Natural Facial and Head Behavior Recognition using Dictionary of Motion Primitives

Qun SHI†, Nonmember, Norimichi UKITA†∥, Senior Member, and Ming-Hsuan YANG††, Nonmember

SUMMARY This paper proposes a natural facial and head behavior recognition method using hybrid dynamical systems. Most existing facial and head behavior recognition methods focus on analyzing deliberately displayed prototypical emotion patterns rather than complex and spontaneous facial and head behaviors in natural conversation environments. We first capture spatio-temporal features on important facial parts via dense feature extraction. Next, we cluster the spatio-temporal features using hybrid dynamical systems, and construct a dictionary of motion primitives to cover all possible elemental motion dynamics accounting for facial and head behaviors. With this dictionary, the facial and head behavior can be interpreted into a distribution of motion primitives. This interpretation is robust against different rhythms of dynamic patterns in complex and spontaneous facial and head behaviors. We evaluate the proposed approach under natural tele-communication scenarios, and achieve promising results. Furthermore, the proposed method also performs favorably against the state-of-the-art methods on three benchmark databases.

key words: facial and head behaviors, dynamical systems, motion primitives

1. Introduction

Existing methods for human facial and head behavior recognition mainly focus on expressive faces with deliberately and prototypical expressions of six basic emotion categories (i.e. anger, disgust, fear, happiness, sadness, and surprise). This practice may follow from the work of Darwin and more recently Ekman [1], who suggested that basic emotions have related prototypic expressions. For analyzing natural conversation scenarios, other facial and head behaviors such as speaking or nodding are also of great importance.

The goal of this work is recognition of spontaneous natural facial and head behaviors of broad age groups in daily conversations. This problem is more challenging than the one using conventional datasets in several aspects. First, different people have characteristic facial structures (especially for elder people), which make recognition difficult. Second, the rhythms of a spontaneous behavior vary in sequences. These behaviors may not always contain the distinct transition phases from neutral to certain expressions and vice versa. Thus it is difficult to align the temporal segments of spontaneous behaviors. Third, multiple facial and head behaviors are likely to occur simultaneously during natural communications, i.e., compound behaviors, such as smiling while nodding. Existing databases rarely contain these compound behaviors, and few methods address these issues.

To tackle the challenging issues discussed above, we propose a novel approach using the dictionary of motion primitives (DoMP). Figure 1 shows the main steps of the proposed method; see the caption for its explanation.

1. The main idea is to first capture the holistic movements on important facial parts via dense feature extraction. Part-wise motion analysis allows us to suppress a bad effect due to different individual facial structures.
2. Next, we spatiotemporally segment these features, construct dynamic models in a segment-wise manner, and cluster them to form a DoMP. With the DoMP, facial and head behaviors can be interpreted based on a distribution of motion primitives. The distribution-based representation is less likely to be affected by different rhythms in natural conversations.
3. The proposed method is evaluated with two different strategies in order to find the one with better recognition accuracy for compound facial and head behaviors.

Recently, the problem setting of facial and head expression recognition has been extended from using still images to using videos as well as from a lab setting to a natural scenario. While recognition in a natural scenario is regarded as a next research step [2], several new datasets and bench-

Manuscript received April 12, 2017.
Manuscript publicized August 28, 2017.
†The authors are with Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
∥The author is with University of California, Merced, USA.
‡Presently with Toyota Technological Institute.
a) E-mail: ukita@toyota-ti.ac.jp
DOI: 10.1587/transinf.2017EDP7128
marks are captured still in a lab setting and contains only six basic emotions (e.g., [3]). The proposed method is evaluated on a challenging daily conversation dataset with seven behavior classes observed during conversation (e.g., speaking while nodding), and demonstrated to effectively recognize spontaneous compound facial and head behaviors. In addition, we evaluate the proposed method on three conventional databases with simple facial and head behaviors against the state-of-the-art methods.

The contributions of this paper are as follows:

- We develop novel statistical motion descriptors for natural facial and head behavior recognition by integrating dense feature extraction, segmentation, and dynamic models.
- The proposed method is evaluated on a challenging tele-communication dataset, which contains compound facial and head behaviors of people with different ages, genders, and races. The proposed method performs well both on simple and compound behaviors.

2. Related Work and Problem Context

Many existing facial behavior recognition methods use still images where each characteristic face is captured at the apex (i.e., the exact moment where facial and head motion is exerted most) [4]–[6]. However, psychological studies [7] have shown that video recognition enables more accurate and robust recognition of facial expressions.

For videos, many methods use anthropomorphically-meaningful local facial regions such as eyes and mouth to describe facial expressions. For example, Avent et al. [8] develop a low-level method to detect and classify local facial features of different expressions based on edges corresponding to eyes, eyebrows, and lips using neural networks. Kumar and Poggio [9] use the support vector machine to track local features to analyze expressions. These features can be tracked under varying facial expressions [10]. Moses et al. [11] use valley contours for tracking local facial features and recognizing facial expressions.

Instead of relying on localized deformations, temporal segments of either full expressions [12], [13] or components of low-level expressions such as facial action units (AUs) are used for recognition [14], [15]. The effectiveness of mid-level temporal representations is also explored [16]. These methods make use of complete 2D image sequences, and track the motion between frames. However, in most cases when 2D facial intensity images are used, it is necessary to maintain a consistent facial pose (preferably a frontal one) for good performance. Even small motion changes in head pose would affect the recognition performance.

To recognize dynamic features, Ravichandran et al. [17] model each video with a collection of motion primitives that describe the dynamics of spatiotemporal video patches. This bag-of-motion-primitives (BoMP) approach is similar to the work on using the bag-of-features (BoF) for object recognition which categorizes images by observing the distribution of a small collection of features. The main difference lies in that the BoMP uses dynamical models as feature descriptors for a temporal representation rather than a spatial representation by the BoF. Note that while the BoMP has been applied to recognize dynamic textures in constrained setups, it has not been extended to recognize more complicated events such as natural facial and head behaviors. On the other hand, Bousmalis et al. [18], Jung et al. [19], and Dapogny et al. [20], use Hierarchical CRF, neural networks, and random forests for representing a temporal sequences of framewise static cues, respectively. Evangelos et al. [21] generates a temporal BoW from framewise static cues. Unlike these state-of-the-art methods, our proposed method encodes the temporal sequences to be the BoMP.

Our BoMP is employed in order to cope with the different rhythms of facial and head behaviors in natural conversations. Typical examples are speaking and nodding. Among sequences, “the number of times of mouth opening and closing movements” and “the number of nods” are different in the speaking and nodding behaviors, respectively. Furthermore, a compound behavior makes it difficult to segment and recognize each element of the continuous speaking and nodding behaviors. Our BoMP approach is not affected by the different rhythms by recognizing a temporally-holistic property in such complex behaviors.

For BoMP construction, the proposed method first captures facial and head motions using dense optical flow. In this work, we use a histogram of oriented dense optical flow features (HOOF) [22] to encode facial and head expressions. The HOOF features have been shown to be effective in action recognition and other problems [22]. These features are then spatiotemporally segmented and modeled by Interval-based Hybrid Dynamical Systems [23], [24], which allow us to discriminate between subtle differences among motion primitives, which are defined by Linear Dynamics Systems (LDSs) in the proposed method. By clustering all those LDSs into a DoMP, an image sequence can be described as the BoMP. Although the proposed method also performs frame-by-frame local feature detection, it is based on the statistical analysis of large amount of dynamic features, instead of relying on precise tracking of certain feature points. Thus, the proposed method performs more robustly under different environments.

After dynamic feature extraction is introduced in Sect. 3, we propose how to generate motion primitives in Sect. 4. Facial and head behavior recognition using the motion primitives is described in Sect. 5. We present experimental results with comparisons to the state-of-the-art methods for facial and head behavior recognition in Sect. 6 with discussions and conclusions in Sect. 7.

3. Extracting Dynamic Facial Features

In order to retrieve sufficient and important motion information of facial and head behaviors, we use the dense optical flow as the dynamic features. The dense flow allows us to capture muscle/skin movements in featureless regions (e.g.,
forehead and cheeks) caused by micro expressions, which are in principle difficult to find by tracking only feature points such as the corners of eyes and mouth. The effectiveness of the dense flows for facial expression analysis has been validated in the literature [25], [26].

As illustrated in Fig. 2, the proposed dynamic facial feature extraction process consists of three steps: (1) detecting facial feature points, (2) segmenting regions-of-interest (ROI), and (3) computing dense optical flow.

First, the Active Shape Model (ASM) [27] is used to detect facial feature points. In this work, 77 feature points are detected from each frame, describing the contours of the facial parts. Based on the positions of the detected feature points, sub-regions of the facial parts of interest (eyes, noses and mouth) are segmented as rectangular windows. The length and width of each ROI are two times the corresponding facial part’s length and width. Finally, dense optical flows are computed inside the ROIs to capture as much as possible information of the facial and head motions.

The feature points are used for ROI segmentation. The ROIs do not have to be precise because they are employed for obtaining dense optical flows around facial parts. This property facilitates the proposed method to perform robustly to the noisy results of feature point localization. It should be also noted that in order to minimize the influences of the global body/head motions on the detection of the detailed local motions, we perform a global motion subtraction process. In this process, optical flows in each frame are obtained by subtracting the mean flow from raw computed flows. Examples are illustrated in Fig. 3.

4. Constructing Dictionary of Motion Primitives

The motivation for using motion primitives is that complex human behaviors consist of many motion primitives, which are often referred to as motion elements, moveemes, and visemes. Although different instances of human behaviors may vary in terms of their overall motions, many of the motion primitives included in the overall motions are similar. Once the set of motion primitives is determined, complex behaviors can be partitioned into temporal segments, each of which is characterized by a motion primitive.

4.1 Feature Descriptors

After computing the dense optical flows in the former step, the flow sequences are divided into $m \times n \times t$ 3D blocks with the same size. $m$ and $n$ corresponding to $x$ and $y$ image axes, and $t$ is a temporal window size (i.e., the number of frames). In this work, $m$, $n$, and $t$ were predefined and fixed throughout experiments. For each block we compute the HOOF [22] features based on the directions and lengths of the dense optical flow to represent the motion. The reason for using HOOF features is that they are independent of the scale of moving object as well as the motion direction, and robust to small noisy optical flow measurements. The number of bins in the histogram is set empirically and fixed in all experiments. After this process, each sampled 3D block contains a feature descriptor as a temporal sequence of HOOFs.

4.2 Dictionary of Motion Primitives (DoMP)

The next process is to cluster the block-based feature descriptors into a DoMP. The descriptors are modeled by the Interval-based Hybrid Dynamical System (IHDS) method [23], [24] to describe the dynamics of block-based HOOF features. The IHDS scheme is extended from switching LDSs [28], [29] and hybrid dynamical systems [30],...
which have been demonstrated in human action recognition. Aside from finite state transitions represented in general switching LDSs (e.g., transitions between walking and jogging in [29]), the IHDS method models a temporal interval at each finite state. Since each dynamic state in a facial and head expression has a relatively-regular interval rather than human activities in general scenarios, the IHDS method performs well for facial and head expression recognition.

The IHDS method consists of a discrete-event system and a dynamical system that is described by differential equations. As shown in Fig. 4, the IHDS method has a two-layer architecture. The first layer illustrated in the upper part of the figure has a finite state automaton that models stochastic transitions between intervals. The second layer consists of a set of multiple LDSs, \( D = \{ D_1, \ldots, D_N \} \).

Each interval is described by \( \langle q_i, \tau \rangle \) where \( q_i \) denotes a discrete state in the automaton and \( \tau \) denotes the physical temporal duration length of the interval. It is assumed that each state, \( q_i \), in the automaton corresponds to a unique LDS, \( D_i \). The state transition in dynamical system \( D_i \) is modeled by the following equation:

\[
\begin{align*}
x_t &= F^{(i)} x_{t-1} + \omega^{(i)}_t, \\
y_t &= H x_t + v_t, \tag{1}
\end{align*}
\]

where \( x_t \) is the interval state vector at time \( t \) and \( F^{(i)} \) is a transition matrix. \( \omega^{(i)}_t \) and \( v_t \) are the process and observation noises modeled by Gaussian distributions \( N(0, Q^{(i)}) \) and \( N(0, R) \), respectively. Note that each dynamical system has its own \( F^{(i)}, g^{(i)} \) and \( \omega^{(i)} \) while all the dynamical systems share a single continuous state space (i.e., \( H \) and \( v_t \)) in order to reduce parameters in accordance with [23], [24].

The goal of the IHDS identification is to estimate the number of LDSs, \( N \), and the parameter set of all the systems \( D_i \). The estimation process is divided into two steps in order to reduce a bad effect of initialization. 1) A clustering process of dynamical systems using a typical training data and 2) a refinement process for all the parameters based on the Expectation-Maximization (EM) method using all the training data. At the same time, we can model all the training data into temporal interval sequences. While readers may refer to [23], [24] for the details of this learning method, its summary is described in what follows.

In the first clustering process, a typical sequence in each class is clustered in regular short intervals. A dynamical system, which is represented by Eqs. (1) and (2), is obtained in each cluster and then the distance between all pairs of the dynamical systems is computed. The distance is represented by the KL divergence [31]. The nearest pair is merged iteratively. The convergence of this merging process is determined based on the ratio of the inter-class variance to the intra-class variance of the KL divergence in accordance with the discriminant analysis. This process obtains the number of dynamical systems and the initial set of frames included in each dynamical system. With this initial set, the initial parameters of each dynamical system and state transition probabilities among the dynamical systems are computed.

The parameters of \( D_i \) in Eqs. (1) and (2) are estimated in our clustering process, the interval of every dynamical system is fixed at each parameter estimation process as follows\(^1\). With the Singular Value Decomposition (SVD), \( H = U \) and \( X_{1:T} = S V^T \), where \( SVD(Y_{1:T}) = [U, S, V] \). \( Y_{1:T} \) and \( X_{1:T} \) are matrices consisting of \( y_i \) (i.e., HOOF features in our case) and \( x_i \) between 1st and \( T \)-th frames in each cluster, respectively. \( F^{(i)} \) is computed by minimizing \( ||F^{(i)} X_{1:T-1} - X_{2:T}||^2 \). \( \omega^{(i)}_t \) and \( v_t \) are estimated from Eqs. (1) and (2) given \( X_{1:T}, Y_{1:T}, F^{(i)}, \) and \( H \).

In the second refinement process, the number of dynamical systems is fixed during the whole process of the following EM method. The EM method is achieved with all training sequences. The E step optimizes the intervals of the clusters (i.e., frames in each cluster) with the fixed parameter set. Then the M step updates the parameter set in each optimized cluster. The EM steps are repeated while the likelihood of the parameters changes greater than a threshold.

After the learning phase, we can model newly-observed data using the learned IHDS. When an observed sequence is given, the IHDS method estimates an optimal interval sequence to describe the observed data based on maximum likelihood using the Viterbi algorithm [23], [24]. With this method, we can represent a newly-observed video with interval sequences using the learned IHDS. By using the IHDS method, the dynamics of HOOF features in each block is modeled by a LDS.

After clustering the parameters of LDSs based on the EM algorithm, we obtain a DoMP as \( K \) clusters of LDSs.

We note that the HOOF features cannot be modeled by LDSs for recognition of human actions since a non-Euclidean feature space is used [22]. However, we use LDSs to model block-wise HOOF features based on the assumption that nonlinear signals or nonlinear feature sequences can be represented by piecewise linear systems when appropriate units of primitives are used. That is, we assume a complex dynamic event comprising a set of sub dynamic

---

\(^1\)See [32], [33] for more details
events, and the observed signals or feature sequences that are describing each of the temporal regions of sub dynamic events can be represented by a LDS.

4.3 Video Interpretation via DoMP

Once a set of \( K \) dictionary words of motion primitives are extracted, each video sequence can be represented using this vocabulary. This is carried out with a weighting vector \( W = \{w_1, w_2, \ldots, w_K\} \).

Assume that the motion primitive word \( k \) occurs \( N_k \) times in an input video and there are total of \( N \) motion primitive words in the input video. A straightforward representation of \( w_k \) is the term frequency defined by

\[
w_k = \frac{N_k}{N}, \tag{3}\]

When a weighting vector \( W \) is computed according to Eq. (2), it is normalized and represented by a histogram, which will be used for classification.

5. Classification

For classification of compound facial and head behaviors, two different strategies are employed. First, we consider compound behaviors as independent classes. Second, we consider a compound behavior as a combination of elemental facial and head behavior classes. For example, “speaking”, “smiling”, and “speaking and smiling” have their own independent classifiers in the first strategy. In the second strategy, on the other hand, only each of “speaking” and “smiling” is modeled by a classifier and “speaking and smiling” is found if both of “speaking” and “smiling” classifiers have higher scores.

We used the support vector machine (SVM) [34] for classification. For the first strategy, the classification task is partitioned into binary decisions, by training the SVM in the one versus all manner. The final facial and head behavior is determined based on the one with highest score. For the second strategy, if the estimated scores of multiple SVM outputs are higher than a threshold, the input video is considered as a compound facial and head behavior composed of simple ones with scores that are higher than the threshold.

6. Experiments

To evaluate the proposed method, experiments are conducted with a natural tele-communication dataset and three benchmark databases with comparisons to the state-of-the-art methods. In all experiments, we used LibSVM [35] with default parameters for classification.

6.1 Tele-Communication Dataset

We aim to recognize natural facial and head behaviors from tele-communication videos of daily conversations. The videos are acquired from the chatting sessions of users using webcams. The behavior of each subject in the conversation is recorded with the resolution of 1280×720 pixels and 25 frames per second.

6.1.1 Characteristics of Tele-Communication Dataset

We captured video sequences of tele-communication dialogues conducted by different people. For data collection, we asked 14 people to freely talk to each other in pairs through webcams, and during each conversation the behaviors of the two participants were recorded. Each video lasts for at least 10 minutes to ensure the amount of facial and head behaviors occurring inside the data. Note that in our dataset, subjects in the natural conversational videos cover different ages, genders and races, as shown in Fig. 5.

In this dataset, each video segment is classified into one of seven classes of facial and head behaviors: Neutral, Speaking, Smiling, Nodding, Speaking while nodding, Smiling while nodding, and Speaking while smiling. The former four facial and head behaviors belong to simple behaviors, while the latter three are compound behaviors. We manually extracted video segments corresponding to these seven classes from the recorded videos and labeled them with one of the classes.

As a pre-processing step, we manually crop short video segments in which the person performs one type of facial and head behavior in the category. In total, 5000 frames are used for learning the IHDS parameters. For facial and head behavior recognition, 150 video segments are collected for each type of facial and head behavior, half used as the learning data and the other half as testing data for the SVM. Different pre-processed video segments have different temporal durations. For validating the generalization capability of the proposed method, subjects included in training video segments were not included in test video segments.
Table 1 Confusion matrix of the proposed approach for the 7-class natural conversation scenario. The row and column indicate the ground-truth and recognition results, respectively. The rightmost and other columns show recognition accuracy and the number of samples recognized as the expression class of each column.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Neutral</th>
<th>Speaking (Sp.)</th>
<th>Smiling (Sm.)</th>
<th>Nodding</th>
<th>Nodding+Sp.</th>
<th>Nodding+Sm.</th>
<th>Sm.+Sp.</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
<td>2nd</td>
<td>1st</td>
<td>2nd</td>
<td>1st</td>
<td>2nd</td>
<td>1st</td>
<td>2nd</td>
</tr>
<tr>
<td>Neutral</td>
<td>72</td>
<td>73</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Speaking (Sp.)</td>
<td>1</td>
<td>1</td>
<td>59</td>
<td>65</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Smiling (Sm.)</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>60</td>
<td>63</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nodding</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>61</td>
<td>69</td>
</tr>
<tr>
<td>Nodding+Sp.</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Nodding+Sm.</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sm.+Sp.</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>20</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

6.1.2 Experimental Results

Table 1 shows the confusion matrix of recognition results in two strategies, which are described in Sect. 5. The accuracy score in each row, which is shown in the rightmost column, is defined to be \( TP/(TP + FN) \), which is also known as Recall, where \( TP \) and \( FN \) denote “the number of correctly-recognized samples” and “the number of samples misrecognized as a wrong class”, respectively.

The second strategy obtains better recognition results in all classes except the “speaking while smiling” class. The second strategy achieves an average accuracy of 85.0% for the recognition of 7 types of facial and head behaviors. It is better than the first strategy (81.7%). While the first strategy uses the classifiers for all behaviors including elemental behaviors (i.e., “Neutral”, “Speaking”, “Smiling”, and “Nodding”) and their compound behaviors, only elemental behaviors are represented by the classifiers. We attribute the higher accuracy of the second strategy to better discriminativity of a small number of elemental behaviors. While lower accuracy of the second strategy in “speaking while smiling” is caused because many samples (i.e., 20 samples) are misrecognized as “Speaking”, further analysis for improvement will be conducted in our future work.

6.2 Benchmark Datasets

In order to further evaluate the effectiveness of the proposed method using the second strategy, we use three benchmark databases with comparisons to the state-of-the-art methods. The experimental setups are described as follows.

- Oulu-CASIA VIS Database: The Oulu-CASIA VIS database [36] includes 80 subjects between 23 and 58 years old, with six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) of each person. Each video starts at a neutral face and ends at the apex of expression. We perform a person-independent 10-fold cross-validation on the total 480 sequences.
- MMI Database: The MMI database [37] includes 30 subjects of both male and female, aging from 19 to 62. In the database, 236 sequences have been labeled with six basic expressions, in which 205 sequences are captured in frontal views. Each sequence captures the entire “onset apex offset” phase of a single facial expression type. In our experiments, we used all of frontal view data and conducted a person-independent 10-fold cross-validation.
- AFEW Database: The Acted Facial Expression in Wild (AFEW) database [38] has been collected from movies which depicts or simulates the spontaneous expressions in uncontrolled environments. The database consists of three sets: training, validation and test. There are seven classes of expressions in the data (six basic classes and a neutral class).

Sample images of these datasets are shown in Fig. 6. Regarding training/testing separation, we followed the protocols in previous works. In the OULU-CASIA VIS and MMI datasets, a person-independent 10-fold cross validation was performed after all data was separated into 10 groups of equal size. In the AFEW dataset, 380 images and 396 images were used for training and testing, respectively.

The experimental results on Oulu-CASIA VIS, MMI and AFEW databases are illustrated in Table 2, Table 3 and
### Table 2  Experimental results on the Oulu-CASIA VIS database.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adal.BP[36]</td>
<td>73.54</td>
</tr>
<tr>
<td>Atlases [39]</td>
<td>75.52</td>
</tr>
<tr>
<td>STM-ExpLet[16]</td>
<td>74.59</td>
</tr>
<tr>
<td>Proposed</td>
<td>77.51</td>
</tr>
</tbody>
</table>

### Table 3  Experimental results on the MMI database.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADL [40]</td>
<td>47.78</td>
</tr>
<tr>
<td>HMM [41]</td>
<td>51.5</td>
</tr>
<tr>
<td>ITBN [41]</td>
<td>59.7</td>
</tr>
<tr>
<td>CPL [42]</td>
<td>49.36</td>
</tr>
<tr>
<td>CSPL [42]</td>
<td>73.53</td>
</tr>
<tr>
<td>STM-ExpLet[16]</td>
<td>75.12</td>
</tr>
<tr>
<td>Proposed</td>
<td>78.83</td>
</tr>
</tbody>
</table>

### Table 4  Experimental results on the AFEW database.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EmotiW [43]</td>
<td>27.27</td>
</tr>
<tr>
<td>STM-ExpLet[16]</td>
<td>31.73</td>
</tr>
<tr>
<td>Proposed</td>
<td>34.32</td>
</tr>
</tbody>
</table>

### Table 5  Detailed experimental results of the proposed method on the AFEW database.

<table>
<thead>
<tr>
<th>Expression class</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EmotiW [43]</td>
</tr>
<tr>
<td>Anger</td>
<td>44.0</td>
</tr>
<tr>
<td>Disgust</td>
<td>2.0</td>
</tr>
<tr>
<td>Fear</td>
<td>14.8</td>
</tr>
<tr>
<td>Happy</td>
<td>43.6</td>
</tr>
<tr>
<td>Sad</td>
<td>20.3</td>
</tr>
<tr>
<td>Surprise</td>
<td>9.6</td>
</tr>
<tr>
<td>Neutral</td>
<td>34.6</td>
</tr>
<tr>
<td>Overall</td>
<td>27.3</td>
</tr>
</tbody>
</table>

Table 4, respectively.

In each experiment, we compare the results of the proposed method with other state-of-the-art results reported in the literature. Whereas Oulu-CASIA VIS and MMI include videos captured in lab settings, natural facial expressions observed in the wild are provided in AFEW. Videos in AFEW are obviously more challenging than other two datasets as can be seen from recognition accuracy shown in Tables 2, 3, and 4. Overall, the proposed method performs equally well or better than state-of-the-art methods on the benchmark databases composed of basic emotion based expression categories. These results demonstrate that the proposed method performs well also for basic facial expressions captured in some different scenarios, while our motivation in designing the proposed method is to cope with difficulty in natural conversation scenarios.

For detailed analysis, recognition accuracy of each expression on the challenging AFEW database is shown in Table 5. For comparison, the results of Dhall et al. [43] are also shown in the table. This comparison reveals that our proposed method is stable among recognition of different expression classes compared with Dhall et al. [43]. For example, similar expressions such as “anger” and “disgust” might be difficult to be classified by [43], while our proposed method gets better results on average.

### 7. Conclusions

This paper presented facial and head behavior recognition using dynamical systems, which consist of three steps, i.e.,

1. dynamic feature extraction, construction of dictionary based on motion primitives, and classification. We propose a novel dictionary based on motion primitives and show its effectiveness for recognition using statistical analysis without temporal alignment. Experimental results on four datasets demonstrate the effectiveness of the proposed method in recognizing natural and exerted facial and head behaviors. Our future work includes automatic extraction of video segments in each of which a natural facial and head behavior is observed in a long video sequence (e.g., video spotting). This work was supported by Yanmar Innovation Lab. 2112.

### References


[26] Q. Shi He received his B.Sc. in Computer Science from Northeastern University, China, in 2008, and M. Sc. and Ph.D. in Informatics from Kyoto University, in 2011 and 2014, respectively. While he worked as a postdoc researcher at Nara Institute of Science and Technology (NAIST), he developed methods for facial behavior detection and recognition. He has been engaged in Computer Vision.
[27] Ming-Hsuan Yang He is an associate professor in Electrical Engineering and Computer Science at University of California, Merced. He received the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign in 2000. He studied computer science and power mechanical engineering at the National Tsing-Hua University, Taiwan; computer science and brain theory at the University of Southern California; and artificial intelligence and operations research at the University of Texas at Austin. He was a senior research scientist at the Honda Research Institute (formerly Honda Fundamental Research Labs) working on vision problems related to humanoid robots. His research interests include computer vision, pattern recognition, artificial intelligence, robotics, and machine learning.
[28] Norimichi Ukita He is a professor at the graduate school of engineering, Toyota Technological Institute (TTI), Japan. He received the Ph.D. degree in Informatics from Kyoto University, Japan, in 2001. After working for five years as an assistant professor at NAIST, he became an associate professor in 2007 and moved to TTI in 2016. He was a research scientist of PRESTO, JST from 2002 to 2006, and a visiting research scientist at Carnegie Mellon University from 2007 to 2009. His main research interests are multi-object tracking and human pose and activity estimation.