Appearance-Based Smile Intensity Estimation by Cascaded Support Vector Machines

Keiji Shimada[1], Tetsu Matsukawa[2], Yoshihiro Noguchi[1] and Takio Kurita[3]

[1] Human Technology Research Institute, Advanced Industrial Science and Technology, Tsukuba, Japan
[2] Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan
[3] Faculty of Engineering, Hiroshima University, Higashi-Hiroshima, Japan

Abstract. Facial expression recognition is one of the most challenging research area in the image recognition field and has been studied actively for a long time. Especially, we think that smile is important facial expression to communicate well between human beings and also between human and machines. Therefore, if we can detect smile and also estimate its intensity at low calculation cost and high accuracy, it will raise the possibility of inviting many new applications in the future. In this paper, we focus on smile in facial expressions and study feature extraction methods to detect a smile and estimate its intensity only by facial appearance information (Facial parts detection, not required). We use Local Intensity Histogram (LIH), Center-Symmetric Local Binary Pattern (CS-LBP) or features concatenated LIH and CS-LBP to train Support Vector Machine (SVM) for smile detection. Moreover, we construct SVM smile detector as a cascaded structure both to keep the performance and reduce the calculation cost, and estimate the smile intensity by posterior probability. As a consequence, we achieved both low calculation cost and high performance with practical images and we also implemented the proposed methods to the PC demonstration system.

1 Introduction

The visual information plays a very important role in our everyday life. Especially, in regard to communication between human beings, we can come to understand deeply and smoothly each other to pay attention to behaviors and facial expressions as well as languages. Facial expression analysis has been approached by several research fields, for example in psychology[1], brain science, etc. In engineering[2] too, many researchers have tried to analyze and estimate facial expressions and human emotions by face images, by voice signals, by biosignals, etc. for a long time. But, it is still difficult to recognize facial expressions only by face images automatically, because there are many problems such as inconsistencies in individuals, lack of criterion to judge facial expressions, disparities between simulation data and practical data and a mismatch between the expressions and the emotions. Therefore, there is no critical solution to work well under the practical environment and active research is still much in progress.

In particular, smile (In a wide sense, facial expressions, which are observed when human beings derive pleasure) is one of the most important facial expression used to communicate well between human beings and also between human and machines. If we can automatically detect smile on real-time and at high accuracy, it will serve a useful function to existing applications like digital still camera and HMI (Human Machine Interface), and also raise the possibility of inviting new applications like rehabilitation and welfare in the near future. Furthermore, we think that such applications, which contain a camera should work on-the-fly, because of privacy issues.

In general, there are two major approaches to detect smile. One is feature-based method[3] and the other is appearance-based method[4]. Feature-based method has the robustness for the variation of face positions and angles, because it can normalize those and analyze more detailed information around facial parts. But it generally requires to find some facial parts such as eyes, mouth, etc. So, if it does not find those facial parts, it can't provide the result. On the other hand, appearance-based method does not need to find facial parts and can provides the result of smile detection without facial parts detection. As a result, although it is susceptible to the position of facial parts and the variation of face angle, it has low calculation cost.

In this paper, we study the method to detect smile and estimate its intensity using only facial appearance information on real-time and high performance, which is robust to the position gap of facial parts and face angle within approximately ±30 degrees of frontal. We have also implemented our proposed methods to on-the-fly PC demonstration system.

2 Smile Detection and Intensity Estimation

We try to detect smile and estimate smile intensity using 256 gray values, where the size of face is 40 x 40 pixels. That means it is not necessary to identify facial parts in our method. Fig. 1 shows the process flow of our smile detection and smile intensity estimation. In this paper, we study three feature extractions,

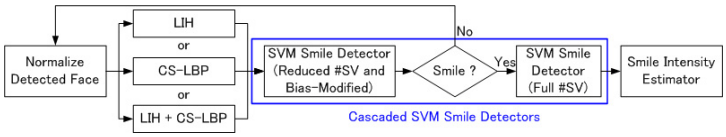

Fig. 1. Smile detection and intensity estimation flow

namely such as Local Intensity Histogram (LIH), Center-Symmetric Local Binary Pattern (CS-LBP)[5] and LIH+CS-LBP, which combines the above two features as facial appearance information. In addition, we consist SVM smile detector as cascaded structure like face detector proposed in [6]. It consists of sub detector, which has a small number of support vectors by applying Reduced

Set Method (RSM)[7] and main detector, which consists of all support vectors. This cascaded structure has the ability to keep high performance, while reducing the calculation cost.
At the end, we estimate smile intensity based on the posterior probability estimated by the output from SVM smile detector.

2.1 Feature Extraction

Generally, face detector does not insure the accuracy of the detected face positions, means that positions of facial parts such as eyes, mouth and etc., are not always corresponding for each detected face. Therefore, the robust features for facial parts positions and face angles are necessary to detect smile only by appearance information, accurately. In this paper, we divide the face image into some grid cells and extract local features for each cell, after that we build the final feature by concatenating all local features. We use Local Intensity Histogram (LIH) and Center-Symmetric Local Binary Pattern (CS-LBP) as local feature and describe how to extract those features in the following subsections.

Local Intensity Histogram (LIH) LIH is build by concatenating the intensity histograms in local regions and the extraction steps are as follows:

1. Divide face image into M x N cells
2. Build an intensity histogram with L bins for each cell
3. Normalize the histogram for each cell
4. Build the final feature by concatenating the normalized intensity histograms of all cells to form a (M x N x L) dimensional vector

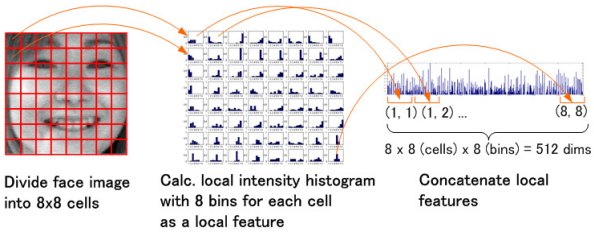

Fig. 2. Example of feature extraction by LIH (8 x 8 cells, 8 bins)

Fig. 2 shows the processing example, where face image is divided into 8 x 8 cells and 8 bins.

Center-Symmetric Local Binary Pattern (CS-LBP) CS-LBP is a simple method and it is also has the ability to extract features, which has robustness for

illumination changes. Additionally, it can also represent a texture information as more compact binary patterns. CS-LBP is calculated by,

$$CS-LBP_{R,N,T}(x,y) = \sum_{i=0}^{(N/2)-1} s(n_i - n_{i+(N/2)})2^i, \quad s(x) = \begin{cases} 1 & x > T \\ 0 & otherwise \end{cases} \quad (1)$$

where T is an encoding threshold, n_i and $n_{i+(N/2)}$ correspond to the gray values of center symmetric pairs of pixels of N equally spaced pixels on a circle of radius R. In this paper, N is fixed to 8 and R is fixed to 1. The following steps show the process to extract CS-LBP feature:

1. Divide face image into M x N cells
2. Calculate a CS-LBP for each cell and build a CS-LBP histogram
3. Buid the final feature by concatenating the CS-LBP histograms of all cells to form a (M x N x 16) dimensional vector.

Fig. 3 shows the processing example, where face image is divided into 5 x 5 cells.

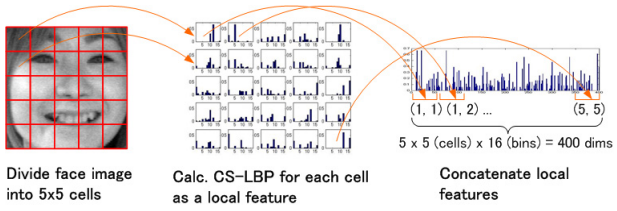

Fig. 3. Example of feature extraction process by CS-LBP (5 x 5 cells)

2.2 Detection and Intensity Estimation

In this paper, we use SVM with RBF kernel function. RBF kernel function is defined as,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma^2}) \quad (2)$$

where σ is a kernel parameter. Then the decision function with kernel is given as,

$$y = \sum_{i \in S} \alpha_i K(\mathbf{x}_i, \mathbf{x}) - h \quad (3)$$

where S and $\mathbf{x}_i$ means a set of support vectors and support vector, K is a kernel function, $\mathbf{x}$ is the input vector. Here, α_i shows weight for support vector and h represent the bias term. In smile detection case, If $y \geq 0$ implies smile and non-smile, otherwise.

In addition, we estimate smile intensity to evaluate the posterior probability of SVM outputs. Using sigmoid function, smile intensity is defined as,

$$si = \frac{1}{1 + \exp(-\lambda y)} \tag{4}$$

where si is smile intensity that ranges from 0 to 1 and λ is the gain. In this paper, we fix λ to 5.

Furthermore, we construct cascaded smile detectors described in Sec. 2. It consists of sub smile detector, which is reduced the number of support vectors constructed by RSM and main smile detector, which has all the set of support vectors.

3 Data Preparation

We constructed the original image database, which consisted of sports games TV programs to train and test our smile detector. In such TV programs, the spectators are often shooted and their expressions and emotions vary with the outcome of games. It is good for us, because the most spectators show same facial expressions with the outcome of their supporting athletes or teams. Therefore, we can collect smile and non-smile images with a high degree of efficiency. Moreover, face direction, gender and age of spectators are so various that we can evaluate the practical performance of our proposed method. As a result, our original smile/non-smile database contained 6,460 faces, which consisted of 2,730 smile samples and 3,730 non-smile samples.

In addition, we also used the pubic databases, "Facial Expression and Emotion Database (FEED)"[8], to evaluate the smile intensity estimation method.

4 Experiments

In this paper, we tested our system by 5-fold cross validation, which was one of several approaches commonly used for evaluation purpose. We compared each performance by Area Under the Curve (AUC), which was obtained by Receiver Operating Characteristics (ROC) Analysis.

4.1 Performance Evaluation by LIH

With respect to LIH, we investigated the optimal number of cells and bins. We first compared the performance according to the number of cells with the fixed number of bins, which is 8 (See Fig. 4 (Left)). Here, we used our original database and smile detector, which consisted of all support vectors (not cascaded). When increasing the number of cells, AUC was gradually improved. So, 8 x 8 cells showed the best performance. Next, we compared AUC by varying the number of bins, while keeping the number of cells constant as 8 x 8 (See Fig. 4 (Right)). 4 and 8 bins showed almost the same better performance, but 4 bins provided

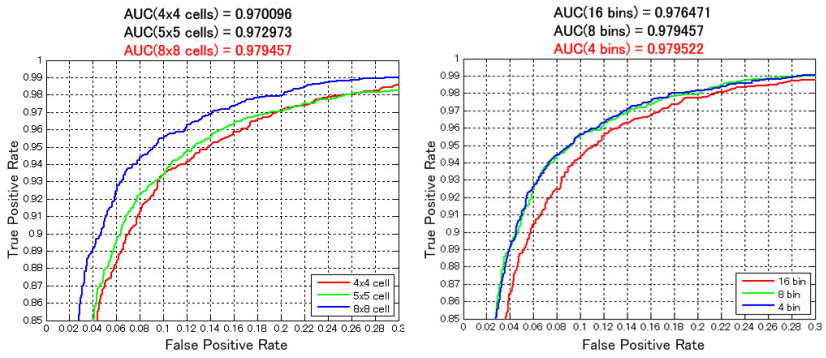

Fig. 4. Comparison of performance by LIH

the best. It means that we need just 4 gray values to detect smile in this experiment. As a result of these experiments, with respect to LIH, the optimal parameters were 8 x 8 cells and 4 bins (That is a 256 dimensional vector) and that performance provided 0.979522 for AUC.

4.2 Performance Evaluation by CS-LBP

With respect to CS-LBP, we investigated the optimal number of cells and the encoding threshold. We first compared the performance according to the number of cells with constant encoding threshold equal to 0.00 (See Fig. 5 (Left)). Here, we used our original database and smile detector, which consisted of all support vectors (not cascaded). As increasing the number of cells, AUC was higher, but

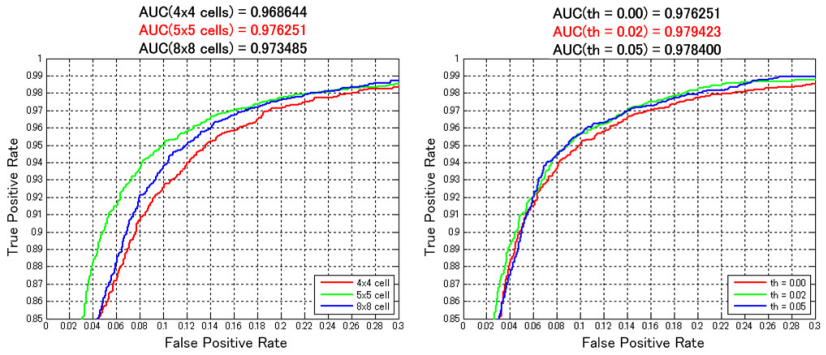

Fig. 5. Comparison of performance by CS-LBP

too match cause to degraded the performance. In this experiment, 5 x 5 cells gave the best. Next, we compared AUC according to the encoding threshold

with constant 5 x 5 cells (See Fig. 5 (Right)). Almost the same performances were shown, but in this experiment, the encoding threshold of 0.02 provided the best AUC. As a result of these experiments, with respect to CS-LBP, the optimal parameters were 5 x 5 cells and encoding threshold of 0.02 (That is a 400 dimensional vector) and it provided 0.979423 for AUC.

4.3 Performance Evaluation by LIH+CS-LBP

In this section, we describe the experiments with LIH+CS-LBP feature, which is combined LIH and CS-LBP. Here, the parameters of LIH were set to 8 x 8 cells and 4 bins and the parameters of CS-LBP were set to 5 x 5 cells and the encoding threshold of 0.02 from the above experimental results. LIH+CS-LBP, which was a 656 (= 256 (LIH) + 400 (CS-LBP)) dimensional vector, improved the performance. Here, it provided 0.982320 for AUC and it was better than using only LIH or CS-LBP (See Fig. 6).

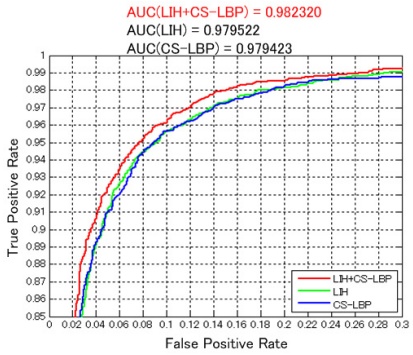

Fig. 6. Performance comparison of all the three features

4.4 Performance of Cascaded SVM Smile Detector

In this section, we describe the comparison of the performance and calculation cost, with our cascaded SVM smile detectors. Here, we used LIH+CS-LBP features and the parameters of LIH were set to 8 x 8 cells and 4 bins and the parameters of CS-LBP were set to 5 x 5 cells and the encoding threshold of 0.02 according to the results of Sec 4.1∼ 4.3. The number of support vectors of sub smile detector were reduced either to 32, 64, 128, 256, 512 or 1024 by RSM. And we also adjusted a bias term (h in Equ. 3) of sub smile detector lower to suppress the miss rejection cases at the sub smile detector. In this paper, we adjusted a bias to achieve True Positive Rate as well as main smile detector's one in advance. Fig. 7 shows the performance according to the several cascaded SVMs. When the number of support vectors of sub smile detector decreased,

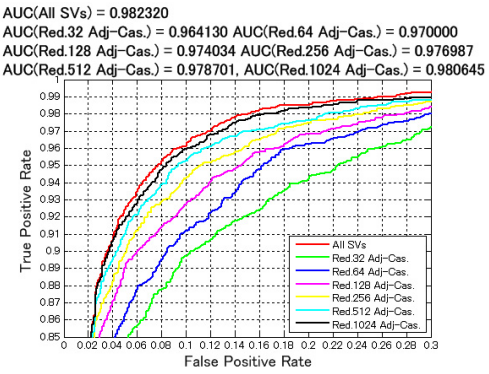

Fig. 7. Performance of cascaded SVMs (feature extraction is LIH+CS-LBP)

the performance degraded. But it kept to provide over 0.98 for AUC with 1024 support vectors in this experiment.

At the end of this section, we showed a comparison of AUC and calculation speed for each smile detector structure (See Table 1).

Table 1. Comparison of the number of SVs, AUC and calculation speed (Matlab@3.0GHz Core 2 Quad)

Classifier	#SVs	AUC (LIH+CS-LBP)	CPU Time (msec)
Normal SVM (with all SVs)	2481	0.982320	9.6418
Cascaded SVMs	32 (& 2481)	0.964130	3.9463
	64 (& 2481)	0.970000	4.0364
	128 (& 2481)	0.974034	4.1089
	256 (& 2481)	0.976987	4.6670
	512 (& 2481)	0.978701	5.4277
	1024 (& 2481)	0.980645	7.5097

on Matlab @ 3.0GHz Core 2 Quad

Cascaded SVMs with 1024 support vectors in sub smile detector achieved comparable in performance ($AUC > 0.98$) to normal SVM (non-cascaded), while reducing over 20% in the calculation cost. As a result of these experiments, our proposed cascaded SVM smile detectors could reduce calculation cost with a little performance degradation.

4.5 Smile Intensity Estimation

The FEED database is suitable to evaluate a change of a certain facial expression, because it has 100 $\sim$ 150 image sequences, which contain the variation from neutral face to a certain facial expression for each subject. Fig. 8 showed the results of our smile intensity estimator to "happy" expression of Subject #0001. This result proved that our smile intensity estimator could track the transition from neutral to smile well. Especially, it could represent the subtle facial expression changes as shown in red rectangle area in Fig. 8.

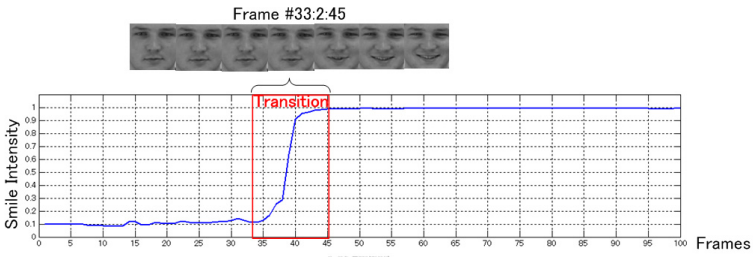

Fig. 8. Smile intensity estimation result of "happy" sequence #1 of Subject #0001

5 Demonstration System

In this section, we introduce the PC demonstration system, which was implemented according to the methods proposed in this paper. Fig. 9 (Top) shows the process flow. At first, we detect face from the input image by face detector and crop, scale and normalize the influence of illumination changes by the histogram equalization. After that, we extract LIH+CS-LBP features as facial appearance information and detect smile and estimate its intensity by cascaded SVMs. Fig. 9 (Bottom) shows examples of detection results by our demonstration system. Our demonstration system roughly spends $48ms$ for face detection

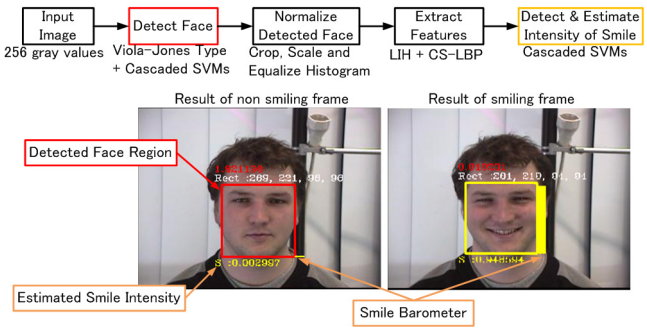

Fig. 9. Process flow of the PC demonstration system (Top) and examples of detection results (Bottom)

and $8ms$ for smile intensity estimation per face on the average (on a Core 2 Quad at 3GHz). It shows that our system can detect face and estimate its smile intensity from input image on semi real-time.

6 Conclusion

In this paper, we studied how to detect smile and estimate smile intensity only by facial appearance information, and proved the validity of our proposed system

through several experiments. We also built the semi real-time PC demonstration system, which was implemented according to our proposed smile intensity estimation methods.

We constructed original smile/non-smile practical database from sports games TV programs, which contained various spectators in both indoor and outdoor. We investigated the optimal parameters for LIH and CS-LBP to detect smile with the above database and compared the performance of smile detection using LIH, CS-LBP and LIH+CS-LBP. In our result, with respect to LIH, we achieved 0.979522 for AUC with 8 x 8 cells and 4 bins and for CS-LBP, achieved 0.979423 with 5 x 5 cells and the encoding threshold of 0.02. Combined feature, LIH+CS-LBP worked better among all the three features. That produced a AUC value of 0.982320 as the best performance. This result indicates that our proposed system is robust and works well even under the practical environment.

In addition, we constructed cascaded SVMs for smile detector, which was composed of a sub detector, consisted of small number of support vectors and a main detector consisted of all support vectors. As a result, we could keep AUC higher than 0.98, while delivering about 20% reduction in the calculation cost.

With respect to smile intensity estimation, we showed that our estimator could track the subtle expression changes from neutral to smile using the FEED database.

In the future, we plan to detect the other facial expressions and estimate those intensities based on the proposed methods described in this paper.

References

1. Ekman, P., Friesen, W.V.: Unmasking the face. A guide to recognizing emotions from facial cues. Prentice Hall (1975)
2. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. PATTERN RECOGNITION **36** (1999) 259–275
3. Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., Movellan, J.: Toward practical smile detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **31** (2009) 2106–2111
4. Deniz, O., Castrillon, M., Lorenzo, J., Anton, L., Bueno, G.: Smile detection for user interfaces. In: ISVC '08: Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II, Berlin, Heidelberg, Springer-Verlag (2008) 602–611
5. Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with local binary patterns. Pattern Recogn. **42** (2009) 425–436
6. Shimada, K., Noguchi, Y., Sasahara, H., Yamamoto, M., Tamegai, H.: Detection of driver 's face orientation for safety driving assistance. Transactions of JSAE **41** (2010) 775–780
7. Scholkopf, B., Burges, C.J.C., Smola, A.J.: Advances in Kernel Methods. The MIT Press (1998)
8. Wallhoff, F.: Facial expressions and emotion database, http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html (2006)