CREATING JOYFUL DIGESTS BY EXPLOITING SMILE/LAUGHTER FACIAL EXPRESSIONS PRESENT IN VIDEO

Uwe Kowalik Kota Hidaka Go Irie Akira Kojima

NTT Cyber Solutions Laboratories, NTT Corporation
1-1 Hikarinooka Yokosuka-shi
Kanagawa 239-0847 Japan
+81 (0) 46 859 2258
E-mail: kowalik.uwe@lab.ntt.com

ABSTRACT

Video digests provide an effective way of confirming a video content rapidly due to their very compact form. By watching a digest, users can easily check whether a specific content is worth seeing in full. The impression created by the digest greatly influences the user's choice in selecting video contents. We propose a novel method of automatic digest creation that evokes a joyful impression through the created digest by exploiting *smile/laughter* facial expressions as emotional cues of *joy* from video. We assume that a digest presenting smiling/laughing faces appeals to the user since he/she is assured that the smile/laughter expression is caused by joyful events inside the video. For detecting smile/laughter faces we have developed a neural network based method for classifying facial expressions. Video segmentation is performed by automatic shot detection. For creating joyful digests, appropriate shots are automatically selected by shot ranking based on the smile/laughter detection result. We report the results of user trials conducted for assessing the visual impression with automatically created 'joyful' digests produced by our system. The results show that users tend to prefer emotional digests containing laughter faces. This result suggests that the attractiveness of automatically created video digests can be improved by extracting emotional cues of the contents through automatic facial expression analysis as proposed in this paper.

Keywords: affective media processing, video skimming, image processing

1. INTRODUCTION

The recent rapid development of the Internet infrastructure, high-performance computing at reasonable cost, and the availability of high-capacity mass-storage devices at low cost, as well as the introduction of high-quality digital video technology into the home consumer market, has created the basis for new network-based multimedia services. Video sharing websites, IPTV- and VOD-services as well as private content collections provide a massive amount of video contents. This means that the user is confronted with the problem of finding, accessing, and managing contents appropriate to his/her needs. In this context, video abstractions provide an effective way of gaining a rapid understanding of a video without the need for watching it in its entirety. In recent years, much video research has been focused on how to efficiently create video abstractions automatically since the vast amount of contents means that manual abstraction is not practical.
There are two types of video abstraction: static key frame abstraction and dynamic video abstraction, so called video skims. Both types aim to provide the user with a compact overview of the contents. Whereas key frame abstraction yields one or more still images extracted from the video sequence, video skims consist of video segments. Our work focuses on the latter approach. There are two types of video skims: video summaries and video highlights, also referred to as video digests. Whereas video summaries intend to provide a compact overview of the *story*, digests aim to attract the user by presenting *highlights* of a video.

Affective media processing has recently gained the attention of many researchers in this field. Especially in applications that demand the personalization of content distribution it is assumed that affective analysis of diverse media contents has great potential. In the future, users are likely to demand content identification through not only objective information (e.g. genre, actors' names, director, etc.), but also emotional Meta information, i.e. looking for e.g. funny or sad movies. Affective labeling of video contents thus will be essential to serve the users' needs. We believe that especially a joyful impression appeals to the user when selecting contents. We therefore focus on automatic creation of *joyful* digest in the present work. In front of this background the present paper proposes a novel method of extracting highlights from video contents for creating emotional digests by detecting joyful events based on the presence of joyful facial expressions such as *smile* and *laughter*. The paper is organized as follows. Section 2 refers related works and explains the idea behind our approach. Section 3 introduces the proposed method in detail. Section 4 presents a performance evaluation carried out with a prototype system implementation using the proposed method. Moreover we present the results of the subjective user experiments. Section 5 finally concludes this paper and gives an outlook to future work.

2. RELATED WORK

Affective content labeling schemes can be roughly

categorized into dimensional approaches and event detection approaches. Dimensional approaches try to map low-level features of the video and/or audio such as motion parameters, cut frequency, color histogram and sound energy into a continuous emotion model space as e.g. the P-A-D model (P-A-D = *Pleasure-Arousal-Dominance*) [1], whereas the type of features and the number of used model dimensions may vary [2][3][4]. The reported results show a high accuracy. However as stated by the authors in [2], the relations between the affect dimensions and low-level features known so far are rather vague and it is currently still difficult to determine the causal relation between low-level features and affect.

Event based approaches in contrast focus on detecting specific events that are related to emotional expressiveness. Highlight detection from audio streams by HMM is proposed in [5]. Three audio events i.e. laughter, applause, and cheering can be detected. In [6], an audio-based method for extracting laughter portions for the specific domain of Consumer Generated Video (CGV) is proposed. This method estimates laughter probability from prosodic parameters of an audio track; it utilizes a generalized state space model consisting of acoustic models and a state-transition model. The segment-based detection result is then used for creating a digest.
Audio-based methods have the drawback that the detected audio event may not reflect the actual emotional situation in the video, for example a situation with background laughter is likely to be classified inappropriately. Another short-coming of audio-based methods is that they can not detect emotional events that appear without audio. For instance, a silent smile on a person's face can relate to a very joyful story event, but will not be detected. However, since audio-based methods such as [6] definitively have their advantages e.g. in situations, where visual features are not present or can not be reliably extracted, they can be complemented by vision based methods improving the overall performance in multimodal systems.

Although low-level visual features have been fairly studied in the context of affective video abstracting, there exist to our best knowledge no work on exploiting high-level features such as facial expressions for the problem. Previous work in the field of emotion research provides some evidence that a human face reveals the emotional state of a person by displaying blends of six basic emotions (*joy, sadness, anger, fear, disgust, surprise*). Furthermore facial expressions are universal and independent of the cultural background [7][8]. Exploiting these properties of human faces in affective media processing seems very attractive and we therefore propose a new method based on *smile/laughter* face detection to achieve our objective of creating *joyful* digests automatically. As suggested by related work in the field of behavioral science, seeing 'happy' faces can evoke a 'happy' feeling in a person [13]. Another idea behind our proposal is that facial expressions are related to events inside the video, i.e. an actor or a person shows a specific facial expression in response to a story event. We assume that video segments containing smiling or laughing faces are likely to contain *joyful* story events, thus detecting *joyful* segments by automatic *smile/laughter* detection and their inclusion into a video's digest can convey the joyful impression to the audience through the created digest.

3. PROPOSED METHOD

Fig. 1 depicts the general outline of the proposed method. A digest created by using only facial expression based segmentation is of limited value to the user since it contains only face images. Moreover the duration of facial expressions tend to be very short. Inclusion of such short segments into a digest is assumed to negatively influence the impression created by the digest. Therefore the video content is first subjected to shot detection for video segmentation. This is to ensure that segments included into the digest will contain some story context. Our system implementation uses the cut-detection method introduced in [9]. In parallel frame wise face detection is performed employing the approach described in [10]. Detected face regions are then classified regarding the displayed facial expression. We developed a neural network based method for classifying face images into two classes namely *smile/laughter* faces and *others*.

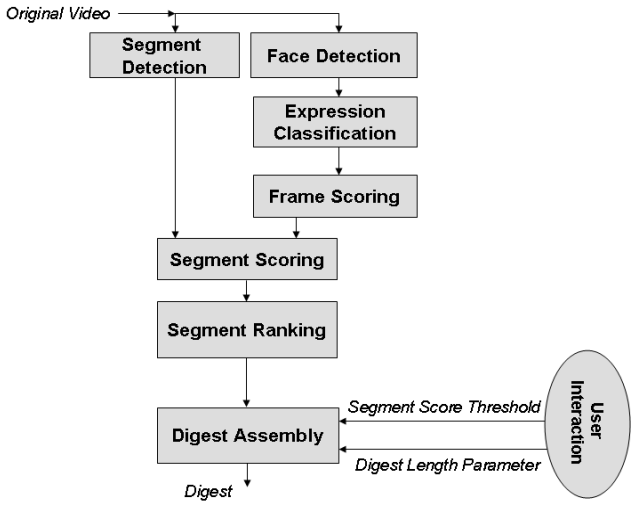

Fig. 1 System Overview

The expression classification result is then used for calculating a frame expression score *(FE score*). Next a segment expression score (*SE score*) is calculated from the *FE score* for affective shot ranking. Shot selection is then carried out based on the target expression score and the target length of the digest (both given by the user). Selected shots are finally concatenated to form the digest. In the following subsections we describe first our method for detecting *smile/laughter* faces, explain then the method of selecting appropriate segments based on *SE score* ranking and describe finally the step of digest assembly.

3.1. *Smile/Laughter* Expression Detection

The facial expression detection applied to extracted face

regions employs a neural network for feature classification and consists of the following processing steps.

1. face image preprocessing
2. feature vector extraction
3. feature vector classification

First color space conversion to 256-gray level pixel format, image size normalization to 128x128 pixels and histogram equalization is applied to the detected face regions. A 512-element feature vector is then extracted by grid-like sampling at 8x8 equidistant face image locations (Fig. 2).

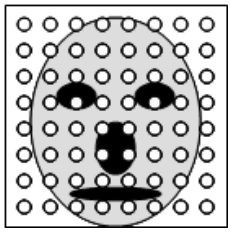

Fig. 2 Equidistant 8x8 grid sampling scheme for feature extraction from face images

The feature vector consists of Gabor-filter coefficients (*Jets*) calculated at each sampling position. In the following we will give a brief review. Given a gray level image $I(\vec{x})$, the *Jet* $J(\vec{x})$ is extracted by transforming the image at the point $\vec{x}$ by

$$J(\vec{x}) = \int I(\vec{x})\psi_i(\vec{x}-\vec{x}')d^2\vec{x}' \tag{1}$$

with a number of Gabor kernels ψ_i where

$$\psi_i(\vec{x}) = \frac{\|\vec{k}_i\|^2}{\sigma^2} e^{-\frac{\|\vec{k}_i\|^2\|\vec{x}\|^2}{2\sigma^2}} \left[e^{j\vec{k}_i\vec{x}} - e^{-\frac{\sigma^2}{2}} \right] \tag{2}$$

Each Gabor kernel describes a plane wave that is enveloped by a Gaussian function. σ defines the width of the enveloping function (here $\sigma = 2\pi$). The first term inside the brackets defines the oscillation whereas the second term compensates against average changes in lighting. The vector $\vec{k}$ describes the wave parameters and is defined as

$$\vec{k} = \begin{bmatrix} f_\nu \cdot \cos\varphi_\mu \\ f_\nu \cdot \sin\varphi_\mu \end{bmatrix} \tag{3}$$

with $f_\nu = 2^{-\frac{1}{2}(\nu+1)}$ and $\varphi_\mu = \mu \cdot \frac{\pi}{8}$ where ν is the frequency parameter and μ specifies the orientation.

We use only one frequency ($\nu = 5$) and eight orientations ($\mu = 0...8$) in our implementation. Furthermore only the magnitude is used to form the feature vector. The extracted feature vector is then subjected to a classification step by a feed-forward neural network classifier (FFANN) in order to determine the presence of a *smile/laughter* facial expression. This paper adopts the TACOMA procedure proposed by Lange et al. to generate a FFANN whose size is optimum for the feature classification task [11]. We chose the logistic transfer function for the neurons and use *QuickProp* learning algorithm. The output was trained to give '1' for smile/laughter face samples and '0' for samples of other facial expressions. The detailed results of a classifier performance evaluation are presented in subsection 4.1.

3.2. Segment Scoring and Segment Ranking

For segment selection we propose calculating a segment score based on the results of face detection and facial expression classification in order to determine joyful segments inside a video. Our proposed method calculates the segment score V_i accordingly to equation (4). V_i is equal to the *SE score* mentioned earlier.

$$V_i = \frac{1}{N} \sum_{k=0}^{N-1} G_k \tag{4}$$

N is the number of frames that actually contain faces inside segment i with $i \in [0...M-1]$ as the segment index for a given number of segments M and $k \in [0...N-1]$ refers to the frame index within a segment. The *SE score* yields $V_i \in [0...1]$. G_k is calculated according to formula (5) and equals to the *FE score* previously introduced.

$$G_k = \frac{L_k}{F_k} \tag{5}$$

The frame expression score G_k is the number of faces showing a the specific facial expression L_k divided by the total number of faces per frame F_k inside the k-th video frame. In the case that no face was detected in a frame ($F_k = 0$), we define $G_k = 0$. After the calculation of the score for each segment a segment candidate list is generated containing all segments in descending score order. Fig. 3 depicts the candidate *shot* list creation process based on the *smile/laughter* expression score implemented in our system for joyful video segment detection. Important to mention here is that the original timing information of the shot appearance inside the original video is maintained in our system for the following digest assembly step.

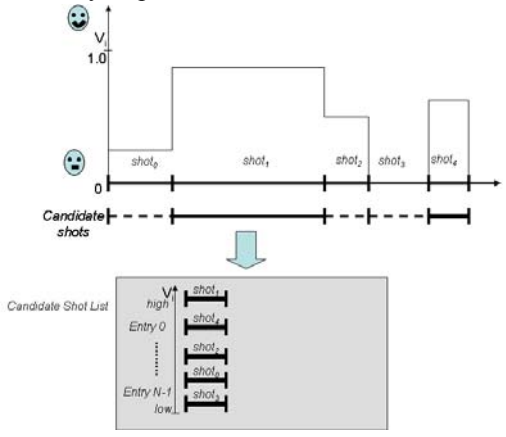

Fig. 3 Candidate shot list creation based on *SEScore*

3.3. Digest Assembly

In general, the final digest assembly step consists of selecting a number of segments from the list of candidate segments with the goal of meeting the user-defined digest length. In our prototype implementation the length of the final digest is controlled by two user-defined parameters:

1. Maximum number of shots to be included
2. Threshold of smile/laughter expression shot score (*SE score*)

Both parameters can be interactively set by the user via a GUI. Fig. 4 illustrates the digest assembly step.

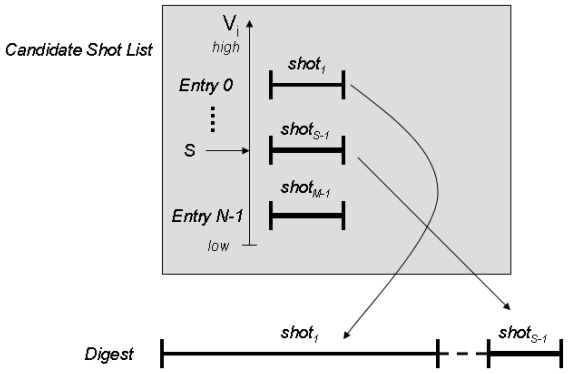

Fig. 4 Digest Assembly

Digest assembly is performed by lookup and concatenation of all those shots having a *SE score* exceeding the threshold. Selection from the candidate list is performed in descending order of their appearance inside the list. This ensures that shots with high smile/laughter expression scores are preferentially included. Given maximum target number S of shots to be included into the final video skim, $shot_{S-1}$ yields the last shot of inclusion. Since the timing information for each shot is maintained throughout the digest creation process, the sequential order of shots is not changed with respect to the original time line in order to keep the story line.

4. EXPERIMENTS AND RESULTS

In this section we first present results of the smile/laughter face classifier evaluation in subsection 4.1. We also performed a subjective user experiment in order to evaluate, whether digests generated by our proposed method can evoke a joyful impression at the audience. In the experiment we focus on the visual impression. In subsection 4.2 we explain the experiment and present the results.

4.1. Smile/Laughter Classifier Evaluation

We prepared two sample sets: one for classifier training and one set for classifier testing. Table 1 shows the approx. number of samples included in each set.

Table 1 Number of face samples used

	Learning	Test
positive	5000	800
negative	10000	1500

As for the sample creation we applied the face detection integrated in our system to numerous video contents captured from TV and movie contents and used directly its output in order to capture possible variations in the face detection result. Face image samples were than manually labeled into *smile/laughter expression* class and *other expressions* class. Labeling was performed by two subjects independently and a labeler agreement of ~96% was calculated partially on the training data set. According to our goal of detecting facial expressions in arbitrary video sequences, we included faces samples with different directions and quality levels. Fig. 5 shows some examples of face samples included in the training data set.

Fig. 5 Examples of face samples used for classifier training

Fig. 6 shows the classifier ROC curve created by applying a decision threshold to the FFANN output and varying the threshold between '0.0' and '1.1'.

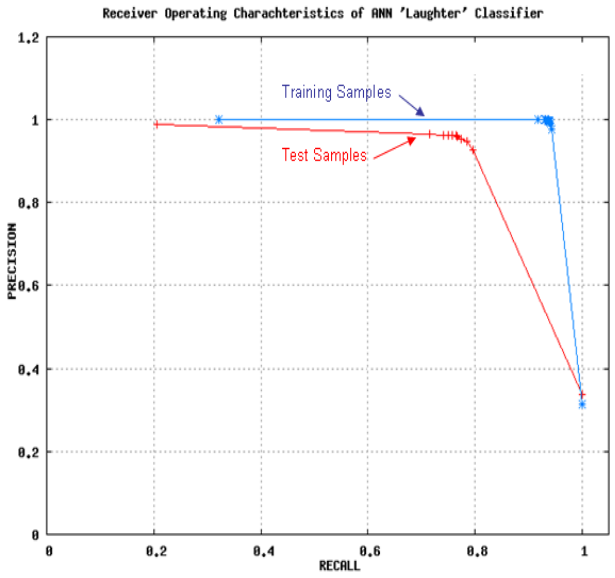

Fig. 6 FFANN classifier ROC curve

We set the classifier threshold at the point of maximum F1-measure value. The F_1-measure is the harmonic mean of precision and recall and calculated as given in equation (6).

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \qquad (6)$$

Table 2 shows the precision and recall for the threshold th=0.2 leading to the highest F_1-measure value of 0.86 for the test data. This classifier configuration is used during the subjective user test described in the following subsection.

Table 2 Classifier Performance (th=0.2, F_1=0.86)

Precision		Recall	
Learning	Test	Learning	Test
0.99	**0.95**	0.94	**0.78**

4.2. Subjective User Test

The goal of the subjective user test was to evaluate, whether the digests created by our proposed method can create a joyful impression at the audience. This time we were interested only in the visual impression, therefore no audio was used. We prepared a set of digests from 3 TV drama contents using our prototype system. Four different types of digests were created (Table 3). *Type I* digests were created by random shot selection. Digests of *Type II* were created by setting a low target *SE score*, i.e. many faces but few or no laughter expressions were present. *Type III* digests yielded high *SE score*, i.e. the digests contained many laughing face scenes. In order to see the impact of face presence on the impression created by a digest we prepared a fourth digest type in addition that contained no or only a few faces. Overall, 12 different digests were created and presented to the subjects. In order to create digest *Type II* and *IV* we modified the system by introducing an upper threshold th_u that defines the maximum value of V_i and a switch controlling a face presence parameter for shots to be subjected to inclusion into the shot selection list and final video skim respectively.

Table 3 Contents used for subjective tests

Title	Orig. length	Avg. digest length
Japanese TV Drama 1	15 min	13 sec
American soap opera	22 min	13 sec
Japanese TV drama 2	1 h 11 min	33 sec

Fig. 7 illustrates the influence of upper bound settings for th_u and the face presence switch on the digest creation. Digest *Type II* (no laughter) is created by setting the upper threshold for V_i to a lower value th_{u2}. The face presence switch is set to false for digests of *Type IV*.

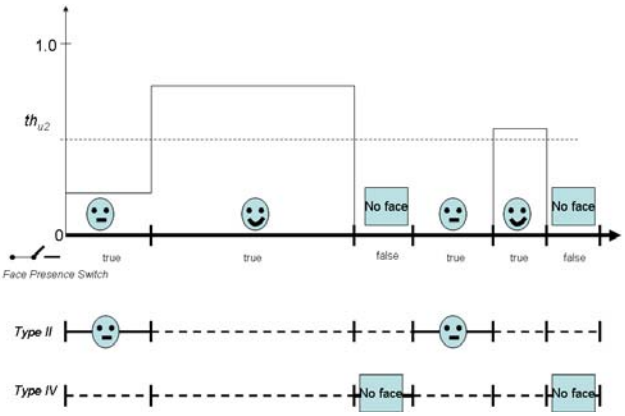

Fig. 7 Creating Type II and Type IV digests

Six subjects, 3 male and 3 female, age between 20 and 40 years, participated in the experiment. During the experiment the digests were presented in random order. Subjects were asked to give their impression rating whether the digest they just saw created a joyful impression. A 0-3 Likert-scale was used for assessment where a score of '0' refers to 'not at all' and a score of '3' to 'very strong'. The score was to be given immediately after watching a digest by each subject. As for the comparison the *Type I* digest served as base line. The average scores and the differential score with respect to the base line are listed in Table 4. *Type III* digests received the highest average score, i.e. the result shows an improvement of the joyful impression for *Type III* video skims containing many laughter scenes over randomly created digests. Digests of *Type IV* (no or little faces contained) received almost same average score as random digests, whereas *Type II* digests in contrast received the lowest average impression score i.e., the joyful impression decreased with respect to the base line. The reason for this result is that faces inside video attract usually the attention of the audience [12]. We interpret that the audience' impression is strongly influenced by the facial expressions since the face gains the focus of attention. Therefore the 'joyful' impression score dropped in our experiment when users were confronted with faces displaying e.g. expressions of anger or sadness. In contrast the joyful impression tends to improve when users see 'happy' faces. As we found all scores fall below the expected average of 1.5. We infer that this result is related to the choice of the contents used for the experiments. In a short interview after each session many subjects claimed that two TV-dramas were actually not very funny in terms of their story but evoke a more sad or heavy impression.

Table 4 Average score and relative score wrt. Type I

	Type I (random)	Type II (no laughter)	Type III (laughter)	Type IV (no face)
Avg. score	1.1	0.78	1.39	1.06
Diff. wrt. base line	0.0	-0.32	0.29	-0.04

Another reason for the low average score can be seen in the fact that the digests were presented without audio for the sake of visual impression assessment. Many subjects commented that they were not able to easily grasp the contents without audio, which influenced their rating in terms of the 'joyful' impression negatively. Moreover many users also said that the inclusion of short shots gave a negative impression. We interpret this also as one reason for the low average score.

5. CONCLUSION AND FUTURE WORK

In this paper we have proposed a novel approach on creating emotional video digests by detecting joyful segments through the presence of facial expressions of joy such as *smile* and *laughter* that are often related to *joyful* story events and displayed by people/actors in response to such events. We presented results of subjective user experiments conducted regarding the visual 'joyful' impression for the evaluation of our method. The results show that the user's impression in terms of 'joyful' can be improved by the explicit inclusion of shots containing *smiling/laughing* faces into the video skim. The results are promising and therefore we will extend our research in order to investigate, whether the present approach is feasible also for other facial expressions such as e.g. *sad* or *surprise*. We plan to investigate different approaches to improve the excerpt selection and assembly process in our current system. Another issue we will address in our future work is the impact of audio information on the created impression.

ACKNOWLEDGEMENTS

We would like to thank our project manager Yoichi Kato, for encouraging us to undertake this research. We appreciate the hard work of our support staff member Megumi Machiguchi involved in data labeling and received great support from Shingo Ando during software integration phase. Our special thanks to all group members for the fruitful discussions and continuous motivation.

REFERENCES

[1] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament", Current Psycho. Vol.14 (4), pp.261-292, 1996

[2] A. Hanjalic, Li-Qun Xu, "Affective video content representation and modeling", IEEE Trans. on Multimedia, Vol. 7, Issue 1, p.p. 143-154, Canada, 2005

[3] S. Arifin, P. Cheung, "A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information", Proc. of the 15th ACM Multimedia, pp. 68 – 77, Germany, 2007

[4] H. Kang, "Affective Content Detection Using HMMs", Proc. of the Eleventh ACM international Conference on Multimedia MULTIMEDIA '03, USA, 2003

[5] R. Cai, L. Lu, H. Zhang, L. Cai, "Highlight sound effects detection in audio stream", Proc. ICME '03, Vol. 3 pp. 37-40, USA, 2003

[6] G. Irie, K. Hidaka, N. Miyashita, T. Satou, Y. Taniguchi, "A Video Skimming Method for Detecting 'Laughter' Scenes in Consumer Generated Videos", Journal of Institute of Image Information and Television Engineers ITE, Vol. 62 No.2, pp. 227-233, 2008 (Japanese)

[7] P. Ekman, D. Keltner, "Universal facial expressions of emotion", in U. Segerstrale & P. Molnar (Eds.). Nonverbal Communication pp.27-46, Mahwah NJ:LEA (1997)

[8] J. D. Boucher, G. E. Carlson, "Recognition of Facial Expression in Three Cultures", Journal of Cross-Cultural Psychology, Vol. 11, No. 3, pp.263-280, 1980

[9] Y. Taniguchi, A. Akutsu, T. Tonomura, "PanoramaExcerpts: Extracting and Packing Panoramas for Video Browsing", Proc. of the 5th ACM Multimedia, pp. 427 – 436, Seattle, USA, 1997

[10] S. Ando, A. Suzuki, Y. Takahashi, T. Yasuno, "A Fast Object Detection and Recognition Algorithm Based on Joint Probabilistic ISC" , Proc. of MIRU2007, Japan, 2007 (Japanese)

[11] J.M. Lange, H.-M. Voigt, D. Wolf, "Growing Artificial Neural Networks Based on Correlation Measures, Task Decomposition and Local Attention Neurons", Proc. IEEE Conference on Neural Networks Vol. 2, pp.1355-1358, USA, 1994

[12] Y.-F. Ma, L. Lu, H.-J. Zhang, M. Li, "A User Attention Model for Video Summarization", Proc. of the 10th ACM Multimedia pp. 533-542, France, 2002

[13] B. Wild, M. Erb, M. Eyb, M. Bartels, W. Grodd, "Why are smiles contagious? An fMRI study of the interaction between perception of facial affect and facial movements", Psychiatry Research: Neuroimaging, Vol. 123, Issue 1, pp. 17-36, May, 2003