FACIAL EXPRESSION RECOGNITION USING HOUGH FOREST

Chi-Ting Hsu[1], *Shih-Chung Hsu*[1], *and Chung-Lin Huang*[1,2]
1. *Department of Electrical Engineering, National Tsing-Hua University, Hsin-Chu, Taiwan*
Email: s9961601@m99.nthu.edu.tw, d9761817@oz.nthu.edu.tw, clhuang@ee.nthu.edu.tw
2. *Department of Applied Informatics and Multimedia, Asia University, Taichung, Taiwan*

***Abstract*—This paper introduces a new facial expression recognition system. Facial expressions analysis encounters two major problems: non-rigid morphing (human facial expression are non-rigid and shape deformation) and person-specific appearance (the facial action features are people-dependent). Our facial expression system analyzes the non-rigid morphing facial expressions and eliminates the person-specific effects through patch features extracted from facial motion due to different facial expressions. Finally, classification and localization of the center of the facial expression in the video sequences are performed by using a Hough forest.**

I. INTRODUCTION

In recent years, the human-machine interface (HMI) applications have been widely applied in many smart devices. Human emotion understanding for friendly HMI is an indispensable and challenging research topic. Human face contains a lot of information and it varies from person to person. Mehrabian[1] indicates that to identify the mood of a human being, the proportion of the language (conversation, words of significance) accounts for 7%, the auditory message (intonation, sound size) accounts for 38%, and the visual message (facial expressions, body movements) accounts for 55%. To know the mood of a person, over half of the message is from the visual observations, and the facial expression is one of the most important information.

The previous works of automatic facial expression recognition can be categorized into two main categories[2]: image-based methods and video-based methods. However, a natural facial expression is dynamic, which evolves over time from the onset, the apex, and the offset. Onset is defined as the time from the start of the expressive episode to a peak of intensity. Apex is the amount of time the expression is held at the peak, and offset time is the time from the first evidence of fading of the expression until it stops fading. The image-based methods take only one shot to capture the image characteristics at the apex of the expressions. They ignore such dynamic feature, so they cannot perform well in most real world settings. Facial expressions analysis encounters two issues: non-rigid morphing (human faces are non-rigid and shape deformation under expression) and person-specific appearance (the locations of the facial features are absolutely not constant for different people).

Existing approaches to facial expression analysis can also be divided into geometric-based and appearance-based approaches. Appearance-based method may detect the edge gradient, or generate the gradient distribution of the objects based on histogram of gradient orientation (*HOG*) [4]. Another description of area texture variation resistant to brightness variations is Local Binary Patterns (*LBP*) [5]. There are other features such as Haar-like [7], Gabor Wavelet [9], for facial expression recognition. The advantage of such features is user discriminative. It is well captured because of facial expressions and facial texture changes.

Geometric-based Features using a number of points to describe the contours of the face information. Usually, facial feature points (both ends of the eyebrows, around the eyes, around the mouth etc.) are regarded as the key points. The variations between the key points can be used for the expression recognition. To achieve the automated expression recognition, most researchers use Active Appearance Model, AAM [8] or active shape model (ASM)[9] to detect feature point and trace these characteristic points to record the displacement of the point. Because it finds the relative position between the points as the features, it is not influenced by the person-specific appearance. The disadvantage is that the location of these feature points is not precise for tracking because of head displacement.

Ekman [10] suggested that all emotions belong to a rather small set of categories. These “basic” emotions (anger, disgust, fear, happiness, sadness, and surprise) are expressed by the same facial movements across different cultures, and therefore represent an appealing choice when designing automatic methods for facial expression classification. Pantic *et al*.[11] conclude that facial expression recognition method is divided into three categories: template-based, and rule-based, learning-based. They generally apply the learning-basis of expression recognition system, and the support vector machine SVM [12] is widely used in [4, 5, 6].

This paper presents a video-based method for the classification of facial expressions into one of the basic emotion labels. We capture the facial images through the camera, analyze the non-rigid morphing facial expressions, and eliminate the person-specific effects through patch features extracted from facial motion. Finally, classification and localization of the center of the expression in the video sequences are performed by using a Hough transform voting method and random forests.

In the preprocessing, the face, nose and two eyes are located for the following face alignment process. Then, we may estimate the motion field due to facial expression changes between consecutive frames. Third, the patches are sampled from the video sequence and the motion in patch is

summed. Finally, a combined random forests and a Hough voting are used to classify the sequence in terms of expression and localization of the center of the expression in the video sequences. Different from the previous Hough forest [17, 18], we apply the Hough voting for expression recognition and propose the ROI filtering to accumulate more effective votes and locate the accurate facial expression in temporal axis.

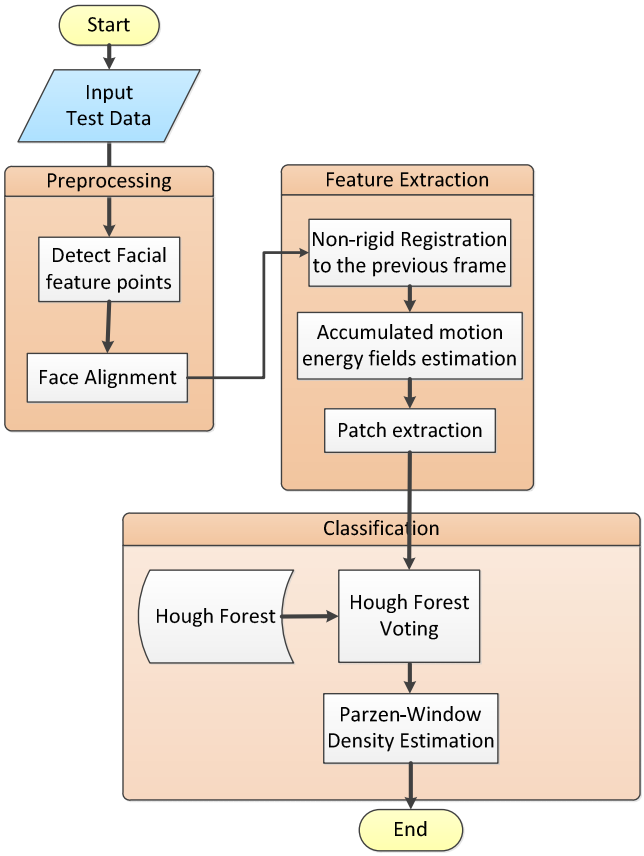

Figure 1 System flow diagram

II. Feature Extraction

Based on previous study [13], we find that facial expression recognition based-on feature extractions encounters two challenges: non-rigid morphing (human faces are non-rigid and undergo shape deformation during facial expression) and person-specific emotional face appearance as shown in Figure 2.

Figure 2. (a) Non-rigid morphing, (b) person-specific emotional appearance effects.

To increase expression recognition accuracy, we analyze the non-rigid morphing facial expressions and eliminate the person-specific effects. The facial expression is similar to certain kind of human action. By applying similar human action analysis, we may eliminate the person-specific emotional appearance. When people want to make a specific action, the muscles will be moving towards a particular direction. Facial muscles will do the similar way to make certain kind of expression. Although everyone looks different, the facial muscle will make the similar movement for facial expression as shown in Figure 3. Our expression recognition system analyzes facial expression by analyzing the motion features extracted from a sequence of frames.

Figure 3. Basic emotions motion direction

A. *Preprocessing*

The preprocessing consists of face detection, eye detection, nose detection, face alignment, tracking and motion field extraction which are shown in Figure 4. The face, nose and two eyes on each frame are located by face detection and tracking module [3]. This area is then in-plane rotated so that the face will have an up-right pose.

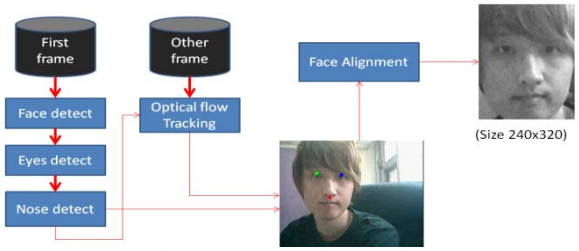

Figure 4. preprocessing overview

After face detection, we may find the feature points of the face (*i.e.*, eyes, nose). We align the faces based on the locations of the feature points. Face are rotated and scaled so that the eyes lie on the same horizontal line and have the same inter-ocular distance. The advantage is to ensure that the faces are fully aligned and located at the same position. The original and the aligned faces are shown in Figure 5.

Figure 5. (a) the original face; (b) aligned face; (c) tilt face.

Optical flow [14] shows the motion field of apparent motion of objects based on the intensity variations of surfaces and edges in a visual scene caused by the relative motion between an observer and the scene. Optical flow techniques such as Lucas-Kanade method assumes: (1) Image brightness in a small region remain the same although their location may change, (2) The image motion of a surface patch changes gradually over time, and (3) Neighboring points in the patch have similar motions. The Lucas-Kanade method assumes that the displacement of the image contents between two nearby instants (frames) is small.

B. Motion Extraction

After preprocessing of each video sequence, we analyze the motion field which occurs due to facial expressions between consecutive frames. We use optical flow method and B-splines interpolation to analyze the motion feature. B-splines curve is a widely used in the application of computer-aided parametric curves and displacement analysis on a large number of excellent local control capacities. It also has been applied in MRI medical imaging [15]. Here, we use optical flow algorithm and find the control points. Then, we interpolate the motion field which into three different kinds of usages: the magnitude, the vertical direction, the horizontal direction as shown in Figure 6.

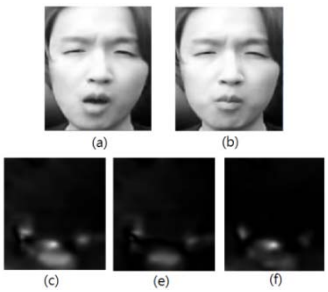

Figure 6. Facial motion feature. (a) (b) two faces; (c) motion magnitude; (e) motion *x* direction; (f) motion *y* direction

Since the amount of motion between consecutive frames is usually small and may not provide enough information for expressions detection, the motion field is accumulated in temporal as

$$I_i^{mag}(x,y,t_i) = \sum_{t=ti-T}^{t=ti+T} \sqrt{u(x,y,t)^2 + v(x,y,t)^2} \qquad (1)$$

where $u(x, y, t)$ and $v(x, y, t)$ are the two motion vector components of pixel located at (x, y) in the t^{th} image frame, and T is the observation window temporal duration. Figure 7 shows the accumulated motion features which include the motion magnitude, the motion over time in the horizontal direction, and the motion over time in the vertical direction.

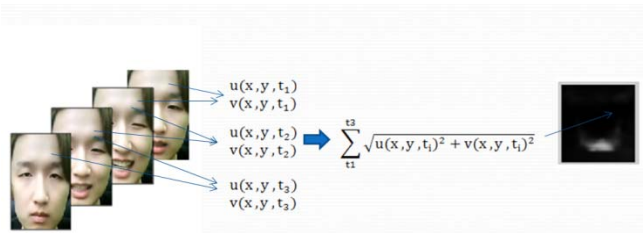

Figure 7. Accumulated motion feature.

C. Patch sample

To eliminate the person-specific effects, we extract the 3D video patch p which contains the motion features of facial expression. The 3D video patches are sampled from the video sequence as shown in Figure 8.

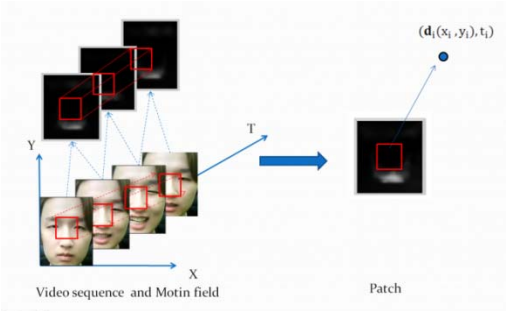

Figure 8. Patch samples.

III CLASSIFICATION

The retrieved video patches can be used to identify the facial expression. However, the number of patches is so large that matching these patches for recognizing the facial expression is very time consuming. To speed up retrieval, Hough forest [17, 18] is proposed which has recently been extended to handle multi-class detection in the spatio-temporal domain and applied to the task of action recognition.

A. Random Forest.

Random forest [16] is an ensemble classifier that consists of many decision trees. Decision trees are commonly used in operation research, specifically in decision analysis, to help recognizing a strategy most likely to reach a goal. If the decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptor for calculating conditional probabilities. The randomization consists of random selection of training data and random test selection of the division hypothesis at each non-leaf node.

Random forest training involves tree construction and a best binary hypothesis of the non-leaf node that divides the training data into two subsets. After training, a binary test is

assigned to each non-lead node. In the testing, a test sample passes the non-leaf nodes of tree and reaches a leaf node. The category of the test sample is calculated by averaging the class probability distributions of the training samples pre-stored at the designated leaf node.

Generalized Hough transform has been used to find the imperfect instances of objects within a certain class of shapes by a voting procedure. This voting procedure is carried out in a parameter space, from which object candidates are obtained as local maxima in a so-called accumulator space that is explicitly constructed by the algorithm for computing the Hough transform. The randomized trees are used to learn the mapping between a 3D video patch and its vote in 4D Hough space to obtain the class label and spatio-temporal location of an action in the sense of generalized Hough transform.

The observation in face expression recognition system is the 3D video patch. The patch parameters are used for mapping between the input patch and its *n*-nearest neighbors with label. The facial expression of the input patch is classified by weighting these neighbors, and the temporal location of the facial expression can be obtained by Hough voting process based on its neighbors in temporal space. Finally, the facial expression can be recognized by the labeled bucket having the largest number of accumulated votes of which the temporal location indicates the apex point of the facial expression.

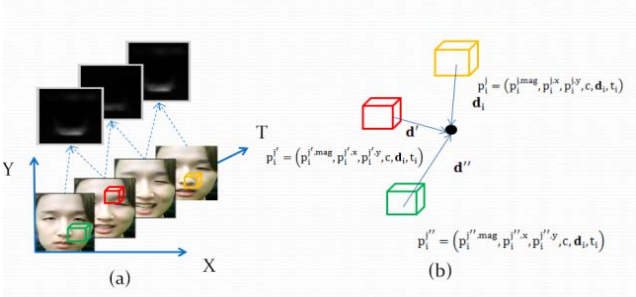

Figure 9. (a) Patches and (b) Hough voting based on patch parameters.

In the training, 3D video patches are semi-randomly selected from the training videos with parameters. The sampled patches will be located at the discriminant portion on the face such as the eyes and mouth. For instance, the patches on the nose are less discriminant than the patches on the other places because the motion field in the nose region has little facial expression related information. So, we only sample the video the patches around discriminant face region. .

B. Hough Tree

Hough tree T_n in Hough forest $\boldsymbol{T}=\{T_n\}$ is constructed from a set of 3D video patches $\{p_i\}$ defined as

$$p_i = \left(p_i^{mag}, p_i^x, p_i^y, c, \boldsymbol{d}_i, t_i\right)$$

where p_i is a randomly selected 3D video patch (20×20×3), p_i^{mag}, p_i^x, p_i^y is the motion feature of the patch, c is the expression label, $\boldsymbol{d}_i$ is 2D spatial displacement from the video patch center to the facial expression center of I_i, t_i is temporal displacement from the video patch to the video sequence center.

Each leaf node L stores a portion of the training patches. p_c^L denotes the proportion of patches per class label reaching the leaf after the training, *i.e.* $\sum_c p_c^L = 1$, and $D_c^L = \{\boldsymbol{d}_i, t_i\}_{c\in C}$ denotes a set of the training patches' spatial and temporal displacement vectors for class label c respectively. Each non-leaf node B of a tree is assigned a binary hypothesis test during training. There are many binary tests to be randomly selected. The binary hypothesis h at non-leaf node B is defined as

$$h_{B,a,\boldsymbol{p},\boldsymbol{q}}(p_i) = \begin{cases} 0 & \text{if } p_i^a(\boldsymbol{p}) < p_i^a(\boldsymbol{q}) \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

where p_i is the 3D video patch, B represents this non-leaf node ID, a is the decision attribute of the patch to be selected by this node (*e.g., mag, x,* or *y*), $\boldsymbol{p}$ and $\boldsymbol{q}$ are any two points in the 3-D video patch.

The random trees in Hough forests are constructed based on a standard random forest framework [17, 18]. Construction begins at the root by choosing a binary test, splitting the training patches according to the test results and then constructing children nodes. At each subsequent child node, the same procedure continues recursively, with each node being designated as a non-leaf node until the termination criteria is met, *i.e.* the child node is of a maximum depth, or there are less than a minimum number of patches remaining. Upon termination as a leaf, the remaining patches information, $D_c^L = \{\boldsymbol{d}_i, t_i\}_{c\in C}$ and p_c^L are stored.

The ideal binary test will split the patches so that the uncertainties of the class label and temporal center offsets are minimized. Here, we develop two measures to evaluate the uncertainty for a set of patches $A = \{p_i\}$. The first measure aims to minimize the class uncertainty:

$$U_1 = -1|A| \cdot \sum_c P_c \ln(P_c) \quad (3)$$

where $|A|$ is the number of patches in set A and P_c is the proportion of patches with label c in set A. Note that the summation expression is the standard definition of entropy for the class labels. The second measure aims to minimize the center offset uncertainty:

$$U_2 = \sum_i \|t_i - \bar{t}_A\|^2 \quad (4)$$

where $\bar{t}_A$ is the mean temporal offset of set A. Note that the offset uncertainty is minimized for all classes at the same time.

At each non-leaf node during training, a pool of binary test $\{h^B\}$ is generated with randomly selected values of a, $\boldsymbol{p}$, and $\boldsymbol{q}$ falling within the constraints of training data. Then, the test will be selected if either class or offset uncertainty is minimized. The set of patches arriving at the non-leaf node

will be evaluated by all binary hypotheses in the pool and the binary hypothesis satisfying the following minimization objective will be chosen as

$$Argmin_k \ (\ U_*(\{A \mid h^B = 0\} \ + \ U_*(\{A \mid h^B = 1\}) \ ) \quad (5)$$

where subscript * indicates the chosen uncertainty measure for the node. We randomly selecting the uncertainty measure for interleaved nodes throughout the tree with decreasing both class and offset uncertainty.

In training, we decide which measurement methods of the current node to use. After randomly generated binary decisions, we choose a group of the best binary decision and node data clustered by eq. (2) to generate two sub-nodes. The tree construction terminates under the following two conditions: (a) the depth of sub-node depth is deep enough, and (b) the number of patches in each cluster is not enough. If you meet one of the child nodes for leaf and store this information, otherwise continue to divide.

IV. Expression Recognition

To recognize facial expression, we extract the patches from the test video. These patches will be tested by each decision tree in Hough forest. Each patch ends up a certain leaf node of the tree in which the pre-stored patches will be used to cast votes for certain expression class and certain temporal center. Each patch will cast one vote to the designate bucket based on its corresponding expression class label c, spatial/temporal displacements $\mathbf{d}$ and t. Based on the votes in the accumulators in the temporal space for different classes, we can recognize the correct facial expression. To begin with, we consider an input patch $p_i(\mathbf{y})$located at $\mathbf{y}\in R^3$, with motion features $f_i(\mathbf{y}) = (p_i^{mag}, p_i^x, p_i^y)$. $c(\mathbf{y})$ is the patch's unknown class label, $\boldsymbol{d}_i$ is the location of the patch in the face, $t(c, \mathbf{y})$ is the displacement of the patch at $\mathbf{y}$ from the unknown sequence center for class c. Let $Q_c(t)$ be the random event corresponding to the possible existence of facial expression with labeled class c and centered at t. We are interested in finding the conditional probability $P(Q_c(t) \mid p_i(\mathbf{y}))$, which can be decomposed as follows:

$$\begin{aligned} &P(Q_c(t) \mid p_i(\mathbf{y})) = \\ &\sum_{l \in C} P(Q_c(t) \mid c(\mathbf{y}) = l, f_i(\mathbf{y})) \cdot P(c(\mathbf{y}) = l \mid f_i(\mathbf{y})) \\ &= P(Q_c(t) \mid c(\mathbf{y}) = c, f_i(\mathbf{y})) \cdot P(c(\mathbf{y}) = c \mid f_i(\mathbf{y})) \\ &= P(t(c, \mathbf{y}) \mid c(\mathbf{y}) = c, f_i(\mathbf{y})) \cdot P(c(\mathbf{y}) = c, f_i(\mathbf{y})) \end{aligned} \quad (6)$$

Suppose that the patch ends up in leaf L of tree T. The first factor can then be approximated as the Parzen-window estimate of D_c^L, the offset vectors belonging to class c, while the second factor can be approximated as p_c^L indicating the probability of the patch belonging to class c. We can then rewrite Equation (6) for tree T as

$$P\left(Q_c(t) \middle| p_i(\mathbf{y}), \boldsymbol{T}\right) = \left(\frac{1}{|D_c^L|} \sum_{t_i \in D_c^L} G((t' - t) - t_i) \right) \cdot p_c^L \quad (7)$$

where $G(\cdot)$ is the *1-D* Gaussian Parzen window function. For the entire Hough forest $\boldsymbol{T}$, we average over all the trees as

$$P(Q_c(t) \mid p_i(\mathbf{y}), \boldsymbol{T}) = \frac{1}{|\boldsymbol{T}|} \sum_n P(Q_c(t) \mid p_i(\mathbf{y}), T_n) \quad (8)$$

where $\boldsymbol{T}=\{T_n\}$ and the $|\boldsymbol{T}|$ indicates the number of trees. Equations (7) and (8) define the probabilistic vote of a single patch $p_i(\mathbf{y})$ for facial expression class c. Votes from all the patches selected from the discriminant regions at frame t are integrated into Hough accumulators in the temporal axis for different classes as

$$V(t, c) = \sum_{\mathbf{y} \in S(t)} P\left(Q_c(t) \middle| p_i(\mathbf{y}), \boldsymbol{T}\right) \quad (9)$$

where $S(t)$ denotes the discriminative region at frame t.

A. ROI filtering

Before voting, we propose ROI filtering to invalidate the votes that point to the incorrect bucket. When the input patch $p(\mathbf{y})$ ends up in the leaf node which provides a set of pre-stored labeled patches with $D_c^L = \{t_i, \mathbf{d}_i\}$. Each patch in the leaf node may cast a vote or not based on its $\mathbf{d}_i$. If the $\mathbf{d}_i$ is not in the ROI, then the vote is invalid. ROI filtering increases the validity of the final accumulated votes, and concentrates the valid votes in the target bucket in the temporal axis. ROI region for each input patch is determined experimentally as shown in Figure 10.

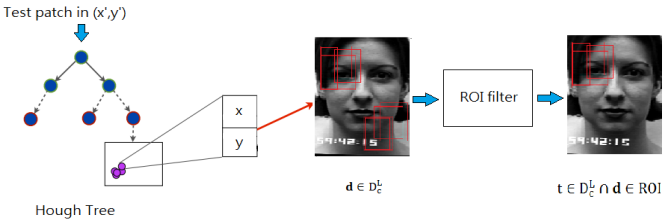

Figure 10. ROI filter

Then, we can rewrite Equation (7) for ROI filter as:

$$P\left(Q_c(t) \middle| p(\mathbf{y}), \boldsymbol{T}\right) = \frac{1}{|D_c^L|} \sum_{t_i \in D_c^L \cap \mathbf{d} \in ROI} G((t' - t) - t_i) \cdot p_c^L \quad (10)$$

The accumulated votes in buckets in the temporal space for six different facial expressions before and after applying ROI filtering are shown in Figure 11.

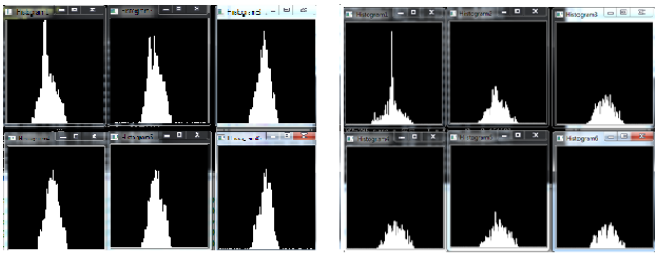

Figure 11. The accumulated votes in the buckets for six facial expressions: (a) before applying the ROI filtering, and (b) after applying the ROI filtering.

B. One-vs.-one method

Here, we decompose six facial expression classification problem into 15 binary-classification problems (*i.e.*, happy vs. surprise, anger vs. fear, sad vs. disgust, and so on). We develop 15 binary-classification Hough trees. Based on these Hough trees, the likelihood of six different classes for each input patch is defined as

$$P\left(Q_c(t)\middle|p_i(\mathbf{y}),\ T\right)=\sum_{j,j\neq c}w_{cj}P\left(Q_{c\leftrightarrow j}(t)\middle|p_i(\mathbf{y}),T_{c\leftrightarrow j}\right) \quad (11)$$

where $T=\{T_{c\leftrightarrow j}\}$, c=1~6, and j=1~6. Let $Q_{c\leftrightarrow j}(t)$ be the random event corresponding to the possible existence of facial expression with labeled class c or j, and centered at t. $T_{c\leftrightarrow j}$ is the binary-classification Hough tree for differentiating classes c and j. The weight w_{cj} for the likelihood generated from each Hough tree are defined as

$$w_{cj}=\begin{cases}-w_c, & p_c^L<p_j^L\\ w_c, & p_c^L>p_j^L\end{cases} \quad (12)$$

where the weights for each class c are normalized as $\sum_j w_{cj}=1$. A positive weight w_c is assigned if the likelihood (or votes) of the designated class is larger than its counterpart, otherwise a negative weight $-w_c$ is assigned. The weight is class-dependent and determined experimentally. For certain pair of ambiguous facial expressions such as angry vs. disgust, the assigned weight will be smaller. Finally, all the weighted likelihoods are added for each class. The one with the largest accumulated weighted likelihood is identified as the correct facial expression $C_{expression}$ as

$$C_{expression}=Argmax_cP\left(Q_c(t)\middle|p_i(\mathbf{y}),\ T\right). \quad (13)$$

V. Experimental Results

Here, we demonstrate our experiments and show the experimental results. We also compare our method with the others based on the same test dataset to analyze the advantages and disadvantages of our method. There are two different video datasets for continuous facial expression recognition: Cohn-Kanade+ AU-coded facial expression database (Cohn-Kanade+) and MMI-Facial Expression Database (MMI). In the experiments, we test our method by using two facial expression database and compare our results with the others. Besides, we also create another facial expression video database by videoing the different facial expressions of our colleagues.

A. Cohn – Kanade+ Dataset

This video dataset [19, 20] is recorded from 210 people with ages from 18 to 50. The gender ratio is that 69% female and 31% male. The racial distribution is 81% Caucasian, 13% African American, and 6% others. The resolution of each frame is 640×490 or 640×480. Each video starts from a natural expression to onset, and finally to apex. Each video present a specific Facial Action Unit (AU) for Facial Action Coding System (FACS). AU represents a designate facial action which does not include sufficient information to indicate a certain expression. In the dataset, they label all the possible expressions for each video. Each facial expression indicates an appearance of certain AUs.

Figure 12. Cohn-Kanade+ expression dataset, from neutral to onset, and to apex.

TABLE 1.The facial expression and action units (AU) [19]

Emotion	Criteria
Angry	AU23 and AU24 must be present in the AU combination
Disgust	Either AU9 or AU10 must be present
Fear	AU combination of AU1+2+4 must be present, unless AU5 is of intensity E then AU4 can be absent
Happy	AU12 must be present
Sadness	Either AU1+4+15 or 11 must be present. An exception is AU6+15
Surprise	Either AU1+2 or 5 must be present and the intensity of AU5 must not be stronger than B
Contempt	AU14 must be present (either unilateral or bilateral)

In this experiment, we choose the facial expression dataset from Cohn-Kanade+ database which has sufficient information and strongly indicates certain expression. Our purpose is to avoid select some videos that provide no indication of certain facial expression. We have selected and re-labeled the videos of six different facial expressions from Cohn-Kanade+ database as shown in Table 2.

TABLE 2. Cohn-Kanade+ hand-labeled database.

Class	Angry	Disgust	Fear	Happy	Sad	Surprise
Quantity	41	55	25	69	28	80

B. MMI Dataset

The MMI video dataset [21] consists of 19 males and females and includes different AUs and expressions. Each

video starts from neutral to onset, to apex and then back to natural expression. The resolution of each frame is 720×480. Each subject is filmed without prior training or head-motion limitation. The length of video is also different for different subject perform the action. Therefore, it greatly increases the complexity of analyzing the facial expression of the videos in MMI dataset. Figure 13 shows MMI dataset. Each video consists of the following transitions: Neutral→Onset→ Apex→Offset→Neutral. In Figure 13, there are three subjects, and each subject has his own way to express the angry expression. We may see that their way of showing the anger and their head motion are different.

Figure 13. MMI dataset of different facial expressions.

TABLE 3. Different expression videos in MMI database

class	angry	disgust	fear	happy	sad	surprise
quantity	32	30	29	38	32	41

C. *Lab708 Dataset*

We have another facial expression video database by videoing the different facial expressions of 10 graduate students. They have not priori guiding to make their facial expressions. However, they are not allowed to move their heads during making their expressions. The resolution of each frame is 720×480, and each subject has recorded 13 video for each expression as shown in Figure 14.

Figure 14. The fear expression of four different subjects in our database.

In the experiments, the outcome is the number of votes. If the number of votes is larger than certain threshold, we identify the location of the accumulator and the class of the facial expression. The threshold is determined as follows:

$$V(c,t) > 1.3 * \frac{1}{|C|}\sum_{c\in C} V(c,t) \quad (14)$$

Based on the threshold, we may identify the most likely facial expression and the frame number that the expression ends.

In the experiments, we determine the size of the patch and the number of selected patches per frame based on the dataset. Here, we train 10 Hough trees and choose 10 videos for testing. In the experiments, the patch sampling rate is fixed at 300 patches per frame. For fixed sampling rate, we can find the best patch size. Then we fix the patch size and vary the sampling rate to find the best recognition rate. In the first experiment, we find that the patch size may influence the recognition rate. The larger patch size will include more temporal information of the facial motion for higher recognition accuracy. However, if the patch size is larger than 20×20, the recognition accuracy decreases. The recognition accuracy vs. the patch size is shown in Table 4.

From Figure 15, we find three patches sizes 10×10, 15×15 and 20×20 have demonstrate the best recognition rate. Then, we illustrate the influence of the number of patch samples selected per frame for facial expression recognition. The larger number of samples in the video will induce more votes, the statistics will become more un-biased, and the recognition accuracy increases. However, more patch samples indicate more complicated voting process and more computation time. We find that the recognition accuracy will not increase once the number of samples per frame is more than 200 as shown in Figure 16.

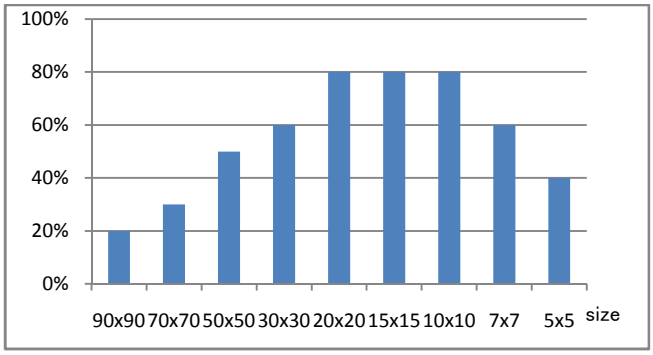

Figure 15.. The recognition rate for different patch size.

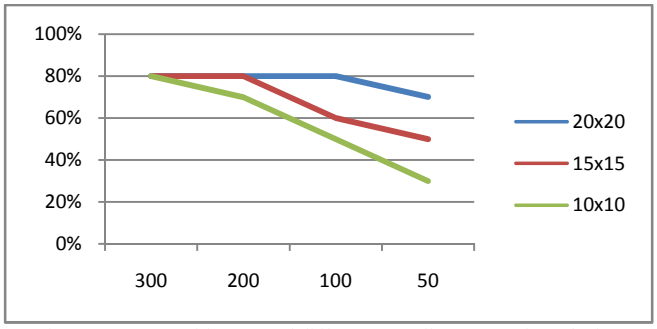

Figure 16. Recognition rate of different sampling rate and patch sizes

In the 2nd experiment, we find that the higher sampling rate will induce a more stable recognition rate, however the recognition time will be increased. When the patch size is 20×20×3 and the sampling rate is 100 per frame, we have the best recognition rate and system performance. Our system can process 700 patches per frame.

One-third of Cohn-Kanade+ dataset are for training, and the rest two-thirds are for testing. MMI dataset includes 5 sequences for each expression. We select 3 videos from Lab 708 dataset for training and the rest for testing. The Cohn-Kanade+ data set are not sufficient. For Cohn-Kanade+ data set, we do not train our classifier by using the modified multi-voting Hough forest but using the conventional Hough forest. The single-person video sequence is not sufficient for single Hough forest training. For MMI and Lab-708 dataset, the multi-voting Hough forest training can be applied. The patch size for training and testing is 20×20×3. In testing, the sampling rate is 100 per frame. Each Hough forest consists of 5 or 3 Hough trees. We train an odd number of Hough trees for the final voting.

TABLE 4. the recognition rate for Cohn-Kanade+ dataset.

	angry	disgust	fear	happy	sad	surprise	Recog. rate
angry	0.75	0.1	0	0.13	0	0	0.75
disgust	0	0.97	0	0.025	0	0	0.97
fear	0	0	0.88.	0.055	0	0.055	0.88
happy	0	0	0.02	0.97	0	0	0.97
sad	0.10	0	0.10	0	0.80	0	0.8
surprise	0	0	0	0	0	1	1

TABLE 5. The recognition rate for the MMI data set

	angry	disgust	fear	happy	sad	surprise	Recog. rate
angry	0.59	0.22	0	0	0.18	0	0.59
disgust	0.12	0.80	0.08	0	0	0	0.80
fear	0	0	0.62	0.04	0	0.33	0.62
happy	0	0	0	0.9	0	0.09	0.90
sad	0.11	0.07	0.14	0	0.66	0	0.66
surprise	0	0	0.13	0	0	0.86	0.86

TABLE 6. The recognition rate for the LAB708 dataset

	angry	disgust	fear	happy	sad	surprise	Recog.rate
Angry	0.6	0.1	0.1	0	0	0.1	0.6
Disgust	0.1	0.8	0.1	0	0	0	0.8
Fear	0	0.1	0.7	0	0	0.2	0.7
Happy	0	0	0	1	0	0	1
Sad	0	0.1	0.2	0	0.7	0	0.7
Surprise	0	0	0	0	0	1	1

D. Comparisons with other methods

In the section, we compare the performance of our method with the other three methods by using Cohn-Kanade+ data set. The features and classifiers used for the other three methods are shown in Table 7.

TABLE 7. The methods of three different methods

Methods	Features	Classifier
Aleksic[22]	Facial animation parameters	HMM
Kotsia[23]	Gabor wavelet	SVM
Yeasin[24]	PCA optical flow	HMM

TABLE 8. Comparison of four different methods.

Recog. rate	Ours	Aleksic	Yeasin	Kotsia
angry	0.75	0.70	**1**	0.86
disgust	**0.97**	**0.97**	0.62	0.87
fear	0.88	0.88	0.76	**0.92**
happy	0.97	**0.98**	0.96	0.95
sad	0.80	0.96	**0.96**	0.89
surprise	**1**	1	**1**	0.96
Avg.	0.89	**0.93**	0.90	0.90

From the experimental results, we find that the recognition rate is not good enough for the three expressions: angry, fear and sad. However, the recognition rate is much higher than the previous three for the other three expressions due to bad feature extraction. The two image features cannot fully reflect the difference between angry and sad. The 3D mouth deformation information is not obtained from the patches as shown in Figure 17.

Figure 17. The angry and sad expressions.

We compare system performance by using three different datasets and find that the recognition rate for the testing data set from MMI is the worst. For the other two datasets, we have better recognition rate. The reason is that we cannot have a very precise face calibration for MMI facial images. MMI facial images have a larger 3-D head motion that complicates the face calibration process. However, under the condition of no-limitation for the test subject, we still can have acceptable facial expression recognition rate due to the multi-voting Hough forest structure.

TABLE 9. The limitation of subjects in the three dataset.

Dataset	Training	Restricted head motion	Avg. Accuracy
Cohn-Kanade+	Yes	Yes	0.89
MMI	No	No	0.73
Lab708	No	No	0.8

Our system demonstrates a pretty good performance for the dataset in which the subjects have head motion limitation. For dataset of larger head motions, the reliability of the extracted spatial-temporal feature reduces and the recognition rate decreases. Our system requires that the input video facial expression evolves over time from the onset, the apex, and the offset. However, the image-based methods have no such limitation. It takes only one shot as observations which capture the image characteristics at the apex of the expressions.

VI. CONCLUSIONS

We introduce a 3-D spatial-temporal local feature extraction for identify the facial expression by applying Hough forest. We have also applied the ROI filtering to reduce the error during the training process and increase the discriminative capacity of the parameter voting. Our facial expression identification method is people-independent. Human beings usually do not make unanimous justification of the video with facial expression. Human facial expression identification is still a difficult un-solved problem.

REFERENCE

[1] A. Mehrabian. Communication ithout Words. *Psychology Today*, Vol.2, No.4, pp. 53-56, 1968

[2] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, vol. 36, pp. 259-275, Sep. 2003

[3] P. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE CVPR*, Dec. 2001, vol. 1, pp. 511-518.

[4] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon,. Emotion recognition using PHOG and LPQ features. the 9th IEEE Int. Conf. on Automatic Face Gesture Recognition and Workshops (FG'2011), Facial Expression Recognition and Analysis Challenge Workshop (FERA), pages 878–883, 2011.

[5] C. Shan, S. Gong, and P. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. I*mage and Vision Computing*, 27(6):803–816, 2008.

[6] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. on PAMI,* 29 (6) (2007) 915–928.

[7] P. Yang, Q. Liu, D.N. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. *CVPR* 2007.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. on PAMI,* vol. 23, no. 6, pp. 681-685, 2001.

[9] Valstar, M.F. and Pantic, M. Fully automatic facial action unit detection and temporal analysis. *Proc. CVPR*, vol. 3, 149, 2006.

[10] P. Ekman and W. V. Friesen, Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, vol. 17, pp.124-129, 1971.

[11] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the Art. *IEEE Trans. on PAMI*, vol. 22, no. 12, pp. 1424-1445, 2000.

[12] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.

[13] S. Yang and B. Bhanu. Facial expression recognition using emotion avatar image," Int. Conf. on Automatic Face & Gesture Recognition, pp. 866 –871, 2011.

[14] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt. Performance of optical flow techniques. Proc. of CVPR, 1992, pp.236-242.

[15] S. Koelstra and M. Pantic. Non-rigid registration using freeform deformations for recognition of facial actions and their temporal dynamics. Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition. 2008.

[16] L. Breiman. Random Forest. *Machine Learning*, 45(1), 5~32, 2001.

[17] J. Gall and V. Lempitsky. Class-specific Hough forests for object detection. IEEE *CVPR*, 2009.

[18] A. Yao, J. Gall, and L. Van Gool. A Hough transform-based voting framework for action recognition. *IEEE CVPR,* 2010.

[19] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. Automatic Face and Gesture Recognition. (2000) 46–53.

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The Extended Cohn-Kande Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression. The 3rd IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2010.

[21] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, Web-based database for facial expression analysis. ICME 2005.

[22] P. S. Aleksic, and A.K. Katsaggelos. Automatic facial expression recognition using facial animation parameters and multi-stream HMMs. IEEE Trans. on Information Forensics and Security (1).

[23] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," Image and Vision Computing, vol. 26, no. 7, pp. 1052-1067, July 2008.

[24] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *Transactions on Multimedia* 8 (2006) 500 – 508