Sometimes the upper boundaries for the accumulation of comfort degree exist. I call this notion as saturation vector which idea is simply shown in Fig.6. And this saturation vector ***s*** means the statues of the user. Therefore the demand vector ***d*** is considered to take the values as below:

$$\boldsymbol{d}_k = \boldsymbol{b} - \boldsymbol{d}_k$$

Where k means the index of discrete time point T_k. Boundary vector **b** is for the moment set as constant vector although the values depend on individual user characteristics.

To be continued.

3) Interaction Vector

To be filled.

4) Scene Change Strategy

Generally, there are no doubt catastrophic changes in peaks of a

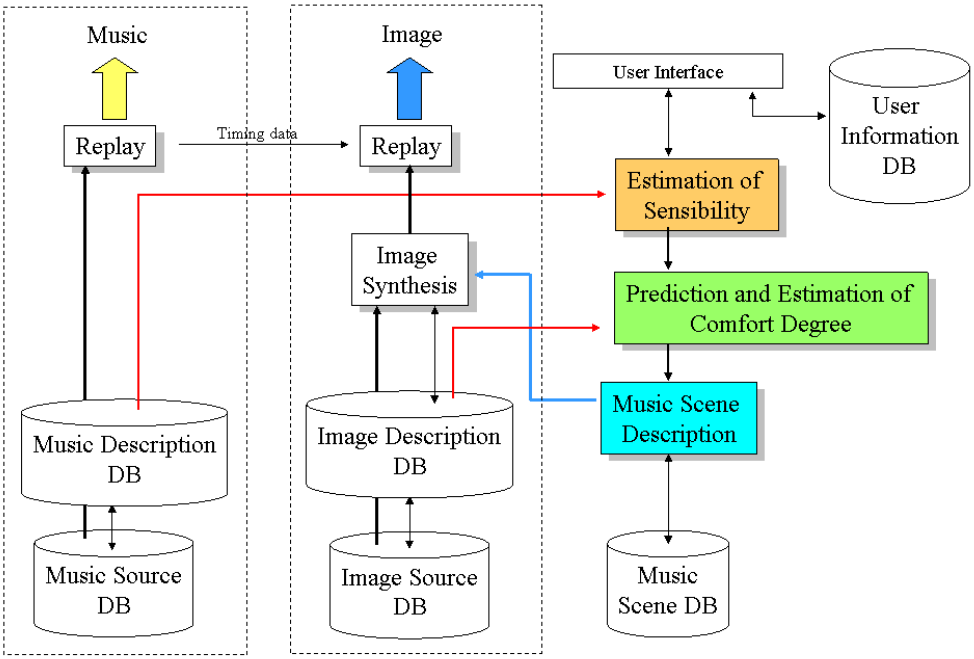


Fig. 1 System architecture.

song. It is natural to perform scene change at the moment. However the following factors are indefinite and should be discussed later in detail.

- Speed of change
- Video effects of change
- Contents before and after of scene change
- Time difference between scene change point and sound change point

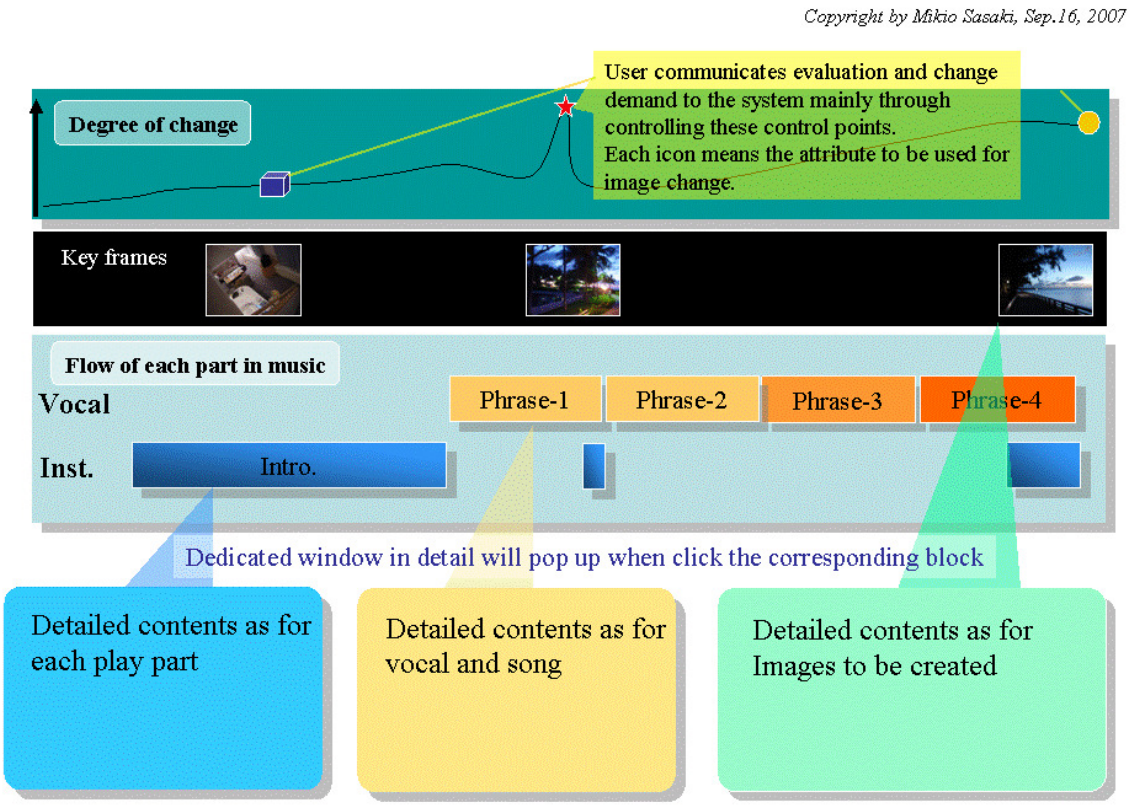


Fig. 2 An example of user interface for evaluation.

IV. Media Description

A. Image Features

In order to acquire the knowledge from images, there would be at least two ways for a human. The first one is that stereo vision by using the human's own eyes. In this case, not so huge scene database for MIS (Music Image System) could be constructed because one ordinary person's physical experiences are fairly restricted. However the person can acquire the basic bases to evaluate any kinds of images and to reason what kinds of objects exist in the images. This process would be performed from the aspects of 3D shape, real color, movement, etc. And often the person can acquire other modalities of information simultaneously. Those are closely combined together and will formalize concept of scene and objects that are included in the scene. Actually, the concept varies from time to time. However the notion of each object has been acquired through the person's experience and communication to and from other people. Therefore the statistical confidence matrix for each scene and its including objects is required. This process would form so-called manifold of multi-variate signals and in some cases it might be analyzed by using KPCA (Kernel Principal Component Analysis) techniques. Moreover, the non-linear mapping which would be used in the KPCA has the possibility of mapping to natural language vocabularies. This idea is partly related to PLSA (Probabilistic Latent Semantic Analysis). From the aspects of multi-variate linear regression analysis, the author has already tested the basic methodologies described above. The most important problem would be to integrate new wave of semantic analysis and conventional 3D measurement analysis. I guess the latter one has been eagerly researched in computer vision area for at least the last three decades. However the first one has become activated only recently, under rapidly growing multimedia technologies in the last ten years.

B. Towards Automatic Indexing

Now we get back to the second way to acquire knowledge from images. This is seeing 2D videos or pictures by human eyes. In this case, usually, shape information would not be acquired through the media itself. But a human brain can guess the information. In order to simulate such surprising functions of human in MIS, the time series analysis like factorization or motion stereo would be possible solutions.

Before going to the scene class, the basic recognition strategy for key objects in the scene is proposed. This idea is organized from the three major triggers such as shape, spectrum, and situation. There are many shape-based recognition methods for a target object and usually, this is the most popular approach. However, they often suffer from serious problems of blur, occlusion, dynamic change, vagueness, individual differences, etc. Secondly, spectrum based recognition is considered. This is applied to both of spatial spectrum and temporal spectrum. Although this spectrum based approach shows the robustness to the problems the shape based methods suffered from, it cannot independently cover the whole of the recognition task. Thirdly, the situation is considered from the knowledge based decision of the pattern classification results. This situation factor is fairly dependent on the human observer and application factor and would supply well hypothesis for particle filter like tracking. On the other hand, unfortunately, it would generate many probabilistic solutions if the shape and the spectrum are insufficiently analyzed.

Therefore I have proposed the dynamic activation scheme of these three triggers (hereafter, I call it 3S model). According to this scheme, the system can switch the main processing strategy at any time. And the confidence degree described later will integrate the series of subtask of recognition at dynamic switching.

On the basis of this idea, I precisely define and propose the following object based scene recognition especially from the aspect of MPEG-based image processing.

C. Confidence Degree

For each block of pixels, we extract color, texture, and motion information directly from MPEG based encoded data without reprocessing decoded images. Then, rule based reasoning, multi-variate regression based reasoning and 3D model based recognition are integrated to calculate the confidence vector that maps each block of pixels to 64 object categories. Brief reviews of the statistical reasoning part in our scene recognition method are as follows:

A confidence vector is defined as

$$\boldsymbol{C} = (C_1, C_2, \ldots, C_i, \ldots, C_N)^T \tag{1}$$

where C_i is i-th component of $\boldsymbol{C}$ and it means a confidence degree of attaching index A_i to one block of pixels in a single picture. Usually, the confidence vector is affected by various situation factors. Largely, there are two classes of situation factors. The first one is regarding situations of scene capture such as time, place, and weather, etc. The second one is regarding image features such as color, texture, motion, two dimensional position, etc.

1) Estimation of Confidence Degrees

Then we can represent the causal relations between the situation factors and the confidence vector by introducing the multi-variate linear regression model such as:

$$\boldsymbol{C} = \boldsymbol{c}_0 + W\boldsymbol{s} \tag{2}$$

Where, $\boldsymbol{c}_0$ means a N-dimensional column vector of initial confidence values. W is a set of regression coefficients ($N \times L$ matrix) corresponding to factor $\boldsymbol{s}$ which means L-dimensional column vector.

Therefore we can estimate B as W^T by utilizing the least squares method as follows:

$$B = (S^{\mathrm{T}} S)^{-1} S^{\mathrm{T}} Y \tag{3}$$

Where, S is represented as the following form:

$$S = [\boldsymbol{s}(1) \ldots \boldsymbol{s}(k) \ldots \boldsymbol{s}(K)]^T \tag{4}$$

$\boldsymbol{s}(k)$ means $\boldsymbol{s}$ at time T_k $(k=1,\ldots,K)$. And Y is a matrix that represents a history of confidence:

$$Y = [\boldsymbol{Y}_1 \ldots \boldsymbol{Y}_n \ldots \boldsymbol{Y}_N] \tag{5}$$

Where, $\boldsymbol{Y}_n$ is a K-dimensional column vector which represents a history of confidence degree corresponding to n-th vocabulary.

Then we can estimate unknown confidence vector $\boldsymbol{C}$ according to various combinations of situation factors. We consider the

index corresponding to the maximum component of $\boldsymbol{C}$ as the first ordered recognized result.

2) Decomposition of Confidence

Confidence degree for a MBK is considered to be decomposed into two factors. First one is regarding scene's situation. Second one is regarding image signals. Then the confidence vector of $MBK(m_r,m_c,k)$ $(m_r=1,..,M_r,\ m_c=1,..,M_c)$ at time T_k can be expressed as follows:

$$C(m_r,m_c,k)=W_S\,\boldsymbol{S}_S(m_r,m_c,k)+W_P\,\boldsymbol{S}_P(m_r,m_c,k) \tag{6}$$

Where, $\boldsymbol{S}_S(m_r,m_c,k)$ means the L_S dimensional column vector which represents the situation of scene projected onto $MBK(m_r,m_c,k)$. And W_S is the corresponding set of regression coefficients ($N \times L_S$ matrix). $\boldsymbol{S}_P(m_r,m_c,k)$ means the L_P dimensional column vector which represents the image signal features relating to $MBK(m_r,m_c,k)$. And W_P is the corresponding set of regression coefficients ($N \times L_P$ matrix).

3) Measurement Matrix

The measurement matrix which contains learning data set is defined as follows:

$$\Psi=[S\ Y] \tag{7}$$

Where, S is a feature factors matrix defined in (4), and Y is a confidence degree matrix. Confidence degrees for learning are set from ground truth or past statistical data. For simplicity, following definition is considered.

$$\boldsymbol{s}=(\ \boldsymbol{S}_P{}^T\ \boldsymbol{S}_S{}^T\ )^T \tag{8}$$

Where,

$$\boldsymbol{S}_p=\left(\boldsymbol{S}_{color}{}^T\,\boldsymbol{S}_{texture}{}^T\,\boldsymbol{S}_{block\_position}{}^T\,\boldsymbol{S}_{motion}{}^T\right)^T \tag{9}$$

$$\boldsymbol{S}_{color}=\left(I_Y,I_U,I_V\right)^T \tag{10}$$

$$\boldsymbol{S}_{texture}=\left(A_{hor},A_{ver},A_{all}\right)^T \tag{11}$$

$$\boldsymbol{S}_{block\_position}=\left(m_c,m_r\right)^T \tag{12}$$

$$\boldsymbol{S}_{motion}=\left(v_x,v_y\right)^T \tag{13}$$

and, $\boldsymbol{S}_{color}$ is the L_{color} dimensional column vector which represents the color information of scene projected onto $MBK(m_r,m_c,k)$. Equation (10) represents the mean color of BLK which is easily acquired from the DC components of 2D-DCT coefficients. $\boldsymbol{S}_{texture}$ is the feature vector made from AC components of 2D DCT coefficients of BLK. A_{hor} means the horizontal strength of AC components which is defined as the absolute sum of specified AC components. A_{ver} is define similarly regarding vertical AC components. A_{all} is the total absolute sum of AC components, namely, AC power. And $\boldsymbol{S}_{motion}$ represents two dimensional motion vector on $MBK(m_r,m_c,k)$.

D. Situation Vector

In this article, $\boldsymbol{S}_S$ is defined as situation vector and the detailed structure is as follows:

$$S_S(m_r,m_c,k)=\left[\boldsymbol{S}_{nenv}{}^T\ \boldsymbol{S}_{location}{}^T\ \boldsymbol{S}_{time}{}^T\right]^T \tag{14}$$

Where, $\boldsymbol{S}_{nenv}(m_r,m_c,k)$ means the L_{nenv} dimensional column vector which represents the natural environment information of scene relating to $MBK(m_r,m_c,k)$.

And, $\boldsymbol{S}_{location}(m_r,m_c,k)$ means the $L_{location}$ dimensional column vector which represents the location information of the scene regarding $MBK(m_r,m_c,k)$. $\boldsymbol{S}_{time}(m_r,m_c,k)$ is the L_{time} dimensional column vector which represents the time information of the scene.

In situation vector, there exist many feature components that cannot be applied to algebraic computation. Then each of them is expressed as the language label (index) and is supposed to have confidence degree for it according to the well-known quantification theory. Examples are as follows:

$$\boldsymbol{S}_{\text{time}}=(\ early\_morning,morning,\ldots,evening,night)^{\mathrm{T}} \tag{15}$$

$$\boldsymbol{S}_{\text{nenv}}=(\ fine,cloudy,rain,\ldots,storm)^{\mathrm{T}} \tag{16}$$

Where, the value of each index $\{early\_morning,\ldots,fine,\ldots\}$ is 1 if the corresponding phenomenon is true, otherwise 0.

E. Evaluation

Each index of MBK is determined by detection of the maximum confidence degree of component in the confidence vector. When the determined index is equal to the ground truth $G_{view}(m_r,m_c,k)$, the recognized result $R_{view}(m_r,m_c,k)$ is set as 1, otherwise 0. Where, G_{view} takes the value of 1 only when the object is defined in the object list according to the specified view, otherwise 0. Therefore the recognition rate Q_{view} is defined as follows:

$$Q_{view}=\frac{100}{K_{view}}\sum_{i=1}^{Mr}\sum_{j=1}^{Mc}R_{view}(i,j,k) \tag{17}$$

$$K_{view}=\sum_{i=1}^{Mr}\sum_{j=1}^{Mc}G_{view}(i,j,k) \tag{18}$$

As this is the measure to evaluate the recognition rate for all of the objects in the view list over a picture plane, it is called scene recognition rate in this article. On the other hand the well known F measure for a single object is defined as object recognition rate.

F. Integration of Shape and Semantics Acquisition

1) Interaction with Semantics

From the viewpoints of broadness of scene database, it is eagerly hoped for MIS to have this simulated function of human scene memories. Then it will become possible for MIS to generate the most appropriate scene from the database.

Therefore I propose the unified scheme for automatic indexing of scene from both aspects of measurements and semantics. The well-known factorization techniques on the basis of excellent works by *Kanade et al.* are briefly described by the next equation derived through SVD (Singular Value Decomposition):

$$\tilde{\boldsymbol{W}}=\boldsymbol{MX} \tag{19}$$

Where, $\tilde{\boldsymbol{W}}$ denotes measurement matrix ($2F\times P$) including mean separated two dimensional positions of P 3D sampled points of the target object which is captured from F cameras. $\boldsymbol{M}$ ($2F\times 3$) denotes camera characteristics under the assumption of orthographic projection. And $\boldsymbol{X}$ ($3\times P$) denotes 3D shape information of target object.

Secondly, semantics acquisition is expressed by using the following formula through multivariate linear regression model:

$$\boldsymbol{C}=\boldsymbol{RS} \tag{20}$$

Where, $\boldsymbol{C}$ denotes the confidence matrix ($N\times P$), and N means the number of object categories. In some applications, each category can be expressed in natural language vocabularies.

Also it can be related to principal component in PCA even if it is nonlinear (KPCA) according to the semantic mapping property of confidence vectors. $\boldsymbol{R}$ ($N\times L$) denotes the regression coefficients matrix. $\boldsymbol{S}$ ($L\times P$) denotes the observed state matrix and L means the number of observed states of each sampled point of the target object. Basically the states are acquired through the projected image information, but these are easily extended to the surrounding environments and related attributes from both of signal and semantics. Also it would be possible to extend to include the user profile in future works, even if it has time-varying properties. In order to take the same form as decomposition, the equation can be converted as below:

$$\boldsymbol{S} = \boldsymbol{R}^{-}\boldsymbol{C} = \boldsymbol{Q}\boldsymbol{C} \tag{21}$$

Where, $\boldsymbol{Q}$ ($L\times N$) denotes the generalized pseudo inverse of $\boldsymbol{R}$.

Remark that the equation (3) is independent of camera number f ($f=1,\ldots,F$).(See ***Appendix-A***)

Then two equations can be integrated into the next formula:

$$\overline{\boldsymbol{W}} \equiv \begin{bmatrix} \tilde{\boldsymbol{W}} \\ \tilde{\boldsymbol{S}} \end{bmatrix}, \quad \overline{\boldsymbol{W}} = \overline{\boldsymbol{M}}\overline{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{M} & \boldsymbol{A} \\ \boldsymbol{B} & \tilde{\boldsymbol{Q}} \end{bmatrix} \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{C} \end{bmatrix} \tag{22}$$

Where, $\overline{\boldsymbol{W}}$ ($(2+L)F\times P$) denotes the generalized measurement matrix. And for the moment, it is assumed that $\boldsymbol{A}=\boldsymbol{0}$ and $\boldsymbol{B}=\boldsymbol{0}$. $\tilde{\boldsymbol{S}}$ ($LF\times P$) and $\tilde{\boldsymbol{Q}}$ ($LF\times N$) are easily composed from F $\boldsymbol{S}$s and F $\boldsymbol{Q}$s, respectively.

According to this scheme, it would become possible for the MIS to acquire simultaneously the shape, the motion, and the semantics of the target object and scene, automatically from tremendous 2D image media database.

However, there still remain the fundamental questions regarding equation (4). First one is that the difference between SVD and multivariate regression (hereafter, I call it as MVR). SVD can derive shape and camera motion at once and those solutions are considered to be unique. On the other hand, MVR relation is basically a projection and there exists plural solutions for Q and C. Second difference is that SVD needs no learning process but MVR needs it.

If the MVR part of (4) have the similarly consolidated property in mental camera like capturing mechanism in Q as the optical camera characteristics M in SVD part of (4), it may become possible to get Q and C simultaneously by the same SVD mechanism. This needs more mathematical investigation.

Second idea is to simply consider (4) as the linear projection problem. In this case, it is possible to solve by using the least squares method even if camera characteristics M has the perspective view transformation properties, which cannot be solved by using SVD because factorization generally assumes the orthographic projection. This second idea means that (4) is considered as a statistical learning problem. However, it may cause the problems of camera number and freedom degrees of camera configuration. If the F cameras are fixed or located on fairly restricted trajectories, this statistical learning would become possible.

-- More precisely defined process and related details are under study --

G. Temporal Characteristics of Confidence Vectors

Three major laws regarding confidence vectors are specified as follows:

1) First law: Change through Communication

When simulating human perception of scene in MIS, number of confidence vector (hereafter, I call it as CV) fields in a scene will increase through communication processes. These processes are logically classified into four classes. First class is man-to-man (namely, agent-to-agent in MIS) and second class is man-to-machine (namely, teaching MIS by user). Third class is machine-to-man which is often used with second class to create interaction process. Fourth class is machine-to-machine which is enabled by communication between plural MISs. In every case, communication will change CV of target categories and it will gradually reach consensus or common sense in MIS.

2) Second law: Change through Movements

For a single observer, each CV varies according to the observer's movements.

3) Third law: Change of Area Size for One CV

The size of area for a single CV is not restricted. For example, In the case of MPEG video, it is not restricted to one block of pixels although the change is on the basis of MPEG block unit. From the aspects of semantics for the area which is to have CV, it is considered that atmosphere, feeling, environment etc. as well as object in the scene can take any considerable size for one CV.

H. Extension

From the aspects of knowledge processing, matrix A and matrix B in (4) have the special roles. $\boldsymbol{A}$ means a mapping from $\boldsymbol{C}$ to $\boldsymbol{W}$, namely, 2D positional information of sample points may be affected by the confidence of object categories to which each point is to belong. Secondly, $\boldsymbol{B}$ means a mapping from $\boldsymbol{X}$ to $\boldsymbol{S}$, namely, shape information may affect observed states.

Moreover, these $\boldsymbol{A}$ and $\boldsymbol{B}$ will have the spatio-temporal evolution process and also will have non-linear mappings to some numerical representations of semantic quantities or human sense properties including feeling and emotion. This expectation means that basic mathematical mechanism like (4) is linear system oriented in my proposed system design. And other non-linear functions of MIS will be realized mainly by using non-linear mapping, logical operations, and state transition like recurrence architecture.

In view of category of aspects, equation (4) can be extended more. Suppose that the extended aspect is U (user), two aspects already discussed above are X (3D scene structure) and C (semantics of contents). Therefore (4) is extended as follows:

$$\begin{bmatrix} W \\ S \\ P \end{bmatrix} = \begin{bmatrix} M_X & A_C & A_U \\ B_X & M_C & B_U \\ C_X & C_C & M_U \end{bmatrix} \begin{bmatrix} X \\ C \\ U \end{bmatrix} \tag{23}$$

For simplicity, it can be denoted as below.

$$W^* = M^* X^* \tag{24}$$

Where,

$$W^* \equiv \begin{bmatrix} W \\ S \\ P \end{bmatrix}, \quad M^* \equiv \begin{bmatrix} M_X & A_C & A_U \\ B_X & M_C & B_U \\ C_X & C_C & M_U \end{bmatrix}, \quad X^* \equiv \begin{bmatrix} X \\ C \\ U \end{bmatrix} \tag{25}$$

Remark that the P in the measurement matrix W^* means practically a user profile.

I. Generalization

Equation (5) can be generalized as follows:

$$\begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_N \end{bmatrix} = \begin{bmatrix} M_1 & A_{12} & \cdots & A_{1N} \\ A_{21} & M_2 & \cdots & A_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ A_{N1} & A_{N2} & \cdots & M_N \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \tag{26}$$

Where, N means the number of aspects for the problem I discuss. So, it is possible to include audio and music aspects in (8). Moreover, the aspects can be considered to include other modalities which can be sensed by touch, heat, etc. Each aspect can have hierarchical layer structure which is expressed by the same form of equation (8). For example, it is possible to express the layer model around brain map (Fig.5) by this formula. (8).

In view of mathematics, it is splendidly beneficial if (8) is decomposition of measurement matrix. However usually, it is also possible to interpret (8) as a linear regression model regarding multivariate time series.

To be continued

J. Application and Vocabulary

If the camera is moving like on-board camera, especially for the case when it is going through, the labeling problem suddenly becomes very difficult. Although a snapshot can be labeled and it will lead to the generation of the scene class (SC), the video scene incessantly changes and creates new SCs. Then even if the observed video scene will have totally one label, it will locally have many different labels.

The definition of index vocabulary is quite dependent on the targeted application. The considered purposes of scene class acquisition are currently retrieval, picture classification, location-aware service, environmental recognition, etc. For a video sequence, the time span of the scene class should be considered. The simple idea is to determine the scene class label for a scene from the composition of the time series of scene class label which is defined for each single frame in the scene as follows:

$$\begin{aligned} \alpha_k &= Index\left\{ p(\boldsymbol{X}^*_k \left| \boldsymbol{W}^*_{1:k}, \boldsymbol{X}^*_{1:k-1} \right.) \right\} \\ \alpha &= SceneClass\{\alpha_{1:K}\} \end{aligned} \tag{27}$$

Where, K is a total number of frames in a scene. *Index* is a mapping from object based confidence vectors to a scene class label at frame time k. *SceneClass* means a mapping from series of index to a single label for the total scene. The basic idea behind this formula is the particle filtering based state tracking [ ]. Because the target application drives the definition and collection of scene database, vocabularies and adjustment logic would be included in these two nonlinear mappings. Moreover, this automatic CSC generation mechanism which enables labeling over a time series features through nonlinear mappings seems quite familiar with the recent excellent work on manifold learning [ ].

Second problem is that the viewpoint effect which is the evaluation side of variation under the constraints of target application. As will be described later, directed graph expression of CSCs is affected by the viewpoint. On the basis of the analogy to natural language understanding, the digraph architecture of semantic network of words is applicable to CSC generation as one of the starting point of visual concept formalization.

K. Scene Class

1) Subjective scene class

Subjective scene class (SSC) is determined by the human defined subjective grouping of scenes through the use of geographical conditions and objects included in the image. At first, I have considered the scene class by focusing the following static factors.

(1) Configurations of major objects in the scene (building, trees, green area, overpass, snow, etc.)
(2) Road class (general road, highway, etc.)
(3) Geographical classification (mountain, town area, beach, lakeside, etc.)

According to empirical investigation on these factors, I can consider several classes in Fig.5. It is relatively simple to describe each of them in natural language expression. In practical scenes, there additionally exist dynamic objects such as vehicles and pedestrians. Moreover, there are other factors which affect the scene such as changes of weather and illuminations, differences in position and attitude of camera, aspects from vehicles, etc.

2) Computational scene class

Computational scene class (CSC) is to classify plural scenes purely according to the machine learning. Now, it is possible to define directional branch which specifies the notion that "If S_1 is learned by the system, it can recognize S_2 ". This means when a scene S_1 is learning data and $S2$ is recognized at the rate of more than the threshold α preset in advance. Simultaneously, if the similar branch can be defined from S2 to S1, the relation is expressed as "S1 and S2 belong to the same computational scene class". This definition of CSC does not mean that the S1 has a unique CSC. Namely, plural

CSCs are possibly defined on S1. On the other hand, if the corresponding subjective scene classes (SSCs) which also can exist simultaneously on a single S1 are known, the relations between SSC and CSC can be constructed automatically.

3) Automatic generation of Scene Class

For simplicity, the view of scene recognition rate is set to *all*. For a combination of learning image I_L and target image I_E, if the scene recognition rate Q_{all} satisfies the following condition:

$$Q_{all} \geq Q_{th\_CSC} \tag{28}$$

I_L and I_E belong to the same CSC, otherwise the different CSC.

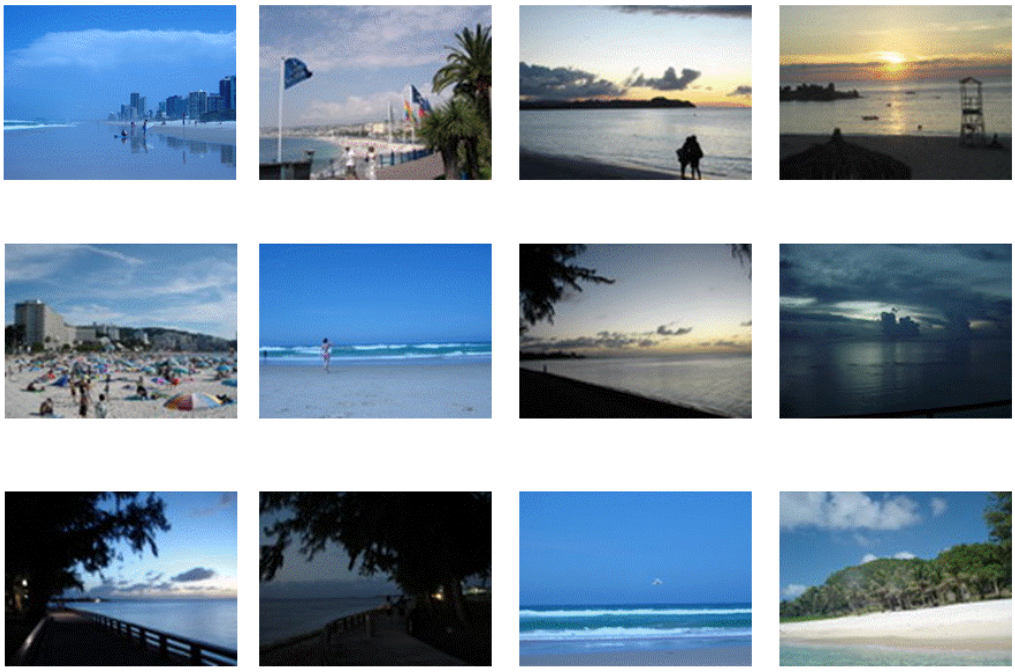


Fig. 3 Database example of ‘beach’.

To be continued

L. Image Database

As a first example, the author’s original various pictures for a vocabulary ‘beach’ is shown in Fig.2.

M. Natures of Scene to be Analyzed

To be continued

V. User Description

A. User Profile

In order to make the machine adaptable to each user's characteristics, the notion of user profile is one of the most popular and easiest ways of techniques.

Firstly the numerical distance between profiles can be defined so as to express the difference of each user's characteristics as follows:

$$\boldsymbol{D}_{prof}(A,B) = \mathrm{DIST}(\boldsymbol{Prof}_A, \boldsymbol{Prof}_B) \tag{29}$$

Where, *Prof* denotes $N_p \times L_p$ matrix each row vector of which means multivariate numerical values for one profile attribute on the basis of confidence vector even if the original attribute takes natural language expression.

To be continued

B. Situation Description

Description of situation or status would be performed as profile. Profile works sometimes as memory of the state that describe the past. And sometimes the profile also works as preset stereotype knowledge for inference. The recurrent use of profile can evolve the profile itself. Therefore the profile will be temporarily modified.

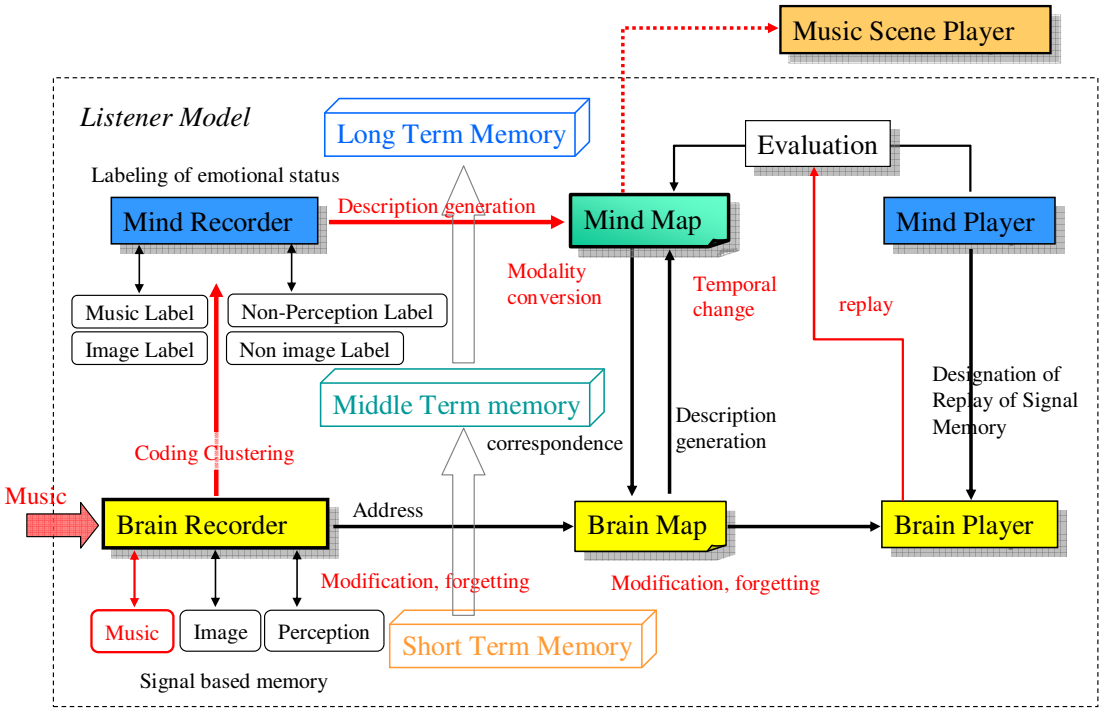


Fig. 4 Creating process from brain recorder to music scene.

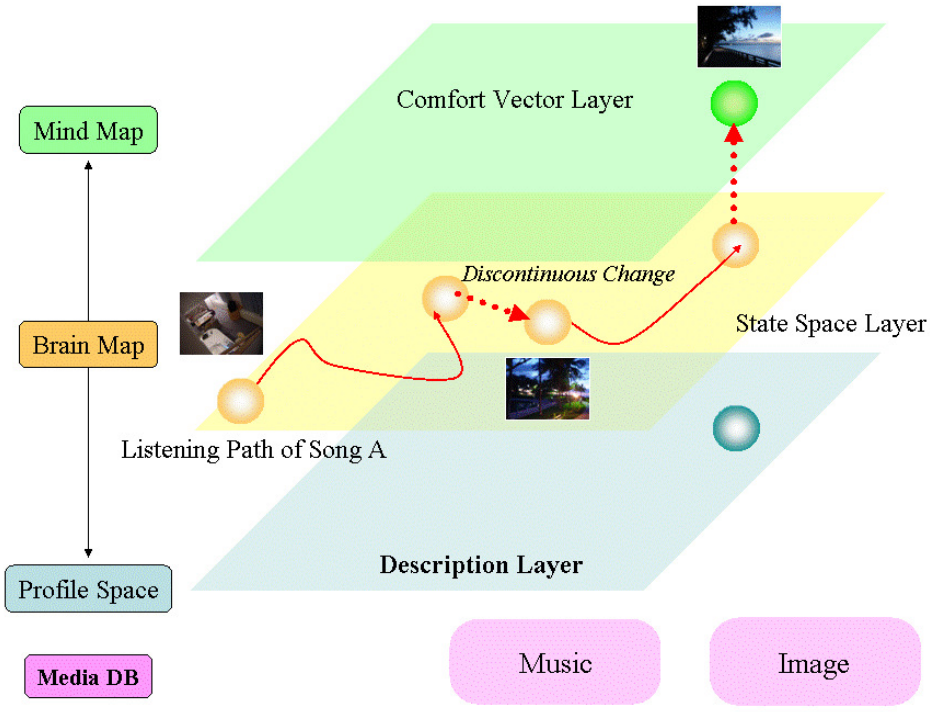


Fig. 5 Layer model around brain map.

VI. PERCEPTION MODEL

A. *Time and space moving model*

To be filled.

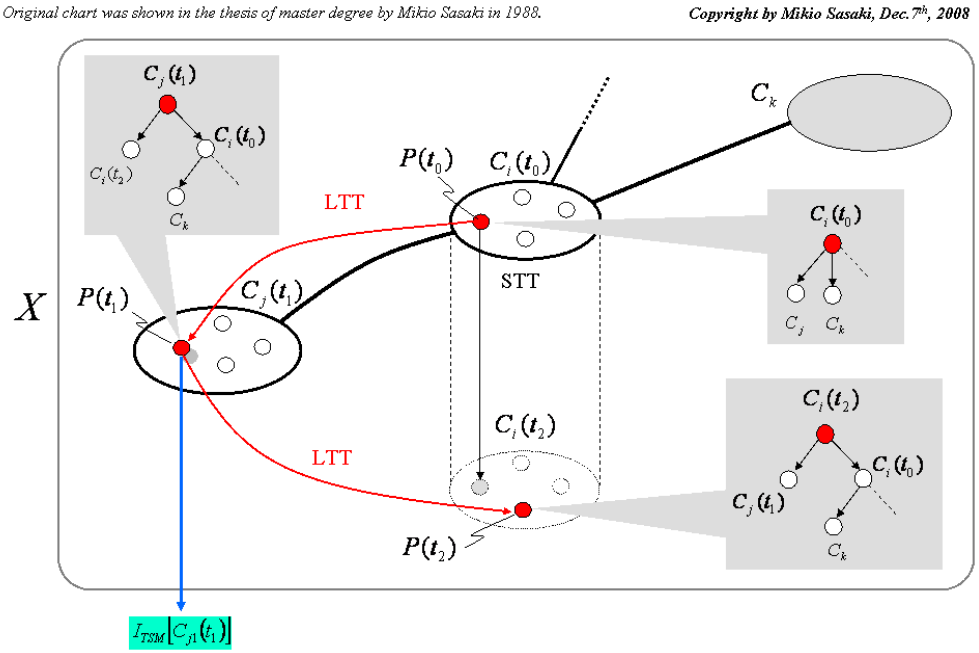


Fig. 6 'Moving' in TSM model and its corresponding hierarchical transition.

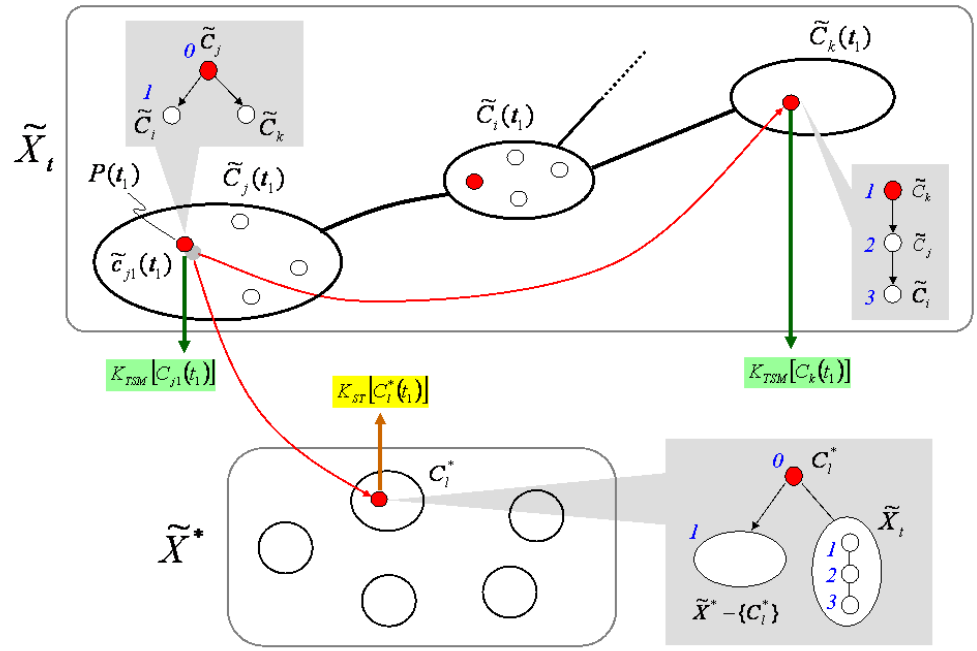


Fig. 7 'Moving' in virtual world expressed in ST model and its corresponding hierarchical transition.

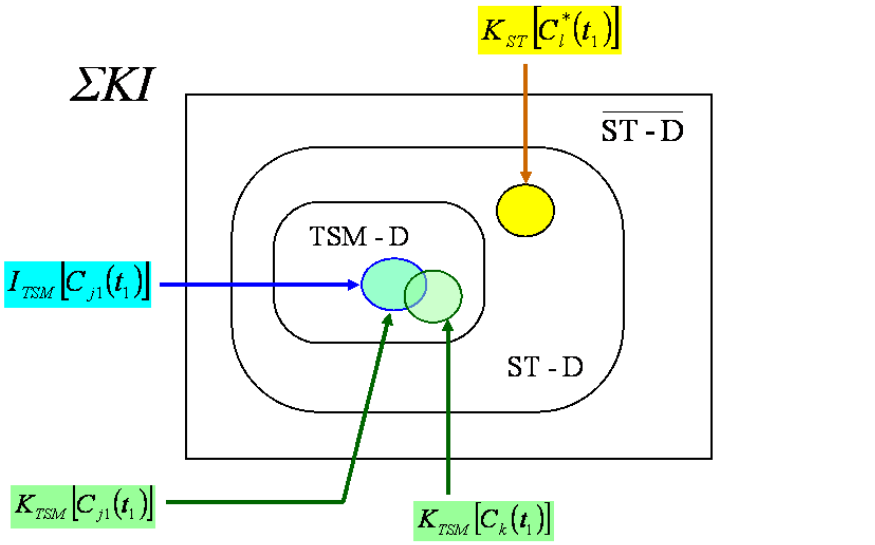


Fig. 8 Inclusion relation of KI according to TSM-ST dependencies.

B. Verification

Verification between real perception or experience and past memory should be iterated in order to keep common perception as compared with other individuals.

To be continued

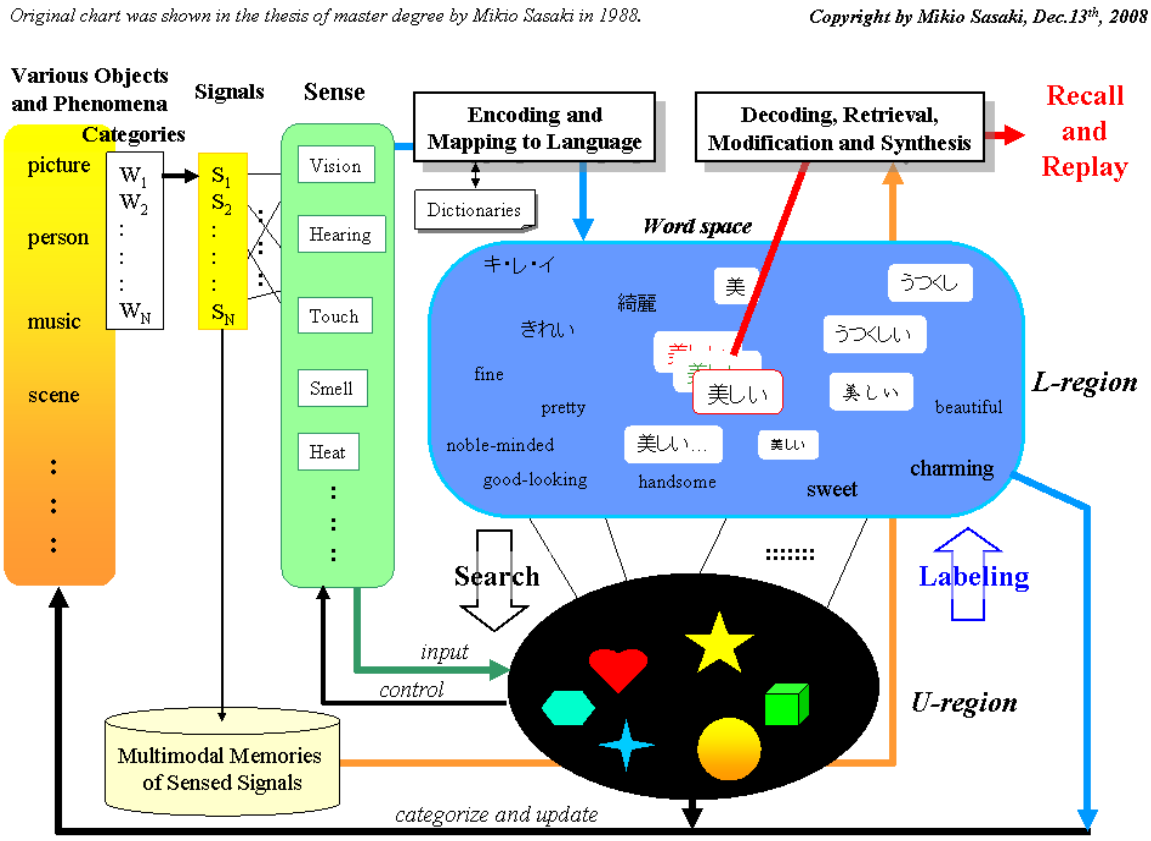
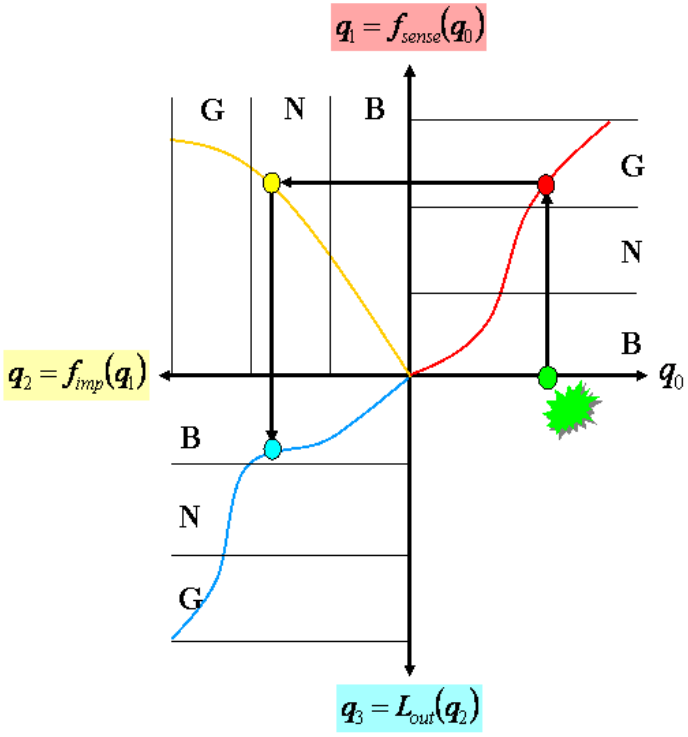
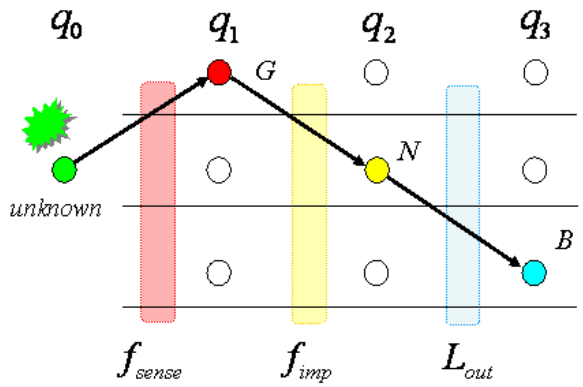


Fig. 9 Labeling process of the sensitivity status relating to a word '美しい'.



(a) Detailed mechanism of linguistic labeling.



(b) Transition in the quantized state space.

q_0	q_1	q_2	q_3
state of dish	taste	evaluation	utterance
unknown	Good	Good	Good
		Normal	Normal
	Normal	Bad	Bad
	Bad		

(c) Summary of labels over the transition.

Fig. 10 A model of utterance on the basis of status labeling (for an example of the taste of dish).

VII. Situation Recognition

In MIS, situation means mainly mental status of user. The mental status is so much affected by the degree of absorption which fairly depends on the environment in which user exists. As many media machines take into account, time and place are major factors of the user environment. However in MIS, it is appropriate to start from the situation under the physically static environment like indoor or some kind of fixed location even if it is outdoor surrounding. After that, it is likely to consider that there exists mobile environment like listening in a vehicle or train or airplane or ship. However because it is not appropriate for users to see entertainment scene when actively driving, those mobile environments are basically restricted to the cases in which those are regarded as simply "the sight-varying static environments". The most important point is how these situations affect the mental status under absorption regardless the static or the mobile environment. Where, remark that the absorption does not necessarily requires shutting out the apparently trivial things which seems to be no relation to the music or scenes directly. This is because those kinds of miscellaneous things often make sense in other layers of mental status.

From the aspects of mental state, ongoing situation under evolvement induced by music and the past mental state is highly estimated. It is natural to involve the effects of scenes and sounds which the user has experienced through the interaction with MIS. In view of this aspect, it is appropriate for MIS to recognize or to track user's mental situation.

A. *Confidence Vector*

Confidence vector is defined as below.

$$\boldsymbol{C} = (C_1, C_2, \ldots, C_i, \ldots, C_N)^T \qquad (30)$$

Where C_i is i-th component of $\boldsymbol{C}$ and it means confidence degree of indexing A_i to one macroblock. Usually, a confidence vector is affected by two classes of various situation factors. The first class is regarding situations of scene capture such as time, place, and weather, etc. The second one is regarding image features such as color, texture, motion, two dimensional position, etc.

To be continued.

B. *Status Judgment by using Confidence*

Suppose that some of the vocabularies retrieved and prepared (or recalled) for a category Ak (k=1,…,K) corresponds to situation S(n) in current observed scene. For each vocabulary, it is possible to assign confidence ckj(n) on discrete time n on the basis of the observation of image features (such as DCT coefficients and motion vectors in MPEG, etc.), by applying individual appropriate rule base.

Only for the plural vocabularies which confidences on time n are over the preset thresholds, next frame which is to be observed on time n+1. Through the index which corresponds to the plural situation candidates S_m *(m=1,..M) that are included in hypothesis* S(n), new knowledge or rules defined in advance are retrieved and activated. Where, situation candidates are the combinations of vocabularies and feature vectors, and index L_m is attached. If the successive contradictions occur near the state around the time n, the index L_m shall be rejected.

VIII. Status Tracking

To be filled.

A. Brain Player

It is supposed that a feeling of 'silent' is sometimes misperception caused by forgetting. Similarly, the following factors may have a chance to affect the feeling of music:

- Audio effect or filtering
- Stereo or not
- Using headphone or not
- low volume or high volume.
- Car stereo which is considered to move.

1) Modification of memory

Not only feeling of real time sensing, but also memory of several modalities have a chance to be modified. Time period , time length, sense of long or short, color, etc. are the actual attributes. Sometimes those lead to misperception if real perception and verification are absent for a long time. However, roughly, TSM-ST effect described above still holds in this layer.

If the verification was not repeated for a long time, the relation between a label and it's corresponding coverage in feature space or the numerical distance will be deviated and sometimes lose reliable measure. This phenomenon could lead to modification of confidence. For example, the word 'silent' could have the sound state of quiet but very low volume of bass exists. The opposite phenomenon would be an exaggeration which is often seen in visual arts or paintings.

The attributes which can be applied to the mechanism above are as follows:

- Color, texture
- Length
- Time period
- Loudness
- Positional relation
- Location
- Impression of human face
- Configuration of objects
- Observation coordinate that might not be restricted to three or lower dimensional characteristics.
- Movements
- Magnitude, height, etc.

In mental space relating to MIS, the accuracy of the object is not a big problem but "what it is" is more important. Then the idea I have shown above in media description does not need the accuracy of industrial measurement devices.

2) Pseudo psychometrics

Also other than attributes themselves, functional mechanisms are simultaneously modified. As the results, the followings could happen:

(a) Inversion of the sign of attribute value

Ex.) The content which a listener holds a bad impression may turn to have a good impression. This might be triggered by a temporal change or deviation of sensibility and relative comparison which may affect the logical decision or logical order.

(b) Non-linearity

Time length no longer holds linear property or sequential property. Also the length and largeness of object does not necessarily match to the ones in the real world.

(c) Exaggeration

Color, shape, texture, brightness, etc. are sometimes exaggerated.

To some extent these effects above are interpreted as personal difference. However the range is usually restricted when assuming the personal difference. On the other hand, there seems to be no restriction in pseudo psychometrics above, however some rules that can be hold exists. At this stage the 'Rules' cannot be specified in detail.

Here, I propose the following hypothesis:

H1) Modality Change:

Only remarkable audio memories remains unchanged, but other memories of audio have been modified to the 'memory of comfort'. In this process, some parts of audio memories disappeared.

H2) Difficulty in simultaneous replay:

In BP, it is hard to perform simultaneous replay of plural sound parts.

H3) Nonlinearity in Time:

Replay time is often affected by the degree of concentration in BP.

H4) Random Access Capability:

Random access replay is very easy in BP. However, for other person's mind, the series of events replayed by random access would not be so natural or comfortable because other mind cannot anticipate the next event in advance. This phenomenon is very much similar to the disgusting feeling when watching video sequences captured by other person's video camera.

H5) Memorizing conditions:

Sound replayed by BP highly depends on sound or acoustic or audio condition when the sounds were memorized by Brain Recorder.

From the aspects of personal difference, it should be discussed regarding sensibility modeling which can represent individual characteristics and temporal difference. Empirically, being absorbed in more than one or two states simultaneously is usually fairly difficult for a single person under ordinary condition, although it s possible in rare case. For example, assuming that you are in three places including one real place and two virtual places simultaneously would be so much annoyed or distractive mental status.

B. Mind Map

The map described above is not restricted to any specific scenes or any specific geographic information. Moreover it does not necessarily contain correct answers. In this aspect, my proposed system is not classified into deterministic problem solving machine. The only restriction is that the relatively defined "corresponding statuses" are verified through labels

acquired as common language, on the basis of mutual communication process. Therefore it could be applied to every kind of perception based information and status other than visual information and physically acquired information. The author calls it as “Brain Map” that is the extended notion of so-called map.

C. Validation Gate

As the prominent works [ ] in the field of particle filtering, it is natural to consider the validation gate in each scene evolvement. Where, there would be at least two stages of validation gates. The first stage is that the validation gate for mental status of user which the MIS predicted or reasoned. The second stage is to validate the image internally generalized as the description level.

To be continued.

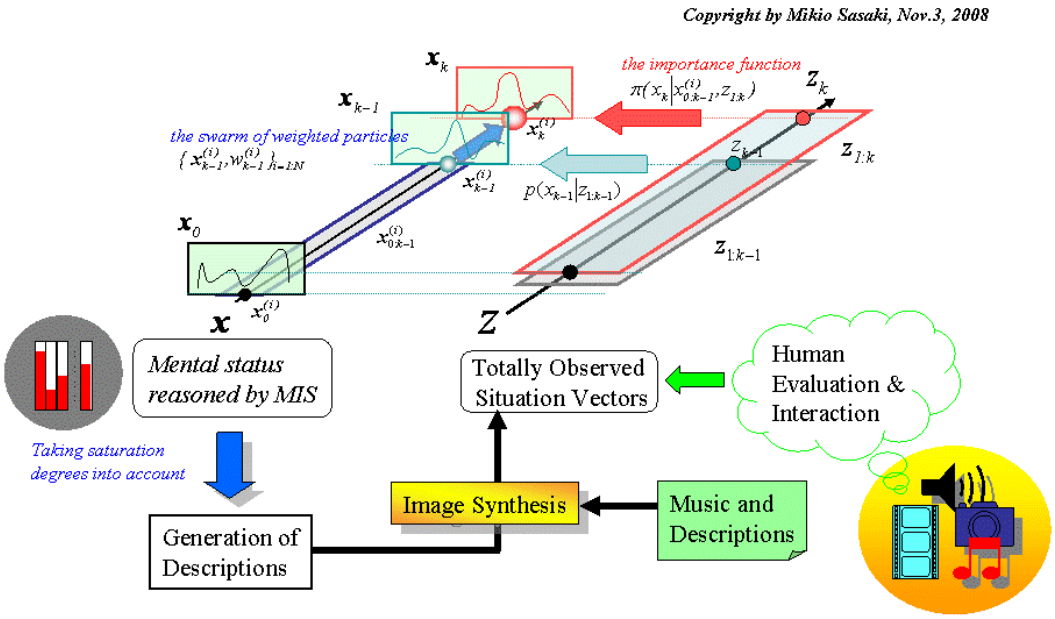


Fig. 11 Concept of particle filtering based tracking of mental status.

IX. Experiments

MPEG based experiments are expected.
These are under study.

A. Image Indexing

B. Image Retrieval

Details of this section will not be disclosed here.

X. Evaluation

Under study.

XI. Conclusion

Under study.

REFERENCES

[1] M. Sasaki, "What Kind of Images Do We Need in the Future Musical Instruments? - a point of contact with Vector Quantization and Knowledge Coding –", *PCSJ87*.

[2] M. Sasaki, "A Proposal for High-Efficiency Coding of the Definite Image Sequence", *PCSJ88*.

[3] P. Viola, M. J. Jones, "Robust Real-time Object Detection", *Cambridge Research Laboratory Technical Report Series CRL2001/01, Feb. 2001*.

[4] J.-M. Odobez, D. Gatica-Perez, and S. O. Ba, "Embedding Motion in Model-Based Stochastic Tracking", *IEEE Trans. Image Processing*, vol.15, no.11, pp.3514-3530, Nov. 2006.

[5] E. Arnaud, E. Mémin, "Partial Linear Gaussian Models for Tracking in Image Sequences Using Sequential Monte Carlo Methods", *International Journal of Computer Vision* 74(1), 75-102, 2007. http://perception.inrialpes.fr/~Arnaud/papiers/Ijcv07Arnaud.pdf .

[6] Xiang, Tao; Gong, Shaogang; "Beyond Tracking: Modeling Activity and Understanding Behaviour.", International Journal of Computer Vision, Vol. 67, No. 1, April 2006 , pp. 21-51(31).

[7] Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[8] "Guidelines for the TRECVID 2006 Evaluation", http://www-nlpir.nist.gov/projects/tv2006/tv2006.html .

[9] http://staff.science.uva.nl/~cgmsnoek/pub/mediamill-TRECVID 2006-final.pdf

[10] G. Pirirou, P. Bouthemy, and J.-F. Yao, "Recognition of Dynamic Video Contents With Global Probabilistic Models of Visual Motion", *IEEE Trans. Image Process.*, vol. 15, no.11, pp.3417-3430, Nov. 2006. http://www.irisa.fr/vista/Papers/2007_ip_piriou.pdf.

[11] Joakim Jiten, Bernard Merialdo, "Semantic Image Segmentation with a Multidimensional Hidden Markov Model". http://www.eurecom.fr/util/publidownload.en.htm?id=2079

[12] J. Vogel, B. Schiele, "Semantic Modeling of Natural Scenes for Content-Based Image Retrieval", *International Journal of Computer Vision* 72(2), 133-157, 2007.

[13] Julia Voge1, Adrian Schwaninger, Christian Wallraven, Heinrich H. B¨ulthoff, "Categorization of natural scenes:local vs. global information", http://www.cs.ubc.ca/~vogel/publications/apgv06_VogelSchwaningerWallravenBuelthoff.pdf .

[14] Julia Vogel and Bernt Schiele, "Performance Evaluation and Optimization for Content-Based Image Retrieval", *Pattern Recognition.* Vol. 39, No. 5, pp. 897-909, May 2006. http://www.cs.ubc.ca/~vogel/publications/PR06_VogelSchiele.pdf

[15] Anna Bosch, Andrew Zisserman, and Xavier Mun˜oz, "Scene Classification Using a Hybrid Generative/Discriminative Approach", *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol.30, no.4, pp.712-727, Apr. 2008.

[16] D. Hoiem, A. A. Efros and M. Hebert, "Recovering Surface Layout from an Image", *International Journal of Computer Vision* 75(1), 151-172, 2007.

[17] G. Pirirou, P. Bouthemy, and J.-F. Yao: "Recognition of Dynamic Video Contents With Global Probabilistic Models of Visual Motion", *IEEE Trans. Image Process.*, vol. 15, no.11, pp.3417-3430, Nov. 2006.

[18] "DIGITAL VIDEO RETRIEVAL at NIST - TREC Video Retrieval Evaluation", http://www-nlpir.nist.gov/projects/trecvid/

[19] Joakim Jitén Söderberg: "Multidimensional Hidden Markov Model Applied to Image and Video Analysis", A Thesis for the degree of DOCTOR OF PHILOSOPHY, Institute Eurecom Sophia Antipolis, February 2007.

[20] Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[21] Ashutosh Saxena, Min Sun, and Andrew Y. Ng, "Make3D: Learning 3D Scene Structure from a Single Still Image", *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol.31, no.5, pp.824-840, May 2009.

[22] Takeo KANADE, Conrad J. POELMAN, Toshihiko MORITA, "A Factorization Method for Shape and Motion Recovery", *IEICE Trans.* J76-D-II No.8, pp.1497-1505, Aug. 1993.

[23] E. Arnaud, E. Mémin, "Partial Linear Gaussian Models for Tracking in Image Sequences Using Sequential Monte Carlo Methods", *International Journal of Computer Vision* 74(1), 75-102, 2007.

[24] Ahmed Elgammal, Chan-Su Lee, "Tracking People on a Torus", *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol.31, no.3, pp.520-538, Mar. 2009.

[25] http://homepage3.nifty.com/MikioSasaki/scene/ref/TMBV.pdf

APPENDIX-A

Usually, multivariate linear regression model take the following form:

$$Y = XB \tag{A1}$$

In this case, it is well known that the estimated matrix of B in the least squares criterion is calculated by using the generalized inverse matrix of X:

$$\hat{B} = \left(X^T X\right)^{-1} X^T Y \tag{A2}$$

However, (2) has the following form:

$$\boldsymbol{Y} = \boldsymbol{FX} \tag{A3}$$

In this case, the estimated matrix of *F* in the least squares criterion is calculated by using the generalized inverse matrix of X:

$$\hat{\boldsymbol{F}} = \boldsymbol{YX}^T \left(\boldsymbol{XX}^T\right)^{-1} \tag{A4}$$

This is easily derived through well known growth-curve model as shown by next formula:

$$\boldsymbol{Y} = \boldsymbol{FBG}^T + \boldsymbol{E} \tag{A5}$$

Where, *Y* (*N*×*P*) is a data matrix, and *F* (*N*×*R*), G (*P*×*R*) are known matrices. B (*R*×*R*) denotes a unknown matrix and E means an error matrix. According to the excellent works by Penrose (1956), the next estimation equation was derived:

$$\hat{\boldsymbol{B}} = \left(\boldsymbol{F}^T \boldsymbol{F}\right)^{-1} \boldsymbol{F}^T \boldsymbol{YG} \left(\boldsymbol{G}^T \boldsymbol{G}\right)^{-1} \tag{A6}$$

By defining C and X as

$$\begin{aligned} \boldsymbol{C} &= \boldsymbol{FB} \\ \boldsymbol{X} &= \boldsymbol{G}^T \end{aligned} \tag{A7}$$

, the equation (A5) can be rewritten as below:

$$\boldsymbol{Y} = \boldsymbol{CX} + \boldsymbol{E} \tag{A8}$$

Therefore by using (A6), the estimation of *C* is obtained as follows:

$$\hat{\boldsymbol{C}} = \boldsymbol{YX}^T \left(\boldsymbol{XX}^T\right)^{-1} \tag{A9}$$

According to this conclusion, it is possible to have the same form of solutions for equation (2) and (3).

Mikio Sasaki was born in Kochi, Japan, in March 1958. He received the B.E. (1981) in Electronics from Kyoto University. After five years working for , he also received the M.E. (1988) in Electronics from the University of Tokyo. In 1991, he moved to . His major interests are media understanding, machine vision and knowledge processing.

This paper is still a draft version and the research is on going. All of the original properties written here do not belong to any specific companies or any other individuals. This document is being written under completely private research activities of the author.

M. Sasaki is with a private research institute in his home in Japan. Although he is working for an automotive parts company, it has nothing related to this research. If you are interested in this research send a e-mail to: (e-mail: ---------).