Classification of Spectators' State in Video Sequences by Voting of Facial Expressions and Face Directions

Tetsu Matsukawa, Akinori Hidaka and Takio Kurita

University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki, 305-8568, Japan
E-mail: {t.matsukawa, hidaka.akinori, takio-kurita}@aist.go.jp

Abstract

In this paper, we proposed a classification method of spectators' state in video sequences by voting of facial expressions and face directions. The task of this paper is to classify the state of the spectators in a given video sequence into "Positive Scene" or "Negative Scene", and "Watching Seriously" or "Not Watching Seriously". The proposed classifier is designed by a "bag-of-visual-words" approach based on face recognitions. First, the multiview (left-profile, front, right-profile) faces are detected from each image in the given video sequence. Then the detected faces are classified into the two expressions, smile or not smile. The classification results of the face directions and the facial expressions are voted to each classes' histogram over the video sequence. Finally, the state of the spectators is classified by using the kernel SVM on the voted histograms. We conducted experiments using spectators' video sequences captured from TV. Our approach demonstrated promising results for classifications of "Positive Scene" and "Negative Scene" or "Watching Seriously" and "Not Watching Seriously". It was also ascertained that the facial expression is important in the classification of "Positive" and "Negative". On the other hand, face direction is important to classify whether the spectators are "Watching Seriously" or "Not".

1 Introduction

In the entertainment industry which treats sports or comedy show, understanding whether their spectators have been satisfied or not is important to evaluate the quality of their services. Currently, questionnaire survey has been used to evaluate the degree of their satisfaction. But great cares of cost and time are required for questionnaire survey. So, it is desired to automatically evaluate spectators' satisfaction degree from video sequences. To realize automatic evaluation of spectators' satisfaction, in this paper we focused on facial information.

Currently, many face detection methods have been proposed by many researchers[1,2,3,4]. Viola and Jones proposed a boosted cascade framework and apply an integral image concept for face detection [1]. Their method enabled us to detect faces with fast speed and high precisions. In a natural scene, we need to recognize various directions of faces. Huang et al. proposed a rotation invariant multivew face detection (MVFD) method by extending Viola and Jones approach with a novel Width-First-Search tree structure and sparse features in granular space[2]. Many methods for automatic facial expression recognition were also proposed by many researchers[5,6,7]. Shinohara et al. proposed a facial expression recognition method with optimal local feature's weight maps using fisher discriminant criterion[5]. Hu et al. evaluated several local features and dimension reduction methods for multivew facial expression recognition[6]. Chen et al. proposed a combined method of face detection and facial expression recognition using selected Harr-like and Gabor-features[7].

Although many recognition methods have been existed for both face detection and facial expression, such existing facial recognition researches were focused on the individual's face and not interested in the integration of face recognition results of multiple peoples. For the purpose of understanding the overall satisfaction of spectators, integration of recognition results of many people's faces are needed.

In this paper, we propose to vote the recognition results of the facial expressions and the face directions to estimate the spectators' state in the given video sequence. The overview of proposed method is shown in Figure 2. At first the multiview (left-profile, front, right-profile) faces are detected from each image in the video sequence. Then the detected faces are classified into the two classes, smile or not smile. The classification results of face directions and facial expressions are voted as histograms over the video sequence. Finally the state of the spectators is classified by using the kernel SVM on the voted histograms. Recently, "bag-of-visual-words" approach which uses the histogram of quantized local features has been proposed in generic object recognition problems [11][10]. Our approach can be regarded as a face classifier based "bag-of-visual-words" which uses facial expressions and face directions as visual-words. The details of the proposed method are described in section 3. The advantages of our method are not only having no assumption of camera settings, but also utilizing the distributions of face directions for state classifications.

There are some researches of estimating of single user's preference [13, 14, 15], but there are some problems to apply for spectators' state recognition. Kakusho at el. proposed "facial expression mapping (FEM)" to transmit facial expressions of the user with

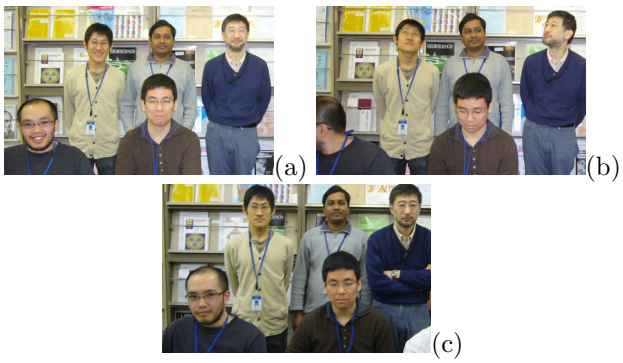

Figure 1: Example of spectators' scene. (a): "Positive Scene", (b): "Negative Scene" - "Not Watching Seriously", (c): "Negative Scene" - "Watching Seriously". These are for explanation and not the actual Olympic scenes used in this paper.

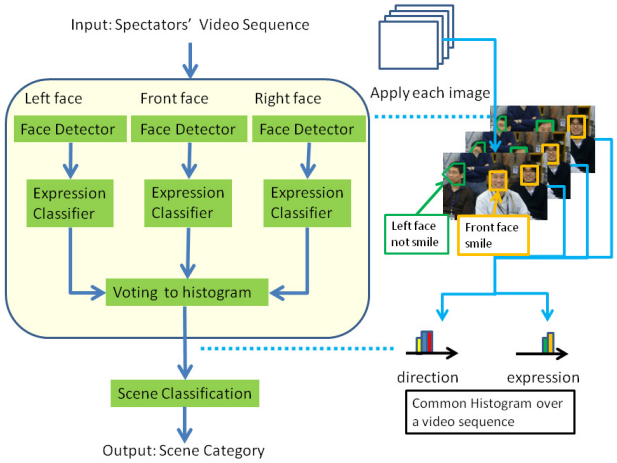

Figure 2: Overview of proposed method.

the face of embodied agent in agent-mediated distance communication [13] by learning the user's preferences in FEM using a RBF network. They used LED markers for clipping face regions, but it is a burden for many users. Aiming at the development of a system for automatically acquiring personal preferences from TV view's behaviors, Yamamoto at el. proposed a method for automatically estimating TV viewer's intervals of interest based on temporal patterns in facial changes with Hidden Markov Models[14]. Miyahara at el. proposed a tagging system for video contents based on a viewer's face expression using Elastic Bunch Graph Matching and Support Vector Machine. These systems have an assumption of using user's front face captured by a camera, but this assumption is invalid in many spectators' face direction in natural camera settings.

2 Classification Task

In this paper, we consider two binary classification tasks for estimation of spectators' state in the given video sequences. Assume the spectators are watching a sports game. One task is to classify the spectators' state into "Positive Scene" and "Negative Scene". "Positive Scene" is the scene where the spectators' encouraging team is leading the game and "Negative Scene" is not leading. The other task is to classify the spectators' state into "Watching Seriously" and "Not Watching Seriously" from the "Negative Scene". We prepared video sequences from TV programs of the Beijing 2008 Olympic Games. "Positive Scene"(50 samples) were collected from the scenes in which the players of the spectators' country was wining or leading the game. "Negative Scene"(50 samples) were collected from the scenes in which the game was not beginning or the spectators' country was not leading. "Watching Seriously"(25 samples) were also collected from the scenes in which the spectators' country was not leading at the game. "Not Watching Seriously" (25 samples) were collected from the scenes in which the game was not beginning. Our final goal is to classify spectators' state into 3 states, "Positive Scene" , "Negative Scene - Watching Seriously" and "Negative Scene - Not Watching Seriously". Combining the two classification problems, this goal will be achieved. So, we didn't include the "Positive Scene" into "Watching Seriously" vs "Not Watching Seriously". The number of persons in each image is different in each scene. One frame contains from about 3 to 30 persons. The number of frames in each sequence is 30 frames at video rate. Figure 1 shows the examples of the target scenes.

3 Scene Classification Method

3.1 Multiview Face Detection

To extract facial direction and expression information, we use multivew (left-profile, front, right-profile) face detectors and view dependent facial expression classifiers. A face detector is a combination of Viola-Jones' method (VJ-detector)[1] and a Support Vector Machine[9] based nonface-filter. SVM based nonface-filter uses different features from VJ-detector to reduce false positive and is used to confirm the detected regions by VJ-detector. The image features and the type of kernel of SVM based nonface-filter are the same as the classifier for facial expression recognition described in 3.2. In the case that several different view faces are detected in close position, we select the face view that has maximum classification score of SVM based nonface-filter. Since the side faces are often detected in the position that contains the half of front face, we trained the nonface-filter using such displacement faces as negative samples to eliminate such displacement faces.

3.2 Facial Expression Recognition

Facial expression classifiers classifies a face region into "smile" or "not". They are prepared for each face direction. Since the detected faces by VJ-detector are not correctly aligned at its correct position, the features should be invariant to such small shift of the detected face regions. As shift-invariant features, we use the histogram of intensity (HI) in the local 5×5 pixels. The face regions detected by VJ-detector are resized to 50×50 pixels. Then 50×50 pixels face regions are divided into 10×10 non-overlap cells by regarding 5×5 pixels $= 1$ cell. Intensity is quantized to 8 levels from 256. The dimension of HI feature is 10×10 (cells) $\times$

8 (intensity-levels) = 800. We also attempted the Histogram of Oriented Gradient (HOG)[8] features which are known as the effective for human detection with the same cell size to the HI feature. But, the classification performance of the HI feature was better than the performance of HOG feature in the preliminary experiment. So, we use HI feature in the following experiments. As classifiers, we used χ^2 kernel SVM which indicates good performance to classification based on the histogram-based image features [9]. The decision function of kernel SVM is defined by

$$f(\mathbf{H}) = sgn(\sum_{i \in SV} \alpha_i y_i K(\mathbf{H}_i, \mathbf{H}) - b)\,,$$

where SV is a set of support vector, $K(\mathbf{H}_i, \mathbf{H})$ is the value of a kernel function for the training sample $\mathbf{H}_i$ and the test sample $\mathbf{H}$, y_i is the class label of $\mathbf{H}_i$ (+1 or -1), α_i is the learned weight of the training sample $\mathbf{H}_i$ and b is a learned weight of threshold parameter. χ^2 kernel $K_{\chi^2}(\mathbf{H}_X, \mathbf{H}_Y)$ for histogram $\mathbf{H}_X = \{H_X(1), .., H_X(N)\}$ and histogram $\mathbf{H}_Y = \{H_Y(1), .., H_Y(N)\}$ is defined as follows:

$$K_{\chi^2}(\mathbf{H}_X, \mathbf{H}_Y) = exp(-\frac{1}{2\sigma}\sum_{i=1}^{N}\frac{(H_X(i) - H_Y(i))^2}{H_X(i) + H_Y(i)})\,.$$

The kernel parameter σ and regularization parameter C of SVM [9] are determined by 5-fold cross validation.

3.3 Voting Method

Multiview face detection and facial expression recognition described in 3.1 and 3.2 are applied by each frame of an input video sequence. The recognition results of face directions and facial expressions over a video sequence are voted into common histograms. The flows of the voting methods are shown in figure 3. We create three types of histograms for facial expressions denoted by $\mathbf{H}^E$ (2 elements of smile or not smile), facial directions denoted by $\mathbf{H}^D$ (3 elements of left, front, right) and co-occurrence of expressions and directions denoted by $\mathbf{H}^{D\wedge E}$ (6 elements of left $\wedge$ smile, left $\wedge$ not smile, front $\wedge$ smile, front $\wedge$ not smile, right $\wedge$ smile, right $\wedge$ not smile). Each histogram is normalized by the number of faces so that the norm of the histogram becomes 1. The face directions are used to evaluate the degree of the concentration to the game of spectators. Since the camera setting is different from scene to scene, the order of the histogram $\mathbf{H}^D$ is replaced as follows. If the number of left faces is the highest, the order of the direction histogram is set to (right, left, front). If the number of frontal faces is the highest, the order of the direction histogram is set to (left, front, right). If the number of right faces is the highest, the order of the direction histogram is set to (front, right, left). The order of the co-occurrence histogram $\mathbf{H}^{D\wedge E}$ is also replaced by the same way. The effect of these replacements is shown in 4.3.

3.4 Scene Classification by SVM

To classify the spectators' state in a given video sequence, we use kernel SVM with multi-channel summation kernel. According to the classification tasks, it is expected that the importance of each of these three

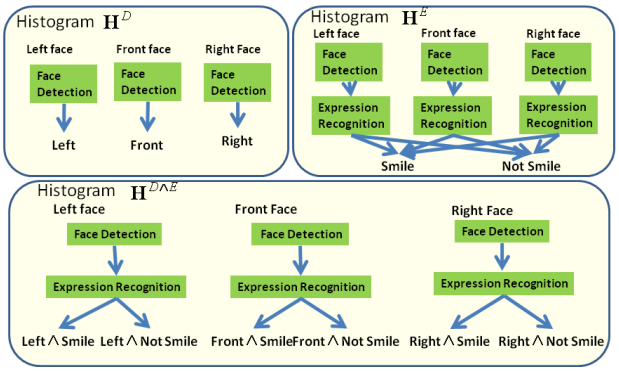

Figure 3: Voting method.

histograms is different. For example, facial expressions are probably the most important to classify "Positive Scene" and "Negative Scene". On the other hand, face directions may be important to classify whether the spectator is "Watching Seriously" or "Not". To control the importance of each histogram in the multi-channel summation kernel, we introduce weights for each histogram as follows:

$$\begin{aligned} K_B(\mathbf{H}_X, \mathbf{H}_Y) &= \alpha_D K_A(\mathbf{H}_X^D, \mathbf{H}_Y^D) \\ &\quad +(1-\alpha_D)K_A(\mathbf{H}_X^E, \mathbf{H}_Y^E) \\ &\quad where\ 0 \le \alpha_D \le 1\,. \end{aligned}$$

$K_A(\mathbf{H}_X, \mathbf{H}_Y)$ is a histogram intersection kernel defined by

$$K_A(\mathbf{H}_X, \mathbf{H}_Y) = \sum_{i=1}^{N} min(H_X(i), H_Y(i))\,,$$

where N denotes the dimension of the histogram.

We also consider the multi-channel summation kernel defined using the intersection histograms with all three histograms as

$$\begin{aligned} K_C(\mathbf{H}_X, \mathbf{H}_Y) &= \alpha_D K_A(\mathbf{H}_X^D, \mathbf{H}_Y^D) + \alpha_E K_A(\mathbf{H}_X^E, \mathbf{H}_Y^E) \\ &\quad +(1-(\alpha_D+\alpha_E))K_A(\mathbf{H}_X^{D\wedge E}, \mathbf{H}_Y^{D\wedge E}) \\ &\quad where\ 0 \le \alpha_D + \alpha_E \le 1\,. \end{aligned}$$

The histogram intersection kernel is Mercer kernel [11]. These kernels K_B and K_C are also Mercer kernels because the summation kernel of Mercer kernels with non negative weights are Mercer kernel [12].

In this paper the weights of each histogram α_D and α_E are automatically determined by using cross validation.

4 Experiment

4.1 Learning Classifiers

The scene classification method described in section 3 has been implemented on linux-2.6. VJ-detector for face detection was implemented using OpenCV. As primitive local features of VJ-detector, we used the extended version of Harr-like feature [3]. For implementing SVM, we used LIBSVM. Since the left face

Table 1: Classification performance of SVM(average of 5-fold cross validation on training data (Dataset-I)).

Direction	NonFace Filter	Expression(HI)
Front	95.62%	92.3%
Side	94.05%	93.5%

Table 2: Accuracy on classification scene (Dataset-II). FP(a): The incorporation rate of false expression, FP(b): The incorporation rate of non face

Classified Category	Accuracy	FP(a)	FP(b)
Front(Smile)	96.3%	2.0%	1.7%
Front(Not Smile)	99.4%	0.3%	0.3%
Left(Smile)	93.8%	3.3%	2.9%
Left(Not Smile)	98.8%	0.2%	1.0%
Right(Smile)	89.7%	4.7%	5.6%
Right(Not Smile)	98.5%	0.2%	1.3%

and right face are symmetric, we prepared only left face classifiers in all classifiers (VJ-detectors, nonface-filters, expression classifiers). The right face recognition was realized by horizontally flipping the input image.

The training face samples of SVM classifiers were collected from the dataset mentioned in section 2 and other video sequences captured from TV program by applying VJ-detectors. They are collected at 10 frame interval. The facial expressions of the training face samples were manually labeled as "smile" or "not smile".

For the expression recognition, 6460 samples (smile 2730, not smile 3730) for frontal faces and 6440 samples (smile 2270, not smile 4170) for side faces were used. For the front nonface-filter, 3000 face images and 3000 non face images were used. For the side nonface-filter, 2000 face images and 2000 non face images were used. The mis-detected regions by VJ-detectors were collected and used for non face training images. We denote these dataset by Dataset-I. The average classification rate evaluated by 5-fold cross validation of each classifiers shown in Table 1.

4.2 Facial Expression Recognition Result

The results of the facial expression recognition by using the constructed classifiers in the actual scenes (50 "Positive Scene" and 50 "Negative Scene" $\times$ 30 frame, we denoted this dataset by Dataset-II) are shown in Table 2. In Table 2, the accuracy means that correctly classified rate for each expression only on detected faces. The recognition accuracy of side faces was little worse than the front faces recognition. This is probably caused by the wide varieties of the side faces. The main reason of false recognition was occlusion by other object such as another person's head or arms and so on. To deal with such situation, local gaussian summation kernel may be better instead of using global gaussian χ^2 kernel [12] when the occluded areas are small. But there were largely occluded faces that cannot be classified by even humans' eyes. So, it may be desirable to reject occluded faces. Except for such occluded cases, the classification performances of expression were high. In next subsection, we show the classification results of the spectators' state.

Table 3: Error rates of scene classification (%), (a): "Positive Scene" vs "Negative Scene", (b): "Watching Seriously" vs "Not Watching Seriously", [ ]: error rate of non aligned direction.

kernel	histogram	(a)	(b)
Linear	$\mathbf{H}^D$	53 [44]	**28** [40]
Linear	$\mathbf{H}^E$	13 [-]	68 [-]
Linear	$\mathbf{H}^{D\wedge E}$	10 [13]	32 [32]
Linear	$\mathbf{H}^D$, $\mathbf{H}^E$	13 [14]	30 [40]
Linear	$\mathbf{H}^D$, $\mathbf{H}^E$, $\mathbf{H}^{D\wedge E}$	10 [13]	30 [32]
K_A	$\mathbf{H}^D$	40 [39]	32 [36]
K_A	$\mathbf{H}^E$	10 [-]	40 [-]
K_A	$\mathbf{H}^{D\wedge E}$	13 [16]	34 [30]
K_A	$\mathbf{H}^D$, $\mathbf{H}^E$	12 [12]	32 [36]
K_A	$\mathbf{H}^D$, $\mathbf{H}^E$, $\mathbf{H}^{D\wedge E}$	12 [13]	32 [34]
K_B	$\mathbf{H}^D$, $\mathbf{H}^E$	**9** [12]	30 [34]
K_C	$\mathbf{H}^D$, $\mathbf{H}^E$, $\mathbf{H}^{D\wedge E}$	**9** [11]	**28** [30]

4.3 Scene Classification Result

The spectators' state was estimated by using SVM with the multi-channel summation kernels and linear kernel. The error rates are calculated by leave-one-out method. The classification error rates in the best parameter of each method are shown in Table 3. By replacing the order of facial direction histogram, the classification performances were improved about 2$\sim$4% in many cases. In both classification tasks, the weighted histogram intersection kernel outperformed the unweighted histogram intersection kernel about 2 $\sim$3%. The 3 channel histograms are almost same performance to the 2 channel histograms.

Weights of linear SVM for each dimension of histograms are shown in Figure 4. By replacing the orders of bins of histograms $\mathbf{H}^D$ and $\mathbf{H}^{D\wedge E}$ as described in 3.3, the bin of "center direction $\wedge$ smile" becomes to support "Positive Scene" mostly. On the other hand, the bin of "right $\wedge$ smile" supported "Positive Scene" mostly when the order was not replaced. The bins of "center direction" and "center direction $\wedge$ not smile" become to support "Watching Seriously" mostly by reordering. In these ways, adequate weights are acquired by replacing the order and error rates became low.

The leave-one-out error curve corresponding to the α_D with same C parameter of SVM are shown in Figure 5. The parameter α_D was searched at 0.05 intervals. In the case "Positive" vs "Negative" classification, the best weight was $\alpha_D = 0.25$, and the error rates were increased as increasing α_D. This indicates that facial expression is more important for this classification task. On the other hand, in the case "Watching Seriously" vs "Not Watching Seriously" classification, the best weight was $\alpha_D = 0.55$, and the error rates were decreased as increasing α_D. This indicates that face directions are more important for this classification task. These observations are consistent with our intuition. The mis-classifications were caused on the scenes that were going off our approach. For example,

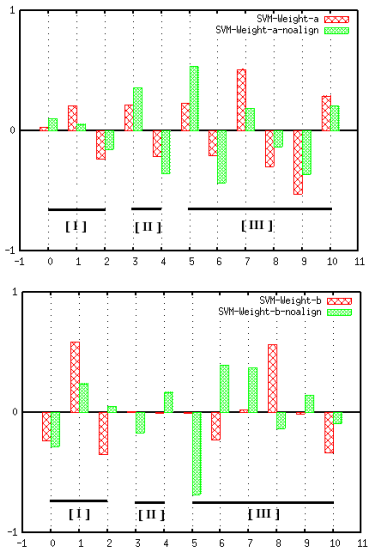

Figure 4: Weights of linear SVM for each dimension of histograms, $\mathbf{H}^D$, $\mathbf{H}^E$, $\mathbf{H}^{D\wedge E}$. [I]: $\mathbf{H}^D$ - the order is left, front, right, [II]: $\mathbf{H}^E$ - the order is smile, not smile [III]: $\mathbf{H}^{D\wedge E}$ - the order is left∧smile, left∧not smile, front∧smile, front∧not smile, right∧smile, right∧not smile, red bar: aligned direction, green bar: non aligned direction, (a): "Positive Scene" vs "Negatice Scene", (b): "Watching Seriously" vs "Not Watching Seriously".

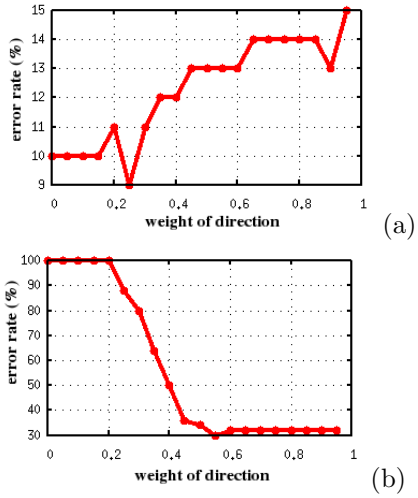

Figure 5: Error rates for K_B corresponding to direction weight α_D (leave-one-out). (a): "Positive Scene" vs "Negative Scene", (b): "Watching Seriously vs "Not Watching Seriously".

the scene that people are exciting but smiling people are few in "Positive Scene" or many people are different direction in "Watching Seriously" scene. Our method is easily able to add other visual-words to histograms, so there is a possibility to improve the performance in such situations.

5 Conclusion

In this paper, we proposed a spectators' state classification method by voting facial expression and direction. We conducted experiments using spectators' video sequences captured from TV. Our approach demonstrated promising results for classifying "Positive Scene" and "Negative Scene" or "Watching Seriously" and "Not Watching Seriously". It was also ascertained that facial expressions are important to classify "Positive Scene" and "Negative Scene" and face directions are important to classify "Watching Seriously" and "Not Watching Seriously".

Future works include to increase the number of classes of facial expressions or face directions and to increase the number of spectators' states.

Acknowledgment

This research was supported by a grant-in-aid for initiatives for attractive education in graduate schools.

References

[1] P.Viola and M.J.Jones, Robust Real Time Face Detection , IJCV 57(2), pp.147-154, 2004.

[2] C.Haung, H.Ai, Y.Li, S.Lao, High-Performance Rotation Invariant Multiview Face Detection, IEEE Trans. on PAMI, Vol.29, No.4, pp.671-686, 2007.

[3] R. Lienhart and J. Maydt, An Extended Set of Haar-like Features for Rapid Object Detection, In ICIP, Vol. 1, pp. 900-903, Sep. 2002.

[4] A.Hidaka and T.Kurita, Non-Neighboring Rectangular Feature Selection Using Particle Swarm Optimization, In ICPR, 2008.

[5] Y.Shinohara, N.Otsu, Facial Expression Recognition Using Fisher Weight Maps, In FG, 2004.

[6] Y.Hu, Z.Zeng, L.Yin, X.Wei, X.Zhou and T.S.Haung, Multi-View Facial Expression Recognition, In FG, 2008.

[7] H-Y.Chen, C-L.Haung and C-M.Fu, Hybrid-boost learning for multi-pose face detection and facial expression recognition, Pattern Recognition, 41, pp.1173-1185, 2008.

[8] N.Dalal and B.Triggs, Histograms of Oriented Gradients for Human Detection, In CVPR, pp.886-893, 2005.

[9] O.Chapelle, P.Haffner and V.Vapnik, SVMs for Histogram Based Image Classification, IEEE Trans. on NN, Volume 10, Issue 5, pp.1055-1064, 1999.

[10] S.Lazebnik, C.Schmid, and J.Ponece, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, In CVPR, pp.2169-2178, 2006.

[11] K.Grauman and T.Darrell, The Pyramid Match Kernel: Discriminative Classification with Set of Image Feaures, In ICCV, pp.1458-1465, 2005.

[12] K.Hotta, Robust face recognition under partial occlusion based on support vector machine with local Gaussian summation kernel, Image and Vision Computing, Vol.26, Issue 11, 1, pp.1490-1498, 2008.

[13] K.Kakusho, L.Li and M.Mihoh, Learning User's Preference in Controlling Facial Expressions of an Embodied Agent by the User's Face, In Proc. of IEEE International Workshop on Robots and Human Interactive Communication, pp.235-240, 2005.

[14] M.Yamamoto, N.Nitta and N.Babaguchi, Estimating Intervals of Interest During TV Viewing for for Automatic Personal Preference Acquisition, In Proc. of IEEE Pacific-Rim Conference on Multimedia(PCM2006), pp.615-623, 2006.

[15] M.Miyahara, M.Aoki, T.Takiguchi and Y.Ariki, Tagging Video Contents with Positive/Negative Interest Based on User's Facial Expression, MMM2008, pp.210-219, 2008.