Scene Recognition using MPEG Features

- Towards MPEG Based Vision -

Mikio Sasaki,
http://homepage3.nifty.com/MikioSasaki/scene/2_1.htm

Abstract In MPEG video, principal feature extractions are almost finished at the time of video coding, especially regarding DCT (Discrete Cosine Transform) coefficients and motion vectors. In this paper, a scene recognition method which is based on the multi-variate linear regression utilizing the MPEG video features is proposed. Firstly, feature vectors are formulated on every block of pixels from the encoded features included in MPEG images and related situation information. Next, confidence vectors regarding object indices are estimated from these feature vectors and an index which has the maximum confidence value is determined as the recognized result. In the estimation process, the multi-variate linear regression method is adopted for the estimation, from the viewpoint of applications that need sequential update and distributed cooperation on the basis of learned results. At this time, some typical landscapes such as park and beach are taken as examples for the investigation on the recognition capabilities towards the MPEG-Based Vision.

Keyword scene recognition, MPEG, DCT, motion, confidence, multi-variate regression, time series analysis.

1. INTRODUCTION

There has been many computer vision related works for scene recognition according to pixel by pixel processing. However, recently, great amounts of compressed pictures are taking place of baseband pictures in various applications. One of the most influential formats for the video compression is MPEG that processes mainly blocks of pixels on the basis of MC-DCT (Motion Compensation - Discrete Cosine Transform) method. Also the corresponding encoding devices have been widely used at low cost by the benefit of the world wide ISO standard of MPEG. In the last two decades, no doubt MPEG has already been backed up by huge amounts of sophisticated technologies including software, hardware, and systems. Therefore video processing techniques which are highly familiar to MPEG based image compression techniques widely used in internet, multimedia, and digital broadcasting are strongly desired in view of interoperability. The standardization activities such as MPEG-7 and MPEG-21 in ISO are the very response to the desire, but the automatic description of the contents is still under research and development. Here, if the image recognition techniques which use MPEG encoded features directly exist, it will become unnecessary to process bitmap data which can be acquired after the decoding operations. Moreover, it would be very much fascinating and helpful in the point that the computational amounts and memory can be reduced and it become possible to utilize communication band effectively.

Conventionally, there have been tendency that the image understanding methods concentrate mainly on the acquisition of 3D structure and region decomposition. So the problems of the mapping to the semantics have been separated as application matters. On the other hand recently, the method to acquire the rough descriptions of the scene is proposed. It uses the finite numbers of vocabularies set registered in advance as the objects which could appear in the scene, so as to relate each block of pixels to the vocabularies by the analysis of color information.[1)4)].This is showing the new direction of the research and hereafter, the system which is to realize the ultimate feature is called MPEG-Based Vision in this article. A method which leads to the system is introduced as follows.

2. RELATED WORK

J. Vogel *et al.* calculated feature vectors of 180 dimensions which contain color, edge, co-occurrence matrix of intensity for a block of pixels (72×48) in a still picture of natural scene which size is 720×480 pixels. Moreover, they classified them into 9 categories by using SVM, and represented the rate of individual category in the total scene as 9 dimensional vectors. They called the feature vector Concept-Occurrence Vector and assigned one for a scene. Next, they classified the vectors into 6 categories such as Coasts, Rivers / Lakes, Forests, Plains, Mountains, Sky, by using SVM and assigned it as an index for a picture[10)].

J. Jiten *et al.* utilized HMM to assign the index to block of pixels in a still picture and applied the idea to tracking of plural people in video scene[15)].

G. Pirirou *et al*. performed semantic clustering of video database by using motion characteristics[13)]. The same kinds of trials are extensively researched at the worldwide consortium TRECVID[14)].

Regarding structure reconstruction, other than the famous factorization based method[16)18)], recently, A. Saxena *et al.* used MRF to infer 3D scene structure from a single till image[17)]. However, they do not refer to the semantic interpretation of the 3D structure of scenes.

In view of MPEG-based recognition, there are several works such as 3D motion extraction, face tracking, scene recognition, human tracking, etc. Although those works are mainly oriented for media description, they have not reached autonomous interaction with semantics.

3. SCENE RECOGNITION

3.1. Confidence Vector

For each block of pixels, the system extracts color, texture, and motion information directly from MPEG based encoded data without reprocessing decoded images. Then, rule based reasoning, multivariate regression based reasoning and simple 3D model based recognition are integrated to calculate the confidence vector that maps each block of pixels to 64 object categories. A confidence vector is defined as below.

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

where, C_i is ith 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, 2D positions, etc.

Then we can represent the causal relations between the situation

factors and the confidence vector by introducing the multivariate linear regression model such as:

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

where, $\boldsymbol{c}_0$ means 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.

4. STATISTICAL INFERENCE

4.1. Multi-Variate Regression Analysis

Therefore it is possible to 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 nth 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.

4.2. Situation vector

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 Ws 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).

In this article, $\boldsymbol{S}_S$ is called situation vector. It can be constructed by the following formula:

$$\mathrm{Ss}(\mathrm{m_r}, \mathrm{m_c}, \mathrm{k}) = [\boldsymbol{S}_{nenv}{}^T\ \boldsymbol{S}_{location}{}^T\ \boldsymbol{S}_{time}{}^T]^T \tag{7}$$

where, $\boldsymbol{S}_{a\$}\ (m_r,m_c,k)$ means the $L_{a\$}$ dimensional column vector which represents information regarding the situation attribute $a\$$ of $MBK\ (m_r,m_c,k)$.

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

$$\boldsymbol{S}_{\mathrm{time}} = (early\_morning, morning, \ldots, evening, night)^T \tag{8}$$

$$\boldsymbol{S}_{\mathrm{nenv}} = (fine, cloudy, rain, \ldots, storm)^{\mathrm{T}} \tag{9}$$

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

4.3. Measurement Matrix

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

$$\Psi = [S\ Y] \tag{10}$$

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.

$$\boldsymbol{s} = [\boldsymbol{S}_{\mathrm{P}}{}^{\mathrm{T}}\ \boldsymbol{S}_{\mathrm{S}}{}^{\mathrm{T}}]^{\mathrm{T}} \tag{11}$$

where,

$$\boldsymbol{S}_{\mathrm{p}} = [\boldsymbol{S}_{\mathrm{color}}{}^T\ \boldsymbol{S}_{\mathrm{texture}}{}^T\ \boldsymbol{S}_{\mathrm{block\_position}}{}^T\ \boldsymbol{S}_{\mathrm{motion}}{}^T]^T \tag{12}$$

$$\boldsymbol{S}_{color} = (I_Y, I_U, I_V)^T \tag{13}$$

$$\boldsymbol{S}_{texture} = (A_{hor}, A_{ver}, A_{all})^T \tag{14}$$

$$\boldsymbol{S}_{block\_position} = (m_c, m_r)^T \tag{15}$$

$$\boldsymbol{S}_{motion} = (v_x, v_y)^T \tag{16}$$

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 defined 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 $MBK\ (m_r,m_c,k)$.

5. EXPERIMENTS

5.1. Generation of Learning Data

Generally speaking, when using statistical recognition method including multi-variate regression, the recognition rate of individual object highly depends on the selection of learning data, even if the learning amounts are fixed. In this article, when processing MPEG-1 full frame video as learning data, the basic unit of learning data is one macroblock. Hereafter, the learning unit is called *event*. Also the fundamental unit of learning amounts is defined for one frame, namely, the maximum number of events per frame is 330 in MPEG-1 video. This is because it is required for the learning operation to consider the area rate for each object in the learning frame. Regarding the method to increase learning data, see reference number 1.

5.2. Evaluation of Recognition Rate

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 equal to 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:

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

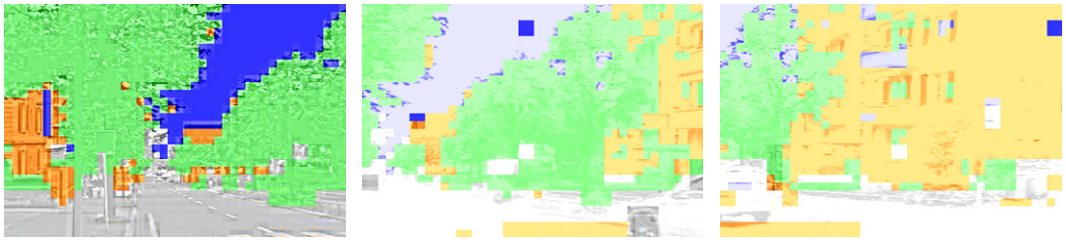
where,

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

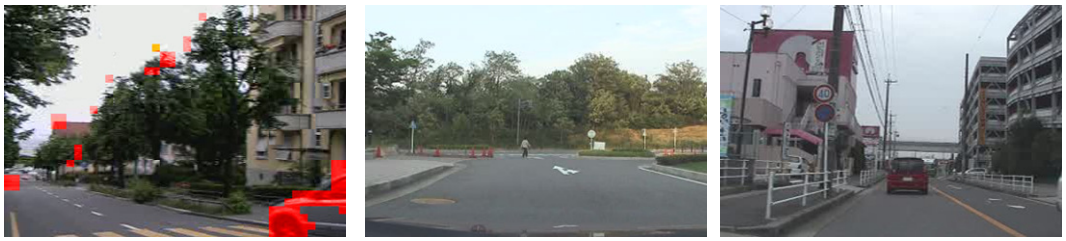
For example, if the view 'sightseeing' is designated, it is possible for the system to compute the recognition rate $Q_{sightseeing}$ by using index set $O_{sightseeing}$= {sky, tree, building, sea, beach} . Note that the view and its corresponding vocabularies of index set are determined by application and user profiles.

As the equation (20) 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

(a) Decoded image (070612_Lausanne, Frame No.1, 51, 74, respectively).



(b) Recognized results (070612_Lausanne, Frame No.1, 51, 74, respectively).



(c) Regions of high AC power (070612_Lausanne, Frame No.51)

(e) Learned scene No.1 (060514-1748_FOREST)

(f) Learned scene No.2 (070628-1744_APITA)

Fig.1 Recognition of street scene.

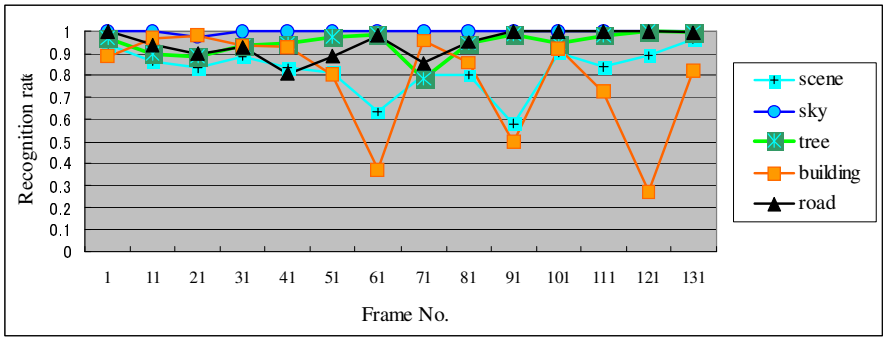


Fig.2 Temporal changes of recognition rates.

well known F measure for a single object is defined as the object recognition rate.

5.3. Experiments

Fig.1 shows the experimental results of street scene recognition. Fig.1(a) and Fig.1(c) shows the MPEG decoded images which are the targets of the recognition. Fig.1(a) is a I picture (intra frame), Fig.1(c) is a B picture (bi-directionally interpolated frame), Fig(b) and Fig.(d) are the recognized results, respectively. In order to discriminate difference block from intra block, the former is colored thickly and the latter is colored thinly for the same index.

For the learning task of equation (3), two MPEG sequences as shown by Fig.1(e) and Fig.1(f) are used. According to the procedure described in the reference article no.1, each MPEG sequence was processed independently to generate individual learning data. Finally the two learning data were merged into a single learning data which includes 2599 events and it was used as a measurement matrix in the equation (12). Fig.2 shows the temporal changes in recognition rates by 10 frame period.

6. INVESTIGATION

Following points are the summary of results of scene recognition experiments by using on-board camera [1].

(K1) The saturation tendency has been seen around 1000 events of learning data in scene recognition rate. When the amounts of learning are more than 2000 events, there exist some cases that the recognition rate decreases in turn.

(K2) The object which dominates large areas such as sky, tree, road, etc. has a tendency to show saturation in recognition rate as the learning data amounts increase.

(K3) Even when in an object which dominates small area, the recognition rate can increase if it is a rigid object which has special features in color and/or location under the increment of learning data.

(K4) For a non-rigid object which dominates small area and changes in shape and location like a pedestrian, the corresponding recognition rate shows extremely low values.

(K5) Regarding temporal changes of recognition rates, the object which continues to dominate large area tends to show small value.

On the other hand, the object which dominates small area or the object which incessantly appears and disappears shows large value of changes in recognition rate.

(K6) Scene recognition rates tend to become stable as the number of objects in the image plane becomes small.

(K7) Even if under the same situation factor, there exist some cases that the recognition rates dynamically change because of temporal variations in number of people, traffic amounts, color of the moving object, and configuration of clouds, etc.

(K8) The recognition rates for the object which dominates large area in the frame such as sky, trees, roads, are relatively stable even if the season changes. Scene recognition rates are not degraded.

(K9) Drastic changes of weather and surrounding views (road constructions, appearance and disappearance of buildings, moving objects which dominate large areas in the image plane, and the extent of sight being captured, etc.) affect the recognition rates.

(K10) Even when the different place is to be recognized as compared with the learned images, scene recognition rates tend to keep high value if the large areas of objects such as sky and road keep high value of recognition rates.

(K11) When a small object is the target of the predefined view, the temporal changes of the scene recognition rate is large. Also the object which occurrence number of appearance, approach, and vanish is high (signboard, billboard, on-coming vehicle, pedestrian) shows the same kinds of unstable behaviors regarding scene recognition rate.

(K12) The average scene recognition rate when the place and time is different from the learned image is low.

(K13) The symmetric property regarding the recognition rate does not exist when exchanging the roles (to be learned and to be recognized) between two images.

The summary above is concerning the case when using an on-board camera, which has difficulty in that many objects appear and disappear in sight from time to time. However on the other hand, there are some advantages that the scene structures can be modeled relatively easily than the case when a human captures the images by the camera in hand[5].

Fig.1(a) and Fig.1(c) show landscape images which were captured by a camera in a human hand. In this case, the camera attitudes drastically change (in horizontal direction, the camera turned almost 180 degrees) because panning and tilting were performed by the human. When the index set of the view is $O_{sightseeing}$= {*sky, trees, building, road*} , the recognized results could follow the real images from frame to frame and it seems relatively well for the first half part. But for the latter half part, the performance has been rather degraded. Totally, the average scene recognition rate resulted in 82.8% (Fig.2). It is considered the degradation was caused by panning, and the next step assignments are to improve the quality of difference block processing.

Another problem is that there are some regions within the building which are misrecognized as sky. It would be actually difficult even for a human to discriminate by the color and the texture alone. The solutions would be to let the system learn more similar images or to let it correct the semantics from the experiences of positional relation and shape information[2)6)].

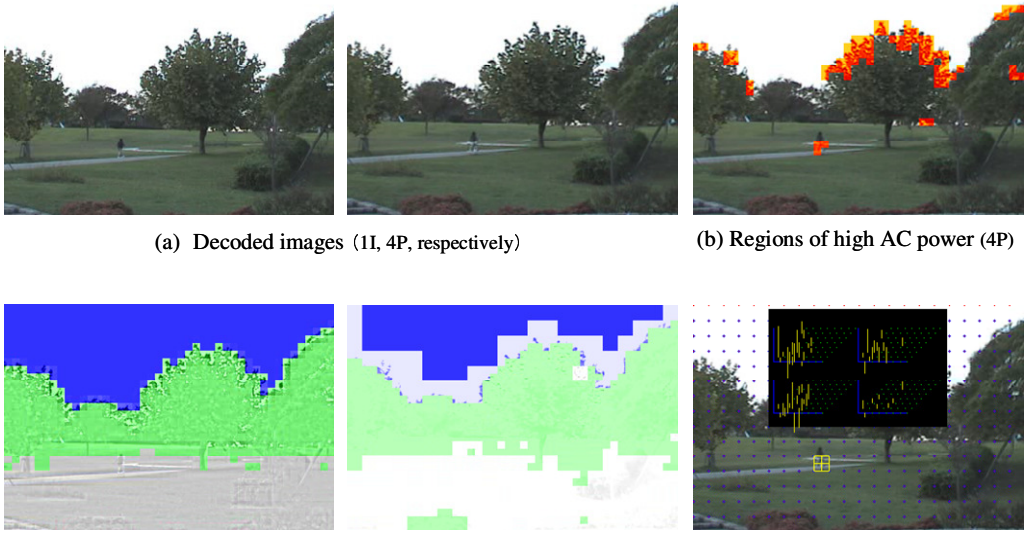
In the proposed method, although the number of feature dimensions are small (25 dimensions), there exists the case when the frame average of scene recognition rates shows 30% higher value than the SIFT （Scale-Invariant Feature Transform） based method[1)]. It is considered that the reason is ascribed to the point that SIFT is not appropriate to classify large area object which has few variations in spatial domain, although SIFT is excellent in extraction of local features on the basis of texture sensitive properties. Moreover, because MPEG can extract inter-frame features, it can be concluded that the MPEG-based method has strong advantages in both of computational amounts and the recognition performance for the scene recognition as the whole video analysis.

7. TOWARDS MPEG-BASED VISION

In order to become very much useful, it is desirable for the MPEG-Based Vision to have the recognition capabilities for far distant human, animals, and landmarks, etc. When restricting to the view of on-board cameras, it is reported that DCT coefficients of motion compensated difference signals which are described in MPEG data can be a trigger to detect pedestrians by incorporating time series analysis [3)]. Also, the tentative experiments to detect faces and to estimate three dimensional motion by using MPEG motion vectors information, and the consortium for image retrieval in TRECVID[9)] are to be focused.

According to the research trend, a trial for recognition of periodical time series changes included in a landscape by utilizing MPEG features is introduced in the following.

At first, a basic evaluation function for AC power within each block of pixels is defined as follows:



(a) Decoded images（1I, 4P, respectively）

(b) Regions of high AC power (4P)

(c) Recognized results（1I, 4P, respectively）

(d) DCT coefficients in the target blocks

Fig.3 Analysis of a scene in a park.

$$L_{AC}(l_{blk}) = \sum_{m=0}^{7}\sum_{n=0}^{7}\left|DCT(l_{blk}, m, n)\right| - \left|DCT(l_{blk}, 0, 0)\right| \quad (19)$$

Where, $DCT(l_{blk}, m, n)$ means two dimensional DCT coefficient in which l_{blk} denotes block number in MPEG macroblock (hereafter, we call MBK). And m and n mean horizontal coordinate, vertical coordinate, respectively.

Then we can define the sets of threshold values for judging the sum of absolute values of DCT coefficients according to the picture coding types that are defined in the ISO/MPEG standard. Basically, there are at least three types of picture coding types in MPEG standard such as I (intra), P (predictive), B (interpolational). In P or B frame, there exists MBKs that are coded as interframe difference or motion compensated difference over 16×16 pixels.

7.1. LPCA

Next, let the feature vector be composed from M_{walk} consecutive predictive frames in MPEG format. Here, the n-th ($n=1,\ldots,N$) feature vector is defined as follows:

$$\boldsymbol{x}_n = \left(x_{n1}, \cdots, x_{nP}\right)^T \quad (20)$$

Where, P means the number of dimensions for the feature vector. As we tentatively set M_{walk} to five, the duration for the time series is 0.5 second in our experiment. Then N becomes 64×5=320.

Corresponding measurement matrix is denoted as below.

$$X = \left(\boldsymbol{x}_1 \cdots \boldsymbol{x}_N\right)^T \quad (21)$$

This matrix should be defined on the data extracted from pedestrian scene in order to perform effective learning of dynamic characteristics of the target object..

Then, the covariance matrix S is defined as below.

$$S = \frac{1}{N} X^T X \quad (22)$$

The p-th ($p=1,\ldots,P$) eigenvector of S is denoted by

$$\boldsymbol{w}_p = \left(w_{p1}, \cdots, w_{pP}\right)^T \quad (23)$$

Then the principal component scores are calculated by using next formula.

$$\boldsymbol{f}_p = X\boldsymbol{w}_p \quad (24)$$

This is easily extended to the normalized matrix form as follows:

$$F = \left(\boldsymbol{f}_1 \cdots \boldsymbol{f}_P\right) = XWdiag(1/\sqrt{\lambda_1} \cdots 1/\sqrt{\lambda_P}) = XW\Delta_S^{-1/2} \quad (25)$$

Where, $W = \left(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_P\right)$ and λ_p denotes p-th eigenvalue of S, namely, $\Delta_S \equiv diag\left(\lambda_1, \cdots, \lambda_P\right)$.

7.2. Experiments

Fig.3 shows the results of scene recognition applied to a P frame of the scene of park landscapes. It can be seen that the judgment of the power of AC components by using some thresholds lead to the detection of boundaries of trees and pedestrians.

Next, by taking Fig.3(a) as a sample data, pedestrians are detected as shown by Fig.4, on the basis of the linear principal component analysis (LPCA) of feature vectors which are mainly composed from time series of DCT coefficients extracted from 10 P-pictures (predictive pictures) a second. The confidence degree in Fig.4 is a relative value which is calculated from top three principal components of the PCA for each block of pixels. It can be seen that the misdetections are reduced as compared with the case when using only the power of AC components, and as the result of that, pedestrians are detected for both of distant location and near location.

By using similar operations, white waves at a beach (Fig.5) and park trees trembling in the breeze (Fig.6) are detected.

Moreover, temporal periodicities (maximum periodicity is 1 second) of spatial spectrum are analyzed by applying FFT to the time series of the first principal components for the observed blocks in Fig.6(d) (Fig.6(f)).

7.3. Investigation

It is expected that the phenomenon which is apparently complicated in shape but showing some repeating behaviors can be judged by using this method described above. According to this expectation, several strategies for further improvements with less increasing the learning amounts are summarized as follows:

S1) Scene class adaptation
S2) Automatic ground truth generation or estimation
S3) Correction by semantics
S4) Utilization of situation factors
S5) High performance learning methods like KPCA, SVM, etc.
S6) Improvement of the recognition performance of P frames and B frames by using motion vectors and other dimensions:
S7) Time series analysis of feature vectors like PCA-FFT.

Particularly regarding S1), although the scene class has not been designated in advance for the experiments above, those are conducted on the assumption that the scenes are on the ground. Namely there are no vocabularies which are related to the water surface in the predefined dictionaries. That is the starting point at this stage. Therefore the recognized results for the input scene for Fig.5 have no index for the sea or water waves. There only exist the sky and road. Then the results in Fig.5 trigger the transition over the scene class relation graphs. This is based on the empirical rule that the surface of the road will not move dynamically and will not make fluid-like texture. There still remains the problem of the proof of the difference between the water surface and trembling trees or something which has complicated dynamic texture. These would be the future study and will not be discussed further here.

In Fig.5, the highest red bar means that this position of block was sampled and used as the self learning in LPCA of time series feature vectors. Then other block which has not a small value of confidence is showing the estimated white waves. For the sand

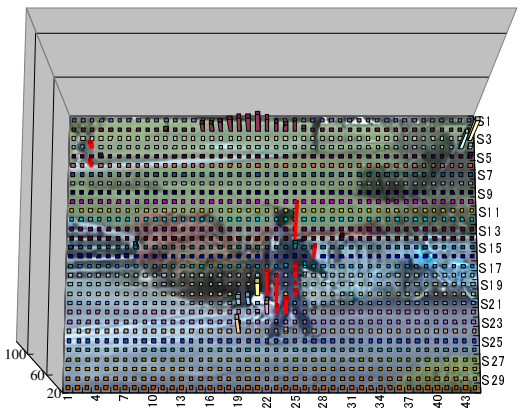


Fig.4 Detection of walkers.

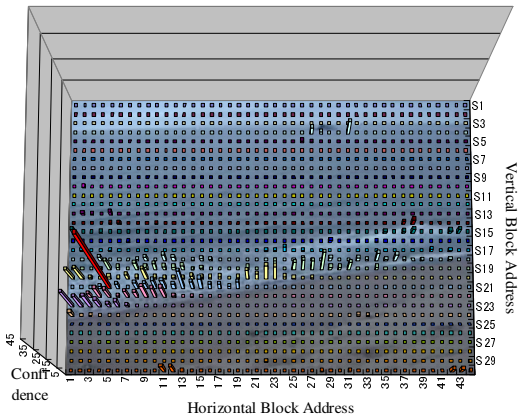
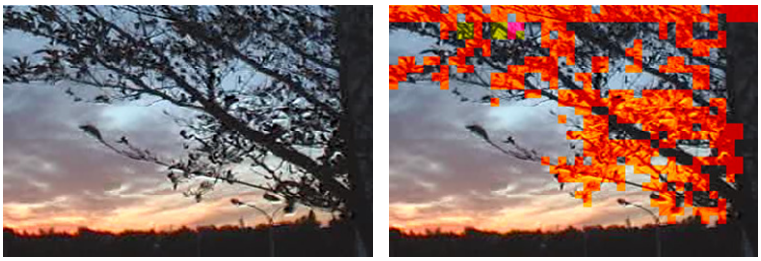
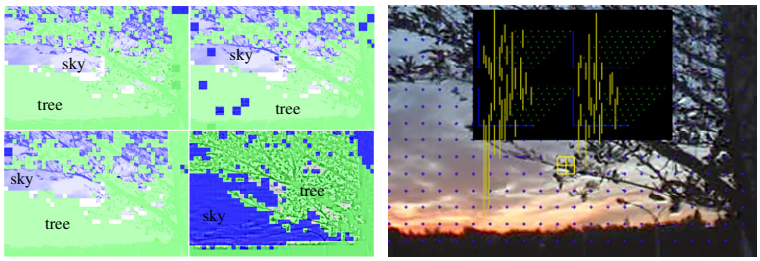


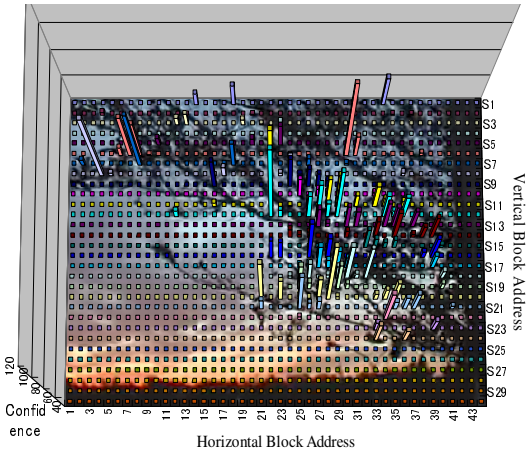
Fig.5 Detection of white waves.



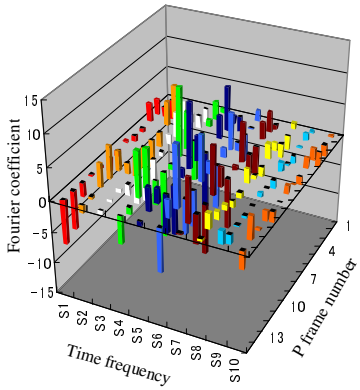
(a) Decoded image（46P: frame no.46, P frame）　(b) Change region on AC power and movement (46P)



(c) Recognized results (46P,47B,48B,49I)　(d) 2D-DCT coefficients in the target block（46P）



(e) Spatial distribution of confidence.（46P）



(f) Temporal changes of the first principal component.

Fig.6　Detection of movements of trees.

surface and the sea plane, almost no block of effective confidence exists. Then the white waves can be identified through the above signal processing and the adjacent surface which is identified as dynamic part by using the AC power would be the sea plane, which is also adjacent to the sky plane above. Remark that the sky plane is recognized in the first stage. On the other hand, the fairly static surface which was identified as the road would be the sand plane through the use of almost linguistic level of common knowledge.

For the experiments shown in Fig.6, the trembling parts of trees are detected through the use of PCA-FFT. It would be effective remedy especially for the boundary parts between the twilight, a little bit cloudy dark sky, and the trees. Those boundaries are partly degraded at the first stage of scene recognition, as can be easily seen in Fig.6(c). The empirical rules on the basis of the Fig.6(b)(e)(f) will improve the results if the appropriate scene class has been selected.

7.4. Next Step

Although nonlinear methods in the multi-variate regressing and the principal components analysis are considered as excellent judging properties, practically, it is rather hard for those nonlinear methods to acquire generic capabilities for the recognition. On the other hand, it is relatively easy for the linear methods to re-utilize the past learned properties and training data. In the statistic learning methods, it is highly heavy task to make ground truth and learning data by human manual operations. So, the reduction of the amounts of the manual operations is eagerly hoped. Then the next step assignments would be the effective database generation of learning data and vocabularies according to the systematic constructions of the application oriented scene class database.

8. CONCLUSION

As a scene recognition method towards MPEG-Based Vision, the author has proposed the multivariate regression analysis based estimation method which uses situation factors and image features as observation values, on the basis of the definition of confidence vectors corresponding to the finite number of vocabularies for a unit of block of pixels. The experimental results show that the average scene recognition rates are roughly good if there are few large changes in object configurations and situation factors (place, time, season, and weather)[1]. The proposed method is effective when recognizing a full frame video as a single cluster of objects in one scene. This nature could lead to automatic judgment of scene class [2]. However, for an object which is difficult to be described within stereotyped architecture learned in advance, or for an object which shows subtle temporal changes like a human, animal, and trembling leaves and trees, and water waves, it is required to add new recognition operations as introduced in this article. The author has not judged to what extent it can become possible within MPEG features. At least, there should exist some kinds of features if a man can observe and judge the target objects in MPEG decoded videos. It would be a very important point when designing algorithm that whether the interpretations are fully ascribed to genuine signal processing or depends on knowledge processing capabilities of a human who observes the

target video.

According to the contents above, the author would like to construct a automatic method to select the optimum learning data and to perform the optimum modification of feature vectors.

References

1) Mikio Sasaki, "A Proposal for Scene Recognition Method Suitable for MPEG Video", *The Journal of the IIEEJ*, vol.38, no.6, pp.890-899 (2009-11).
2) Mikio Sasaki: "MPEG-based Scene Class Recognition", *IEICE Technical Report*, vol.109, no.148, IE2009-55, July 2009.
3) Mikio Sasaki: "MPEG-based Scene Class Recognition", http://homepage3.nifty.com/MikioSasaki/scene/090724_IEICE_SceneClass.pdf
4) M. Sasaki, "Investigation on Scene Recognition on the basis of Statistical Reasoning", AVM-54-8, *IPSJ SIG Technical Reports*, Sep. 2006.
5) M. Sasaki, "Indexing of On-Board Captured Images Based on Color Information and Knowledge Processing", *DENSO TECHNICAL REVIEW*, Vol.12 No.1 2007, p.58-68.
6) Mikio Sasaki, Kenji Muto, Kunihiko Sasaki: "Estimation of Driving Safety from On-Board Captured Scene by using Probe Image Database",*ITST2007*, June 2007.
7) Mikio Sasaki, Hideaki Nanba: "Pedestrian Tracking by using MPEG-based Video Signal Processing", *IEICE Technical Report*, vol. 108, no. 344, IE2008-106, pp. 17-22, Dec.
8) M. Sasaki, H. Nanba, "Intelligent Information Assistance in ITS based on Mixed Technology of Vision and Communication - Towards Knowledge Sharing Type of Environment Recognition –", *JSAI 2006*, Jun. 2006.
9) "DIGITAL VIDEO RETRIEVAL at NIST - TREC Video Retrieval Evaluation", http://wwwnlpir.nist.gov/projects/trecvid/
10) J. Vogel, B. Schiele, "Semantic Modeling of Natural Scenes for Content-Based Image Retrieval", *International Journal of Computer Vision* 72(2), 133-157, 2007.
11) 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.
12) D. Hoiem, A. A. Efros and M. Hebert, "Recovering Surface Layout from an Image", *International Journal of Computer Vision* 75(1), 151-172, 2007.
13) 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.
14) "DIGITAL VIDEO RETRIEVAL at NIST - TREC Video Retrieval Evaluation", http://www-nlpir.nist.gov/projects/trecvid/
15) 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.
16) Jianbo Shi and Carlo Tomasi. "Good features to track", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.
17) 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.
18) 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.
19) 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.
20) M. Sasaki, "A Proposal for High-Efficiency Coding of the Definite Image Sequence", *PCSJ88*.
21) 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.

Mikio Sasaki received the B.E. (1981) in Electronics from Kyoto University, the M.E. (1988) in Electronics from the University of Tokyo. His major interests are media understanding, machine vision, knowledge processing, and MPEG related technologies. He is the member of IEICE, IPSJ, JSAI, ITE, IIEEJ, and IEEE Computer Society (PAMI). This article is based on his private research works. Accordingly, no other official information is described here.