Characterizing Subtle Facial Movements via Riemannian Manifold

XIAOPENG HONG, Xi’an Jiaotong University and University of Oulu
WEI PENG, University of Oulu
MEHRTASH HARANDI, Monash University and Data61-CSIRO
ZIHENG ZHOU, VeChain Foundation
MATTI PIETIKÄINEN and GUOYING ZHAO, University of Oulu

Characterizing subtle facial movements from videos is one of the most intensive topics in computer vision research. It is, however, challenging, since (1) the intensity of subtle facial muscle movement is usually low, (2) the duration may be transient, and (3) datasets containing spontaneous subtle movements with reliable annotations are painful to obtain and often of small sizes.

This article is targeted at addressing these problems for characterizing subtle facial movements from both the aspects of motion elucidation and description. First, we propose an efficient method for elucidating hidden and repressed movements to make them easier to get noticed. We explore the feasibility of linearizing motion magnification and temporal interpolation, which is obscured by the architecture of existing methods. On this basis, we propose a consolidated framework, termed MOTEL, to expand temporal duration and amplify subtle facial movements simultaneously. Second, we make our contribution to dynamic description. One major challenge is to capture the intrinsic temporal variations caused by movements and omit extrinsic ones caused by different individuals and various environments. To diminish the influences of such extrinsic diversity, we propose the tangent delta descriptor to characterize the dynamics of short-term movements using the differences between points on the tangent spaces to the manifolds, rather than the points themselves. We then relax the trajectory-smooth assumption of the conventional manifold-based trajectory modeling methods and incorporate the tangent delta descriptor with the sequential inference approaches to cover the period of facial movements. The proposed motion modeling approach is validated by a series of experiments on publicly available datasets in the tasks of micro-expression recognition and visual speech recognition.

CCS Concepts: • Computing methodologies → Computer vision tasks; Appearance and texture representations;

Additional Key Words and Phrases: Micro-expression recognition, visual speech recognition, motion magnification, motion description, video representation

We express deep gratitude to National Basic Research Program of China (Grant No. 2015CB351705), National Major Project (Grant No. 2017YFC0803905), and Natural Science Foundation of China under Contracts No. 61772419, No. 61572205, and No. 61601362, the Academy of Finland ICT 2023 (Project No. 313600), Infotech, Tekes Fidipro Program (Grant No. 1849/31/2015), Tekes project (Grant No. 3116/31/2017).
Authors’ addresses: X. Hong; email: hongxiaopeng@mail.xjtu.edu.cn; W. Peng; email: Wei.Peng@oulu.fi; M. Harandi; email: mehrtash.harandi@monash.edu; Z. Zhou; email: peter.zhou@vechain.com; M. Pietikäinen; email: Matti.Pietikainen@oulu.fi; G. Zhao (corresponding author); email: guoying.zhao@oulu.fi.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2019 Association for Computing Machinery.
1551-6857/2019/12-ART94 $15.00
https://doi.org/10.1145/3342227

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 15, No. 3s, Article 94. Publication date: December 2019.
1 INTRODUCTION

“Only the dynamics of a movement is unambiguous and convincing.” (Auguste Flach, 1921).

One of the most notable aptitudes of human beings is the capability to communicate verbally and to express emotions. Facial movements not only transmit communication contents but also contribute to ongoing expressing and processes of emotion-relevant information. As a result, characterizing facial movements from videos plays a critical role in various computer vision tasks like Micro-Expression Recognition (MER) \[10, 36, 37, 63, 83\] and Visual Speech Recognition (VSR) \[6, 93\], which have a wide range of potential applications like clinical diagnosis, lie detection, and human-computer interaction.

It is, however, challenging to capture subtle facial movements. The causes are manifold. Take micro-expression recognition, an emerging affective analysis task \[12, 90\] to distinguish subtle and spontaneous facial expressions, as an example. Recognizing spontaneous micro-expression is difficult even for human beings. Psychological studies of micro-expression have indicated that even a trained human may only achieve an accuracy lower than 50% \[14\]. First, the intensity of subtle facial muscle movements is usually low, as people tend to repress their muscle movements in case that their true feelings are being found. Moreover, micro-expressions, unlike normal facial expressions, only cover a small area of faces and thus the facial movements are too slight to be detected. Second, the duration may be transient. As suggested by \[64\], the temporary period of a micro-expression only lasts approximately from 1/25 to 1/3 seconds. Finally, the induction and capture of spontaneous micro-expression are difficult. As a result, datasets containing subtle involuntary movements with reliable annotations are painful to obtained and often of small sizes, approximately a couple of hundreds of samples per dataset.

In this article, we focus on addressing these problems in characterizing subtle facial movements. Our strategy is to solve these challenges from the following two aspects. First, we aim at designing efficient approaches to elucidate the transient and repressed facial movement so that the subtle dynamics can be observed and processed in a more relaxed manner. Second, we devote considerable attention to powerful descriptions to model the dynamics of facial movements, reconciling the difficulty brought by the small-sample-size issue.

1.1 Subtle motion pre-processing

Recently, a growing number of studies in processing subtle dynamics employ motion magnification or temporal interpolation as pre-processing steps \[23–25, 35–37, 41, 46, 56, 63, 83, 84, 94\]. Motion MAGnification (MAG) \[43, 80, 84\] amplifies the intensity of motion and makes subtle motion much easier to observe. Li et al. \[36\] amplify subtle facial movements by employing MAG before extracting features and report a substantial improvement in accuracy on the CASME II \[83\] database. Obvious improvements by introducing MAG are also observed in References \[46, 55, 56\]. Besides motion magnification, temporal interpolation also attracts increasing attention. Zhou et al. \[92, 94\] apply a graph spectral TI approach to normalize sequences for visual speech recognition. The approach is then extended to the MER task, which allows to obtain sufficient frames for presenting transient facial movements \[63\]. Fueled by their success, methods used by References \[23–25, 35–37, 41, 83\] also follow a similar path and report promising results.
Fig. 1. Illustration of the proposed dynamic elucidation method. (a) The traditional methods treat motion magnification and temporal interpolation as two separate modules and perform sequentially. Note that there is an intermediate output: the amplified sequence without temporal extension. (b) Our method provides a consolidated linear solution to realize motion magnification and temporal extension jointly. (c) By dynamic elucidation, a representative frame of the output sequence becomes easy to recognize, while the emotional states behind the one in the original input sequence can hardly be identified.

Nonetheless, there is a clear limitation in the pipeline of using MAG and TI in existing methods. More concretely, MAG and TI are usually treated as two separate modules [24, 36, 83] (see Figure 1(a) for an illustration). Although such an assumption eases the complexity in implementation, it brings several problems. First, it may introduce the diluted problem\(^1\) into the following process. Second, the calculation of intermediate products increases the computational costs. Finally, and what makes it worse, ignoring the underlying connections will inevitably result in the loss of recognition accuracy.

To address these issues, we explore the feasibility of linearizing MAG and TI in the context of digital video processing and analysis, which has been obscured from the existing implementation in the majority of cases. On this basis, MAG and TI can be explicitly reformulated within a linear system. As a result, we propose a consolidated framework for elucidating the subtle dynamics in facial movements. The proposed model, which is called MOTion ELucidation (MOTEL), accomplishes facial muscle motion magnification and temporal interpolation simultaneously, as illustrated by Figure 1(b). Compared to traditional MER systems, our MOTEL method reduces the computational time by avoiding unnecessary separation of the two modules and pre-computing auxiliary structures that are only dependent on the numbers of frames. More surprisingly, omitting the intermediate results also eliminates the side effects and thus leads to the improvement of accuracy. The influence of MOTEL will be validated on two popular, spontaneous MER databases: SMIC [37] and CASMEII [83]. Finally, we notice that existing works usually perform MAG first and then TI, but few explain why. To achieve a satisfactory understanding of the two, we need to identify not only the phenomenon but also the reason why it has to be arranged in such an order. Thus, we also investigate the influence of switching the order of motion magnification and temporal interpolation, which is beyond the scope of our preliminary research [60].

1.2 Subtle motion description
Analyzing and recognizing human facial movements is a challenging task. Facial actions like micro-expression and speaking are the results of complicated dynamic processes of muscles in the tongue, lips, jaw, orbicularis oculi, and other facial muscles. Although human facial movements have been taxonomized by the Facial Action Coding Systems [12], different individuals still have varied appearance of organs and also diverse habits. In addition, there are considerable environmental influences on the poses of subjects, illumination, and the quality of the images or videos captured. As a result, the overall appearances of two instances of facial movements belonging to

\(^{1}\)A phenomenon that the facial muscle movements fall into decline [35].
Fig. 2. Conceptual visualization of our motion description method, where two individuals pronounce the same word “Excuse.” Though (a) the corresponding descriptions of two image sequences (i.e., $M_1^{(1)}$ and $M_2^{(1)}$ for subject A, $M_1^{(2)}$ and $M_2^{(2)}$ for subject B) are away from each other on the manifold, (b) the differences between adjacent frames within each sequence (i.e., $\Delta^{(1)}$ and $\Delta^{(2)}$) are relatively consistent and stable. Our idea is thus to model the distributions of $[\Delta]$ instead of $[M]$.

The same category may vary considerably, though their intrinsic dynamics of movements should be consistent. Such variations are demonstrated by the two image sequences making the same action by two individuals in Figure 2(a), where the corresponding appearance descriptions of two sequences are apart from each other on the manifold. Therefore, one major unresolved problem is how to characterize the intrinsic dynamic process of a movement (i.e., the spatial-temporal variations directly caused by the movements), regardless of the extrinsic variations caused by different individuals and various environmental conditions.

In this work, we focus on efficiently extracting and modeling the intrinsic dynamics of movements as well. We observe that most movements of human facial muscles are too complex to be fully quantified in low-dimensional vector spaces. It leads us to adopt non-Euclidean spaces, i.e., Riemannian structures [47, 61, 62, 78, 79] to encode the dynamics of movements. Although conventional Covariance Matrix descriptors (CM) are capable of encapsulating the spatial-temporal coherence within cuboids on Riemannian manifolds, the distributions of points on manifolds belonging to the same kind of movements still vary as extrinsic variations exist [see Figure 2(a) for an illustration]. To diminish the influence of such diversities, we propose to characterize the dynamics of subtle movements via the differences between points on the tangent spaces of the manifold as illustrated in Figure 2(b). We further relax the trajectory-smoothness assumption of conventional manifold-based methods and rely on statistical models to attain more resilience to the perturbation of points. Finally, we incorporate the tangent delta descriptor with the sequential inference approaches and present a hierarchical representation architecture to cover the period of a facial movements occurrence.

The contributions of this article are threefold: First, we propose a consolidated video motion revealing model for MER, where MAG and TI are, for the first time, jointly formulated. We accelerate MAG and TI methods by exploring their underlying linear forms, which results in computationally efficiency and higher accuracy. Second, we design a spatial-temporal descriptor, which is named the tangent delta descriptor. Tangent delta reflects the intrinsic dynamics of facial movements and disregards the diversity, which is extrinsic to the facial movements and usually caused by different individuals and environments variations. To the best of our knowledge, this is the first time that differences between points on tangent spaces of Riemannian manifolds are used as descriptors for facial dynamics description. Third, we develop a hierarchical framework, which comprehensively...
Fig. 3. Illustration of the proposed motion modeling approach: (a) Through the consolidated motion elucidation module, the hidden motion within the input sequence becomes easier to observe; (b) the intrinsic dynamics between sequential frames are captured via point differences on the tangent space to a Riemannian manifold; (c) long-term dynamics are modeled using state-based sequential inference methods.

extracts the dynamics of movements from adjacent frames over the period of a facial movement occurrence. The proposed framework and its key modules are illustrated in Figure 3.

2 RELATED WORK

We briefly review relevant studies about characterizing subtle facial dynamic, especially in the context of micro-expression recognition [36, 63] and visual speech recognition [6, 93].

One way of modelling temporal dynamics is the sequential inference approach. Sequential inference approaches usually treat dynamics of facial movements as a sequence of particular observations (a.k.a., feature vectors extracted from frames), and recognize it via state-based sequential inference models. Various methods have been successfully applied to extract information from single images [49, 52, 65, 67]. State-based sequential inference models are then utilized to model the connections between sequential frames and model how likely a sequence belongs to a specific kind of movement, such as Hidden Markov Models (HMM) [33, 65, 66], dynamic Bayesian networks (DBN) [1, 51, 67], recurrent neural networks such as recurrent convolutional neural networks (RCNN) [39, 81], and Long short-term memories (LSTM) [27]. Nevertheless, a large amount of training samples are usually required to optimize plenty of model parameters to achieve accurate and stable description of any important movement within the sequence.

To capture dynamics of movements more directly, Spatial-Temporal Descriptors (STD) are popularly used. Some early works include temporal derivatives [9], optical flows [17, 41, 42, 46, 48], color tensors [76], and the spline representations in temporal axis [75]. Moreover, there has been a deluge of research on designing local spatial-temporal descriptors (STD) [11, 26, 28, 31, 32, 37, 53, 83, 86, 87], such as the local cuboid descriptors using both the 2D spatial filters and 1D temporal filters [11], the histograms of 3D gradient orientations [28], the space-time interest point descriptor [32], the covariance matrix-based representations to capture the dynamics of movements [16, 68, 71], and the histograms of the spatial-temporal local binary patterns [86, 87], their fast implementation [20], as well as plenty of variants [23–25, 35–37, 83]. Comprehensive reviews on local STDs in the task of human activity recognition are provided by [1, 74]. The rationale for STD approaches is that once the locations of cuboids are determined (either by interest points detection [31] or by dense sampling [86]), their local neighborhoods contain contextual information to characterize the dynamics of movements. Generally, most STDs are arranged in a bag-of-words manner to represent the distribution in sequences. More specifically, each feature vector is transformed to a unique code index through vector quantization (usually treated as the partition of the
local feature space). Frequency histograms of the codes are then used to represent corresponding sequences. These local spatial-temporal descriptions are robust against background clutters and camera movements. However, spatial-temporal relations among descriptors, which are crucial for non-periodic and complex dynamics analysis [1], are usually ignored. In addition, covariance matrix descriptors [21, 22, 70–73, 82] are common ways of modeling the spatial-temporal coherence on Riemannian manifolds. Reference [45] represents each video clip with three types of image set models. Then different Riemannian kernels are employed to fuse multi-level information for further boosting the performance. However, the distributions of points on manifolds belonging to the same kind of movements still vary as extrinsic variations exist. Besides, in Reference [44], on the basis of temporal alignment and semantics-aware dynamic representation, a kind of variation manifold modeling method is proposed to obtain a group of expressionlets from well-aligned spatio-temporal regions. More recently, learning-based descriptors, especially the deep-learning approaches, have also been applied to capture the subtle dynamics of movement [6, 7, 18, 19, 29, 30, 34, 38, 57, 69, 77, 96].

Subtle emotion analysis has also been studied in other emotion analysis tasks. Zhao et al. made remarkable progress on emotion distribution learning to tackle the subjectivity challenge of emotion perceptions [88, 89, 91]. In Reference [91], the emotion distribution are modelled as a Gaussian mixture model (GMM) and proposed multitask shared sparse regression to predict the parameters of GMM, which is the first investigation on continuous emotion distribution learning. In References [88, 89], multiple features are fused via a novel weighted multi-feature shared sparse learning algorithm to predict the probabilities of discrete emotion categories, which achieves the state-of-the-art performances.

Our approach offers a different way to extract dynamics of facial movements. It characterizes the subtle temporal connections by the difference between sequential points (i.e., cuboid descriptors) in the tangent spaces to the manifolds and captures long-term dynamics via the state-based sequential inference methods. It is thus expected to be more reliable and stable than just capturing temporal coherency at the pixel or cuboid descriptor level, like many of the previous methods do.

3 SUBTLE DYNAMIC ELUCIDATION

In this section, we describe the proposed consolidated motion elucidation model, MOTEL, to boost subsequent recognition and analysis tasks. Given a sequence \( V = [I(1), \ldots, I(T)] \), where \( I(t) \in \mathbb{R}^d \) denotes the \( t \)th frame of a video for \( t = 1, \ldots, T \), our goal is to produce a sequence \( \hat{V} \) where the subtle motion can be easily observed, as illustrated by Figure 1(c).

For achieving such a goal, a general model can be formulated by a video-specific function \( f \) with a set of parameters \( \Theta \) as follows:

\[
\hat{V} = f(V; \Theta).
\]

The function \( f : \mathbb{R}^{d \times T} \mapsto \mathbb{R}^{d \times T'} \) maps the input sequence to \( \hat{V} \) of a fixed-length \( T' \), where the subtle movements in \( V \) are elucidated.

Representative motion magnification and temporal interpolation methods, such as the Eulerian-based MAG method [80] and the graph spectral TI method [4], are non-linear in general cases. To decrease the difficulty in implementation, existing MER architectures [24, 36, 83] usually treat MAG and TI as two separate modules and deploy them successively. Obviously, such separation of the two modules does not take into consideration the peculiarity of subtle motion analysis in digital sequences. It thus raises a series of problems: First, it slows down the processing speed substantially. Second, it also increases the risks of accuracy loss caused by the side effects of intermediate process and storage, as indicated from Figure 1(a).
To the contrary, in this section, we revisit the nature of MAG and TI in the context of digital video analysis like micro-expression recognition, and explore the underlying relationship between MAG and TI. We will show that both MAG and TI can be jointly characterized by a linear system in a consolidated manner under certain circumstances.

The feasibility of getting the right hand side of Equation (1) into a linear form mainly comes from the following two aspects. First, though the Eulerian motion magnification model [43, 80, 84] is a complicated system as it is multi-scale and non-linear, we realize that it is feasible to implement it as a series of scale-wise linear functions. At each scale, it approximates movements using the first-order Taylor expansion. Second, we will show that the interpolated video can be obtained by multiplying the input video with a matrix, without explicitly solving the generalized eigenvalue problem as the graph spectral TI method [95] does. Thus, without loss of generality, in the following context, we formulate our model with a "single-scale" setting and consider $f(\cdot)$ as a linear function.\(^2\)

That is,

$$\hat{\mathbf{V}} = \mathbf{VW}. \tag{2}$$

To obtain the parameter matrix $\mathbf{W} \in \mathbb{R}^{T \times T'}$, we decouple it as follows:

$$\mathbf{W} = \Upsilon \Xi, \tag{3}$$

in which $\Upsilon \in \mathbb{R}^{T \times T}$ and $\Xi \in \mathbb{R}^{T \times T'}$ are the magnification matrix and the interpolation matrix, respectively. The order of the right-hand side of Equation (3) is consistent with the one in Reference [36], while a comprehensive investigation in the influence of order will be performed in Section 5.1.4. In what follows, we will provide the concrete solutions for $\Upsilon$ and $\Xi$, respectively.

### 3.1 Solution to the Magnification Matrix

In this subsection, we provide the solution to $\Upsilon$. For convenience, we introduce $\mathbf{V}_m = [I_m(1), \ldots, I_m(T)]$ to describe the intermediate magnified output.\(^3\) Then, we have $\mathbf{V}_m = \mathbf{VY}$, as per Equations (2) and (3). Let $\Upsilon_{i,j}$ be the entry of $\Upsilon$ at $(i, j)$ for $i, j = 1, \ldots, T$. The magnified frame $I_m(t)$ can be represented as a linear combination of the input ones:

$$I_m(t) = \sum_{i=1}^{T} I(i) \Upsilon_{i,t}. \tag{4}$$

To get the solution of $\Upsilon = \{\Upsilon\}$, we revisit the Eulerian magnification model (EVM) [80], which is the most popular MAG method in the arena of MER. It regards the first frame in a video as a digital signal $s(x)$. Thus (within a short period of time), a subsequent frame is regarded as a displacement of the signal, i.e., $I(x,t) = s(x + \delta(t))$, where $I(x,t)$ is the pixel intensity at location $x$ of the $t$th frame and $\delta(t)$ is a tiny displacement function. Then the first-order Taylor expansion is employed to form an approximation:

$$s(x + \delta(t)) = s(x) + \delta(t) \frac{\partial s}{\partial x},$$

which represents every frame in an ME video as an addition of the base signal and the multiplication of the displacement and the first derivative. The first-order component of the Taylor expansion $\frac{\partial s}{\partial x}$ describes the ME motion. In EVM, the motion is amplified through multiplying $\frac{\partial s}{\partial x}$

\(^2\)Multi-scale analysis can be easily reached by applying the model to each single scale.

\(^3\)The intermediate magnified output is inevitable in conventional methods as illustrated in Figure 1(a). However, such an intermediate result will be absorbed by Equation (3) to form a single $\mathbf{W}$ on the left-hand side of our MOTEL model finally, as illustrated in Figure 1(b).
by a magnification factor $\alpha$. As a result, the magnified signal becomes
\[ s_m(x + \delta(t)) = s(x) + (1 + \alpha)\delta(t) \frac{\partial s}{\partial x}. \]

On this basis, the motion $\frac{\partial s}{\partial x}$ is implemented by \textit{linear} interpolation of adjacent frames in Reference [80]. More concretely, it is fulfilled using a \textit{recursive} algorithm to take all the preceding frames into account and get the “historical” motion $B(t) = L_1(t) - L_2(t)$, where $L_1(t)$ and $L_2(t)$, for $t = 2, \ldots, T$, are auxiliary variables defined as follows:

\[
\begin{cases}
L_1(t) = w_1I(t) + (1 - w_1)L_1(t - 1) \\
L_2(t) = w_2I(t) + (1 - w_2)L_2(t - 1),
\end{cases}
\]

where the $L_1(1) = L_2(1) = I(1)$. The hyper-parameters $w_1$ and $w_2$ are in the range $(0, 1)$ and $w_1 > w_2$. Correspondingly, the magnified frame becomes
\[ I_m(t) = I(t) + \alpha B(t). \]

The EVM method [80] as shown in Equations (5) and (6) is recursive and thus complicated and inefficient in execution. However, we still find strong evidence that the core of the Eulerian magnification can be realized by a linear system, since the key components, including the first-order Taylor expansion, the introduction of $\alpha$, and the motion approximation module, are linear. Such insights thus contribute to breaking the barriers in implementation. As a result, we can solve out a parameter matrix operated on the input video by explicitly expanding the recursive form in Equation (5). By combining Equations (4) and (6), we have the solution for $\Upsilon$:

\[ \partial_{t,j} = \begin{cases} 
\alpha(1 - w_1)^aw_1^b - \alpha(1 - w_2)^aw_2^b & \text{if } j > i, \\
\alpha(w_1 - w_2) + 1 & \text{if } j = i, \\
0 & \text{for otherwise},
\end{cases} \]

where $a = j - i$ and $b = \min(1, i - 1)$. $\Upsilon$ is an upper triangular matrix, since ME motion is only related to the preceding frames in sequences.$^4$ The superiority in reducing computational costs of using an explicit solution of $\Upsilon$ by Equation (7) is apparent when compared with the recursive implementation like Equation (6) in Reference [80].

### 3.2 Solution to the Interpolation Matrix

In this subsection, we will derive $\Xi$ in Equation (3) to form our efficient temporal interpolation module for extending the intermediate magnification output $V_m$ with $T$ frames to a $T'$-frame elucidated video $\hat{V}$:

\[ \hat{V} = V_m \Xi. \]

The popular TI approach [92, 95] employs graph spectral [4] to find bijective functions by solving a generalized eigenvalue problem. By such functions, the input points (namely, the frames of the input video) are first projected into a low-dimensional latent space uniformly, where the topology of neighboring frame adjacency is maintained. The interpolated signal can then be reconstructed from the low-dimensional latent space by different sampling intervals. More concretely, $\hat{\xi}(\frac{t}{r}) = F(I_m(t))$, where the function $F: \mathbb{R}^d \mapsto \mathbb{R}^r$ is a projection from an image in the input space to a point $\hat{\xi}(\frac{t}{r})$ in the $r \leq T - 1$ dimensional latent space.$^5$ According to Reference [4], the optimal projection is a group of sine functions. Every frame thus corresponds to a series of discrete points in $\mathbb{R}^r$:

$^4$The factor $\alpha$ in Equation (7) may be further truncated to realize non-linear response situations as Reference [80].

$^5$W.l.o.g., we set $r$ to its maximum value, i.e., $T - 1$ to maintain a minimum reconstruction error.
\[
\left[ \xi\left(\frac{1}{T}\right), \xi\left(\frac{2}{T}\right), \ldots, \xi(1) \right] = \mathcal{A}(V_m) \left( V_m - \left[ \bar{I}_m \bar{I}_m \ldots \bar{I}_m \right] \right) T \text{ times},
\]

where \( \bar{I}_m \) is the average over the frames in video \( V_m \). \( \mathcal{A}(V_m) \) is a transformation matrix dependent on \( V_m \) to fulfill the map \( \mathcal{F} \). In Reference [4], \( \mathcal{A}(V_m) = (M^{-1}V^TU^T) \), where \( U \) is a unitary matrix by performing the singular-value decomposition (SVD) [15] on the input sequence. \( M \) and \( V \) are the graph Laplacian diagonal matrix and the one of corresponding eigenvectors, respectively. More details can be found in References [92, 95].

**Theorem 1.** \( V_m \) can be linearly transformed to the elucidated sequence \( \hat{V} \).

**Proof.** As \( \mathcal{F} \) is bijective, we can reconstruct the input by sampling from the latent space and projecting the points back to image space. According to Equation (9), the intermediate magnification video can be reconstructed by

\[
V_m = \left[ \mathcal{A}(V_m) \right]^{-1} Y + \hat{V}_m,
\]

where \( \hat{V}_m = V_m (\frac{1}{T} 1_{T \times T}) \). \( Y = [\xi\left(\frac{1}{T}\right), \xi\left(\frac{2}{T}\right), \ldots, \xi(1)] \) are \( T \) points in the latent space. Each \( \xi(t) \), which relates to one point in the input space (i.e., one frame), is sampled from the \( r \) sine functions.

As \( Y \) is known, Equation (10) suggests another way of computing \( \mathcal{A}(V_m) \) or \( \left[ \mathcal{A}(V_m) \right]^{-1} \), by applying the right inverse matrix to \( Y \):

\[
\left[ \mathcal{A}(V_m) \right]^{-1} = V_m \left( I - \frac{1}{T} 1_{T \times T} \right) Y^T (YY^T)^{-1}.
\]

Equation (11) ensures that \( \left[ \mathcal{A}(V_m) \right]^{-1} \) can be computed using a linear transformation matrix independent of \( V_m \). It thus plays an essential role in the following derivations. In contrast, \( \left[ \mathcal{A}(V_m) \right]^{-1} \) has to be obtained by applying SVD to \( V_m \), i.e., \( \left[ \mathcal{A}(V_m) \right]^{-1} = U(V^{-1})^T M \) in Reference [4], where the relation of \( \left[ \mathcal{A}(V_m) \right]^{-1} \) and \( V_m \) cannot be simplified as a linear transformation. \( V^{-1} \) exists as the matrix of corresponding eigenvectors \( V \) is full rank.

In an analogous manner, the TI method also provides an interpolation result for any point in the \( r \) dimensional latent space. Sampling \( T' \) groups of points, i.e., \( Y' = [\xi\left(\frac{1}{T'}\right), \xi\left(\frac{2}{T'}\right), \ldots, \xi(1)] \), the video is extended to \( T' \) frames. That is,

\[
\hat{V} = \left[ \mathcal{A}(V_m) \right]^{-1} Y' + \hat{V}_m,
\]

where \( \hat{V}_m(t) = V_m (\frac{1}{T} 1_{T \times T}) \). Since \( Y' \) is a group of discrete points, it satisfies the requirement for a linear TI model.

By substituting Equation (11) into Equation (12), we have

\[
\hat{V} = V_m \left( I - \frac{1}{T} 1_{T \times T} \right) Y'^T (YY'^T)^{-1} Y' + \hat{V}_m = V_m \left[ \left( I - \frac{1}{T} 1_{T \times T} \right) Y'^T (YY'^T)^{-1} Y' + \frac{1}{T} 1_{T \times T} \right].
\]

This concludes the proof of Theorem 1 and provides us with \( \Xi \).

The condition for Theorem 1 to hold is that both the lengths of the input \( V_m \) and output \( \hat{V} \) are limited and discrete-valued. Thus, \( Y \) and \( Y' \) exist and Equations (10) and (12) hold so that the final matrix \( \Xi \) can be deduced. Obviously, this condition is naturally satisfied in tasks of digital video processing and analysis, such as the micro-expression recognition and visual speech recognition tasks we focused on. \( \square \)
3.3 Solution to the Elucidation Matrix

We use Equation (3) to absorb $\Upsilon$ and $\Xi$, and get the elucidation matrix $W$.

As a result, the proposed consolidated model is associated with the only parameter matrix $W$. MOTEL is thus a unique model for both video magnification and temporal interpolation. Moreover, the relation of MOTEL to the magnification and interpolation method is clear. When $\Xi$ is an identity matrix, the model reformulated in Equation (2) is a MAG model. When $\Upsilon$ is an identity matrix, it degenerates to a TI model.

The proposed MOTEL model is computationally economic. First, its efficiency comes from the discovery of the underlying linear nature of the MAG and TI modules. Moreover, it avoids intermediate I/Os and complicated computation of graph Laplacian. In addition, expanding the recursive forms and avoiding decoupling model parameters from the input also contribute to the improvement of speed. Finally, the computational costs can be further substantially reduced in practice, since there exist parameters that are independent of the input. More specifically, Equation (7) indicates that as soon as the number of frames of the input sequences ($T$) and the magnification factor $\alpha$ are given, $Y$ is fully determined. By the same token, $\Xi$ in Equation (13) is fixed once the lengths of the input and output sequences (i.e., $T$ and $T'$) are known. As a result, for acceleration purposes, the elucidation matrix $W$ on popular settings of the pair of ($T$, $T'$) and frequently used values of $\alpha$ can be computed in advance and stored by, for example, a look-up table (LUT) prior to the arrival of the input. It is also worth noting that the previous solution to $\mathcal{A} (V_m)$ has to be explicitly dependent on the singular value decomposition of the input video in the spectral graph derived solution [4]. It is thus impossible to perform any computation prior to the input video. Instead, our solution ensures that both $\Upsilon$ and $\Xi$, and again $W$, can be computed before being applied to the input video.

4 DYNAMIC DESCRIPTION

As indicated by Figure 2, the differences $\Delta^{(1)}$ and $\Delta^{(2)}$ on tangent spaces characterize the intrinsic dynamics of movements by disregarding the variations caused by the characteristics of the individuals and environments. We thus utilize them as our subtle dynamic descriptors. The following subsection details the formulation leveraging the Riemannian framework of matrix manifolds [47, 62, 72].

Let $\mathcal{R}(d)$ denote the manifold of $d \times d$ Symmetric Positive Definite (SPD) matrices. For any $M \in \mathcal{R}(d)$, there exists a tangent space $T_M \subset \mathbb{R}^{d \times d}$ (at $M$), which is a vector space of symmetric matrices (for instance, the planes $T_{M_1}$ at $M_1$, in Figure 3(b)). To calculate the difference between two points $M_1$ and $M_2$, under the affine-invariant metric, we map $M_2$ to $T_{M_1}$ (the tangent space of $M_1$) and get its projection $T_2$ via the logarithm map, which is defined as the inverse of the exponential map. More specifically, given $M_1 \in \mathcal{R}(d)$, for any $M_2 \in \mathcal{R}(d)$, the logarithm map $\Psi_{M_1} : \mathcal{R}(d) \to T_{M_1}$ is given as follows:

$$\forall M_1 \in \mathcal{R}(d), \quad \Psi_{M_1} (M_2) = Q \left( \Psi \left( Q \left( M_2, M_1^{-1/2} \right) \right), M_1^{1/2} \right),$$

(14)

where $M_1^{1/2}$ is the symmetric matrix square root of $M_1$ and $Q (A, B) = B^T A B$. We define the directed differences $\Delta$ between two SPD matrices $M_1$ and $M_2$ as

$$\Delta (M_1, M_2) \triangleq \Psi_{M_1} (M_2) .$$

(15)

We note that $\Delta (M_1, M_2) = \Psi_{M_1} (M_2) - \Psi_{M_1} (M_1)$ as $\Psi_{M_1} (M_1) = 0$, resembling the differences in Euclidean spaces. $\Delta (M_1, M_2)$ is abbreviated as $\Delta_{1,2}$ for clarity hereafter.

---

$^6$Not to be confused, the maps $\exp (\cdot)$ and $\Psi (\cdot)$ are ordinary matrix exponential and logarithm operators and $\exp_M (\cdot)$ and $\Psi_M (\cdot)$ are point specific operators, respectively.
The difference $\Delta$ calculated through Equation (15) reduces the diversity of $M_i$ or $M_{i+\Delta t}$ caused by extrinsic variations such as illumination changes and subject differences. Moreover, it declines the accumulation of errors in extracting information of the observations along the temporal direction. Furthermore, it preserves the chronological order of frames, which is highly related to the dynamics of movements, since the difference $\Delta$ by subtraction on Riemannian manifolds using Equation (15) does not satisfy the commutative property, i.e., $\Delta (M_i, M_{i+\Delta t}) \neq \Delta (M_{i+\Delta t}, M_i)$. The resulting differences therefore characterize the consecutive connections from one frame to its next frames.

4.1 Problems

Despite the potential of using Riemannian manifold difference as descriptors, there arises a critical theoretical problem. The set of direct differences $\{\Delta_i\}$ calculated by Equation (15) for different pairs of points in sequences are going to be on different tangent spaces to the manifold. The problem is that one cannot compute metrics between tangent vectors across different tangent spaces directly. To address this problem, in Reference [68], parallel transport is applied to move all tangent vectors to a vector space. A (warping) rate-invariant distance is further defined and an algorithm for calculating the mean curve of a group of curves is used. On this basis, one can categorize a given unknown curve to the class, the mean curve of which has the shortest distance to the given one [68]. As a result, the rate-invariant method solves the problem in theory, by defining the rate-invariant distance, termed the transported square-root vector field (TSRVF), and developing an algorithm to calculate the mean curve on Riemannian manifold. Nevertheless, it appears with low accuracy in practice, as shown by the experimental results in Reference [68].

With a careful analysis, we realize that there are three main limitations for the TSRVF method. First, the smooth assumption, which the rate-invariant method requires, does not hold well in practice. More concretely, there is one strong assumption that the trajectory $\beta [t]$ for $t \in \mathbb{R}$ should be smooth enough (i.e., $C^\infty$ smooth) the calculation of the rate-invariant distance in Reference [68], as illustrated by Figure 4(a). However, in practice, especially when subtle motion occurs, we only obtain a sequence of discrete points: $\beta [1], \ldots, \beta [N]$, as illustrated by Figure 4(b). Second, the TSRVF method directly uses the trajectory as a whole, which is too deterministic to handle the variations at each time step. The rate-invariant distance is sensitive to any perturbation. Figure 4(c) provides an intuitive example. When there is perturbation that causes points $b_2$ and $b_3$ on $\beta (t)$ to move away from their positions to $b'_2$ and $b'_3$ on $\beta' (t)$, there will be a distinct rate-invariant distance between these two curves, though in fact they belong to the same class. Finally, its core step, namely, the rate-invariant distance calculation, relies on the dynamic programming (DP) algorithm to solve and thus it is computationally expensive.

4.2 Solution

The problems of existing manifold trajectory approaches mainly come from the deterministic modeling manner. To reflect the dynamics of movements efficiently over a longer period of time, it is thus more reliable to merge the subtle dynamic statistically rather than to use the trajectory as whole in a deterministic manner. As a result, we propose to incorporate the subtle dynamic descriptor with the sequential inference technique with statistical observation models. The distribution of points on manifolds is usually too complicated to be modeled by a single regular Gaussian distribution [5, 61]. Thus, finite mixture models, such as the mixture of Gaussians (moG), is used, since potentially arbitrarily complex probability density functions can be captured given enough Gaussians [13].

We compute our subtle dynamic descriptor by parallel transporting the tangent differences between two SPDMs to the tangent space of a common reference point on a Riemannian manifold. A
Fig. 4. Illustration of the problems of TSRVF, a typical Riemannian manifold-based trajectory approach. (a) There is a strong trajectory-smooth assumption made by TSRVF. (b) However, in practice, only sequences sampled at discrete time points are available. (c) The existing TSRVF method is sensitive to the perturbation caused by variations.

natural choice of the reference point is the identity matrix. More concretely, for any $M_1$ whose determinant equals 1, given a difference $\Delta (M_1, M_2)$ on the tangent space of $M_1$, the parallel transport $U$ of $\Delta (M_1, M_2)$ from $M_1$ to the identity matrix $I$ is given by [85] as follows:

$$\hat{\Delta}_{1,2} = U_{M_1 \rightarrow I} (\Delta (M_1, M_2)) = M_1^{-1} \Delta (M_1, M_2).$$

(16)

To further reduce the computational cost, we adopt the differences on the tangent space of the identity matrix to approximate the parallel transported differences on the tangent space of the identity matrix in implementation.

The proposed dynamic description method has the following advantages. First, we significantly relax the strong trajectory-smooth assumption of the TSRVF method. As a result, our method is able to model discrete time-indexed states and is thus appropriate to model subtle motion in sequences; Moreover, it eliminates the common variations caused by subjects and illuminations in one sequence and further takes into account the variation of observations by modelling them by the Gaussian mixture model within sequential inference methods.

4.3 Implementation

To characterize the temporal coherence of movements comprehensively, we design a principled hierarchical framework, as illustrated in Figure 5.

As shown in Figure 5(a), the 2D spatial filters and 1D temporal filters extract local spatial-temporal connection from adjacent frames. In particular, ten features are adopted to represent a point within a sequence, including: the 3D coordinates $(x, y, t)$; the intensity $I$; the absolute values of the first and the second derivatives in both horizontal and vertical directions, respectively, i.e., $\left| \frac{\partial I}{\partial x} \right|$, $\left| \frac{\partial I}{\partial y} \right|$, $\left| \frac{\partial^2 I}{\partial x^2} \right|$, and $\left| \frac{\partial^2 I}{\partial y^2} \right|$; and the first and the second derivatives along the temporal direction, $\frac{\partial I}{\partial t}$ and $\frac{\partial^2 I}{\partial t^2}$. All 2D spatial filter responses are unsigned whereas the 1D temporal filter responses are signed.

Like many local STD approaches [11, 26, 28, 31, 32, 53, 86, 87], a sequence containing an execution of an movement is regarded as a particular 3D ST volume constructed by concatenating 2D (XY) spatial images along the time axis (T). Each of the 3D ST volumes is divided into several overlapping segments, i.e., cuboids, each of which includes $L$, a small number of consecutive frames. As shown in Figure 5(b), a covariance matrix $M$ is then employed to capture the spatial temporal correlation of the ten features selected over all pixels within a cuboid $Q$:

$$M = c \sum_{x \in Q} (f - \mu)(f - \mu)^T,$$

(17)
where \( x = [x, y, t]^T \), \( \mu \) is the mean of feature vectors within \( Q \), and \( c \) is a normalization factor. To achieve enhanced robustness, especially against the diversity among different capturing environments, \( M \) is further normalized to a correlation matrix, having only unit variances in diagonal.\(^7\) It is commonly used to achieve enhanced robustness in consideration that the normalization is equivalent to the 0-mean-1-variance normalization on the (10-D) feature vectors selected.

We use the tangent delta descriptor by Equation (16) to capture subtle dynamics of facial movements. \( \hat{\Delta}_{1,2} = \mathcal{U}_{M_1 \rightarrow I} (\Delta (M_1, M_2)) \) can be calculated between two points referring to any two frames that are not necessary adjacent. One feasible way is to calculate the differences between two points with a certain short-time interval \( \Delta t \) and a spatial step \( \Delta s \), then obtain a sequence of differences, \( \{ \Delta_1, 1+\Delta t, \Delta_1+\Delta s, (1+\Delta s)+\Delta t, \ldots \} \) before parallel transport.

Finally, we use the continuous hidden Markov model [33] for sequential inference, by utilizing the first-order Markov assumption. In specific, the Baum-Welch algorithm is exploited to determine the parameters that best fit the given training observation sequences belonging to a specific class of movements. During evaluation, the Viterbi algorithm is employed to calculate the probability that a model generates the given testing sequence \( O_{Tst} \). For classification, the class label \( \hat{C} \), which maximizes the posterior probability of \( O_{Tst} \), is then assigned to \( O_{Tst} \).

5 EXPERIMENTAL RESULTS

In this section, the dynamic elucidation model is evaluated on the MER task, which is described in Section 5.1. The dynamic description method and the whole framework are validated in Section 5.2.

5.1 Evaluation of Dynamic Elucidation

We evaluate the proposed dynamic elucidation model on two widely used spontaneous ME databases: SMIC [37] and CASMEII [83]. The SMIC database contains 306 ME video clips belonging to four sub-sets, which are SMIC-HS, SMIC-VIS, SMIC-NIR, and SMIC-subHS. There are three classes for all the ME clips: Positive, Negative, and Surprise. The CASMEII database contains 247 ME video clips. The baseline method of CASMEII chooses classes Happiness, Surprise, Disgust, Repression, and Others for training and testing.

To evaluate the proposed MOTEL extensively, we perform four experiments: (1) consolidation versus separate; (2) computational cost evaluation; (3) comparisons to the state of the arts; and

\(^7\)Singular CMs are avoided by adding a small positive figure to the diagonal entries of CMs as Reference [71].
Fig. 6. Accuracy comparisons under the LOSO protocol on the SMIC-HS dataset for micro-expression recognition using three mainstream feature descriptors: (a) HIGO-TOP, (b) HOG-TOP, and (c) LBP-TOP.

Table 1. Computational Cost Comparison Between our Model and Li et al. [36]

|                     | SMIC-HS |     | CASMEII |     |
|---------------------|---------|-----|---------|-----|
|                     | Li et al. [36] | Ours | Li et al. [36] | Ours |
| Time (in seconds)   | 125.5 s | 27.4 s | 420.9 s | 41.6 s |

(4) influence of the order. By grid searching different values of the hyper-parameters $T'$, $\alpha$, $w_1$, and $w_2$, they are set to 10, 16, 0.4, and 0.05, respectively, which are consistent to those fixed in [36, 80]. As suggested by [36], unless further specified, a linear support vector machine (SVM) [3] is used as classifiers\(^8\) in all experiments. The Leave-One-Subject-Out (LOSO) protocol is adopted, and the micro-average accuracy [2] is employed as evaluation metrics.

5.1.1 Consolidation Versus separate: Accuracy Test. We compare MOTEL with the current best method [36], which contains both the MAG and TI modules but treats them as two separate ones, as illustrated in Figure 1(a). For fair comparison, we use the same classifier and three mainstream spatial-temporal feature descriptors: LBP-TOP [87], HOG-TOP [8], and HIGO-TOP [36], which are extensively evaluated and compared with the state of the art [36].

The comparative results are reported in Figure 6. There are several important observations. (1) For the HIGO-TOP descriptor, the accuracy is improved by almost 10.0% on the SMIC-NIR dataset. On the SMIC-subHS dataset, our accuracy is 87.32%, while the current best is only 80.28%. (2) For the HOG-TOP descriptor, our model improves the accuracy by 3.0% on average. (3) For the LBP-TOP descriptor, MOTEL is slightly better than the baseline results of SMIC, while it improves by almost 4.5% on the CASMEII. (4) The largest improvements in accuracy are almost 13%, 6%, and 4.5% with the HIGO-TOP, HOG-TOP, and LBP-TOP descriptors, respectively. From the results, it is safe to conclude that our model achieves superior accuracy to the current best [36], regardless of the video descriptors chosen.

5.1.2 Consolidation Versus separate: Computational Cost Evaluation. We compare the computational cost of MOTEL to the current best work, HIGO-TOP [36]. Both MOTEL and HIGO-TOP use the descriptor of HIGO-TOP and the SVM classifier. The only difference lies in the pre-processing module. In MOTEL, MAG and TI are performed in a consolidated manner by Equation (2), whereas HIGO-TOP performs MAG and TI separately. The experiments are implemented using MATLAB R2017b platform on a desktop with an Intel i5-6500 (3.20 GHz) CPU and a RAM of 8 GB. Results in Table 1 show that our model is more efficient on both SMIC-HS and CASMEII databases. More

---

\(^8\)The LIBSVM package from www.csie.ntu.edu.tw/~cjlin/libsvm is used.
Table 2. Experimental Results
Under the LOSO Protocol on
SMIC-HS Dataset for
Micro-expression Recognition

| Methods            | ACC   |
|--------------------|-------|
| Baseline [37]      | 48.80%|
| Liong et al. [40]  | 53.56%|
| Huang et al. [24]  | 60.98%|
| DiSTLBP-RIP [23]   | 63.41%|
| CNN-Sel [57]       | 53.60%|
| RCNN [81]          | 57.10%|
| HIGO-TOP [36]      | 68.29%|
| MOTEL (Ours)       | 68.90%|

Table 3. Accuracy Comparisons Under the LOSO Protocol
on the SMIC-VIS, SMIC-NIR, and SMIC-subHS Datasets
for Micro-expression Recognition

| Methods     | SMIC-VIS   | SMIC-NIR   | SMIC-subHS |
|-------------|------------|------------|------------|
| Baseline [37]| 52.10%     | 38.00%     | —          |
| CNN-Sel [57]| 56.30%     | —          | —          |
| HIGO-TOP [36]| 81.69%    | 67.61%     | 80.28%     |
| MOTEL (Ours)| **83.10%** | **69.01%** | **87.32%** |

specifically, the running time on SMIC-HS for all 164 video clips is 125.5 seconds (s) for Li et al. [36] and 27.4 seconds for our model. On the CASME II dataset, for the 247 video clips, the total running time for the two methods is 420.9 and 41.6 seconds, respectively. Our model is thus approximately ten times faster than Reference [36].

5.1.3 Comparison to the State of the Arts. The current best method [36] provides a standard pipeline for MER tasks. As the HIGO-TOP descriptor is recommended by Reference [36], we also use it as the video descriptors.

We compare our method to eleven state-of-the-art approaches, including three deep-learning methods, including the pre-trained CNN with feature selection (CNN-SEL) [57], the pre-trained CNN on apex frame (CNN-Apex) [38], and recurrent CNN (RCNN) [81], and eight state-of-the-art hand-crafted feature approaches [23, 24, 36, 37, 40, 56, 76, 83]. The results are summarized in Tables 2, 3, and 4. For all the databases, our method achieves the best accuracy against all the other methods. More specifically, the accuracy on SMIC-HS, SMIC-subHS, SMIC-VIS, SMIC-NIR, and CASMEII are 68.90%, 87.32%, 83.10%, 69.01%, and 70.85%, respectively, while the second-best results achieved by Reference [36] are only 68.29%, 80.28%, 81.69%, 67.61%, and 67.21%, respectively. It indicates that the proposed MOTEL is robust to the variations in appearance brought by different sensors.

The comparison results listed in Tables 2 and 4 imply that recent deep-learning methods such as References [38, 57, 81] are still inferior to the state-of-the-art handcrafted features in the task of MER under the leave-one-subject-out protocol. That is mainly because the small amount of training data limits the capability of the neural networks. Hence, there are also other deep-learning
Table 4. Comparisons Under the LOSO Protocol on the CASMEII Dataset for Micro-expression Recognition

| Methods            | ACC    |
|--------------------|--------|
| Baseline [83]      | 63.41% |
| Adaptive Mag [56]  | 51.90% |
| TICS [76]          | 62.30% |
| Huang et al. [24]  | 64.37% |
| DiSTLBP-RIP [23]   | 64.78% |
| CNN-Sel [57]       | 47.30% |
| CNN-Apex [38]      | 63.30% |
| RCNN [81]          | 65.80% |
| HIGO-TOP [36]      | 67.21% |
| MOTEL (Ours)       | 70.85% |

Table 5. Computational Costs Evaluation of Using Different Orders to Decouple W

| Orders | SMIC-HS  | CASMEII |
|--------|----------|---------|
| $I \circ M$ | 27.29 s  | 46.63 s |
| $M \circ I$ | 27.42 s  | 41.61 s |

Table 6. Accuracy Evaluation of Using Different Orders to Decouple W

| Orders | SMIC-HS | SMIC-subHS | SMIC-VIS | SMIC-NIR | CASMEII |
|--------|---------|------------|----------|----------|---------|
| $I \circ M$ | 52.44%  | 78.87%     | 77.46%   | 59.15%   | 59.92%  |
| $M \circ I$ | 68.90%  | 87.32%     | 83.10%   | 69.01%   | 70.85%  |

MER methods, which try to solve the small-sample-size problem by using additional training data. For example, Peng et al. [59] mix up the samples from the CASME and CASME II datasets to perform the leave-one-subject-out test. There are also works that perform leave-one-sample-out tests, rather than the recommended leave-one-subject-out tests, to have more samples in the learning state [50]. As the experimental protocols are different, we cannot compare with these works directly. Therefore, we leave them out of the experiments for fair comparisons.

5.1.4 Influence of the Order. We investigate the influence of swapping the decoupling order to compute $W$ in Equation (3). More concretely, $M \circ I$ indicates that $W$ is computed with the exact order with Equation (3), i.e., magnification first and then interpolation, while $I \circ M$ signifies that $W$ is computed in an inverse order, i.e., $W = \Xi \Upsilon$. First, for speed comparison, the results are reported in Table 5, with all the experimental settings the same to Section 5.1.2. It can be easily found that the order has little influence on speed, and they are both much faster than the method in Reference [36]. Second, we compare the accuracy of $M \circ I$ and $I \circ M$ over five datasets. The results listed in Table 6 indicate that $M \circ I$ substantially outperforms $I \circ M$. Specifically, on all the databases, the $M \circ I$ wins the accuracy by over 10% in average, with 16-point lead on SMIC-HS.
as the highest. From the results in Tables 5 and 6, we may safely reach the following observation: Though changing the order from $M \circ I$ to $I \circ M$ influences little on speed, it leads to a drastic drop of accuracy. The main reason is that in the $I \circ M$ order, the artifacts introduced by temporal interpolation would be amplified further by the following magnification module. The resulted sequences thus become seriously distorted.

5.1.5 Discussion. It is a bit surprise at first glance to observe that MOTEL also improves accuracy in the experiments. More concretely, when compared with the method with separate deployment of MAG and TI followed by the HIGO-TOP descriptors [36], MOTEL increases the accuracy by 0.61%, 1.41%, 1.40%, 7.04%, and 3.64% on the SMIC-HS, SMIC-VIS, SMIC-NIR, SMIC-subHS, and CASMEII datasets, respectively. Nonetheless, on closer consideration, it should not be unexpected. There are mainly two reasons for MOTEL to improve accuracy: First, a consolidated framework avoids unnecessary separation. Thus, it is less likely to introduce side effects like the diluted problem or accuracy loss due to intermediate processing. For instance, previous models like Reference [36] have to rescale the intermediate output to a range of $[0, 255]$ for visualization purpose and thus inevitably lead to the loss of data accuracy. Second, The Euler magnification method [80] adopted by other MER methods uses a bandwidth truncation that is designed for general cases of motion magnification. Such a common bandwidth truncation may decrease the accuracy in the MER task. In contrast, by controlling the process of magnification to a reasonable extent via setting the magnification factors, we can remove the bandwidth truncation step in Reference [80], which is well supported by the enhanced accuracy as shown in the experimental results.

5.2 Evaluation of Dynamic Description
In this section, we take visual speech recognition, a.k.a., lip-reading, which is to “read” the words uttered by speakers visually, as an example to evaluate the proposed approach on characterizing the dynamics of facial movements. We carry out experiments on the commonly used OuluVS dataset [86]. The OuluVS dataset includes 20 persons, each of whom pronounces ten phrases from once to nine times. All image sequences are pre-segmented, with the mouth regions determined by manually labeled eye positions in each frame.

There are two evaluation protocols commonly adopted [86]: the Speaker-Independent Test (SIT) by using the leave-one-subject-out protocol and the leave-one-sample-out (LOSMPO) cross-validation test. SIT is usually more challenging than LOSMPO, since all sequences from the speaker under evaluation are excluded in the learning stage in SIT. We thus pay more attention to the speaker-independent test to evaluate the robustness of the proposed approach against extrinsic variations caused by different speakers. More concretely, in each run, testing is performed on all sequences from one particular subject, whereas training is done on the sequences from the remaining 19 ones. Two measurements are employed to quantify the performance: the micro average accuracy over all experimental rounds (ACC), i.e., $\frac{\#\text{correctSeq}}{\#\text{totalSeq}}$; and the macro Average Mean Rank (AMR) over all experimental rounds of the correct rank for each testing phrase.

5.2.1 Implementation. All experiments carried out in this section are implemented using MATLAB. We use the software toolkits,\textsuperscript{10,11} for CHMM and the dynamic Bayesian networks, respectively.

SPDMs are calculated by Equation (17) to represent either frames or cuboids on Riemannian manifold. We set the the divisions in horizontal and vertical axes to 1 to maintain a low dimension and

\textsuperscript{9}LOSMPO separately evaluates the sequences of each speaker, alleged as subject dependent test in Reference [86].

\textsuperscript{10}http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html.

\textsuperscript{11}https://code.google.com/p/bnt/.
avoid employing any additional dimension reduction method, such as Principal Components Analysis (PCA) [65]. As a result, each sequence is divided along the temporal axis into cuboids with a frame length of \( L = 3 \) and with an overlapping step of 1; and set both \( \Delta t \) and \( \Delta s \) to 1. We normalize SPDMs to corresponding correlation matrices. Vectorization are performed for correlation matrices and the proposed tangent delta descriptors by Equation (16) to be used in sequential learning and inference. More specifically, correlation matrices are vectorized as feature vectors by selecting all off-diagonal upper triangular entries, while the tangent delta descriptors are vectorized by selecting all upper triangular entries. The resulting feature representations are then normalized by constraining their L2 norms to be 1. We simply set \( M \), the number of mixtures of moG, to 10 and \( N \), the number of states for CHMM, to 8 in the following experiments.

### 5.2.2 Ablation Study

Five ablations are performed to show the effect of different components in our framework. We compare five descriptors as the observations of CHMMs: (1) frame-level covariance matrices (\( M_7 \)) by removing \( t \), \( \frac{\partial I}{\partial t} \), and \( \frac{\partial^2 I}{\partial t^2} \) from the ten features in Section 4.3; (2) frame-level CMs (\( M_9 \)) removing \( t \); (3) cuboid-level CMs \( M(Q) \) using all ten features listed in Section 4.2; (4) cuboid-level tangent delta descriptor \( \hat{\Delta} \) by Equation (16); and (5) \( \hat{\Delta}^* \) following our motion elucidation component MOTEL with \( T = 40 \) and \( \alpha = 1.0 \).

Table 7 lists the accuracy and micro average mean rank with the corresponding normalization scheme (Norm.), and dimension of the descriptor (Dim.). There are several important observations:

1. Local ST connection among adjacent frames brought by the 1D temporal filters in \( M_9 \) brings substantial improvement to that only using local spatial filters \( M_7 \); (2) the local cuboid descriptor \( M(Q) \) is able to improve ACC by over 10%; (3) the proposed tangent delta descriptor \( \hat{\Delta} \) enhances robustness to non-intrinsic variations and increases ACC to over 88%; and (4) incorporation of the motion elucidation step MOTEL \( \hat{\Delta}^* \) further improves the accuracy to 94.1. It again supports to show that the proposed MOTEL does improve the performance compared with using the raw data.

### 5.2.3 Comparative Results with State of the Art

We compare our method with state-of-art approaches in VSR. The experimental results of spacial-temporal descriptors are directly from [86, 94]. We also implement commonly used appearance descriptions in visual speech recognition such as Discrete Cosine Transform (DCT) [65], Histogram of Local Binary Patterns (LBP) [87], and Histograms of Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) [86]. The dimensions of DCT, LBP and LBP-TOP are all reduced to 75 via PCA for fair comparisons. The descriptors, either frame-level or cuboid-level, are then fed into CHMMs for sequential inference. Likewise, we implement the articulated feature-based DBN approach [67], whereas the six articulated features are learnt via six radial basis function kernel support vector machines on the vector spaces of DCT, respectively. Moreover, we also compare with the representive manifold-based trajectory methods: the transported square-root vector field (TSRVF) [68] and random forest manifold alignment (RFMA) [58]; the learning-based descriptors, like Ong and Bowden [54]; and deep-learning methods, including 3D Convolution with Multiple Towers (MT) [7].

| Desc. | \( M_7 \) | \( M_9 \) | \( M \) (\( Q \)) | Tangent delta \( \hat{\Delta} \) | Tangent delta w. MOTEL \( \hat{\Delta}^* \) |
|-------|----------|----------|----------------|----------------|----------------|
| Norm. | L2       | L2       | L2             | L1             | L2             |
| Dim.  | 21       | 36       | 45             | 55             | 55             |
| ACC   | 20.07%   | 68.4%    | 78.9%          | 88.7%          | 94.1%          |
| ARM   | 4.07     | 1.83     | 1.61           | 1.22           | 1.17           |
Table 8. Comparative Results on the OuluVS Dataset for Visual Speech Recognition

| Method       | Feature | Model    | SIT  | LOSMPO |
|--------------|---------|----------|------|--------|
| DCT          | HMM     |          | 48.8%| 2.31   |
| DCT          | DBN     |          | 52.3%| 2.44   |
| LBP          | HMM     |          | 31.7%| 3.20   |
| LBP-TOP      | HMM     |          | 71.2%| 1.58   |
| Zhao et al.  | SVM     |          | 62.4%| N/A    |
| Zhou et al.  | MKL     |          | 92.8%| N/A    |
| Zhou et al.  | SVM     |          | 85.6%| N/A    |
| RFMA         | SVM     |          | 89.7%| N/A    |
| TSRVF        | NN      |          | —    | 70.6%  |
| Ong & Bowden | SVM     |          | 65.6%| N/A    |
| MT           | CNN     |          | 91.4%| N/A    |
| Ours         | HMM     |          | 94.1%| 1.17   |

Table 8 reports the comparison results of SIT. The proposed methods achieve an ACC of 94.1%, i.e., a substantial improvement (over 23%) over other HMM/DBN-based visual speech reading systems. In addition, our method outperforms other methods using SVM as classifiers. The one closest to our method by Zhou et al. [94], adopts cuboid-level LBP-TOP histograms and multiple kernel learning, and results in a much higher dimension than ours. Furthermore, the proposed method also outperforms the deep-learning-based solution like MT. The reason mainly lies in the small size of datasets for subtle dynamic analysis research. Nevertheless, it still well demonstrates that our method is efficient for capturing subtle dynamics.

Besides the subject-independent test, we carry out LOSMPO (leave-one-sample-out) cross validation as well for comprehensive comparisons, as reported in Table 8. Our approach still defeats other methods and achieves the best result. The articulatory features-based DBN method [67] gets poor performance in LOSMPO. Moreover, it is not surprising that the TSRVF method achieves inferior performance. The result manifests the analysis in the limitations of TSRVF in Section 4.1. Our hierarchical motion description framework outperforms TSRVF by over 24%. It thus demonstrates that our method provides a feasible solution to them.

The experimental results in this section validate our subtle dynamic descriptor and the hierarchical motion capturing framework. Besides its promising results in accuracy, the proposed cuboid-level descriptor only has a dimension of 55 due to the symmetric property. There are therefore three benefits brought by low dimension: (1) low storage requirements; (2) reduced computational complexity; (3) no dimensionality reduction method, such as PCA, is required to make stable training of CHMM feasible given a limited number of training samples. Using PCA as dimension reduction component brings additional computational costs, especially in the leave-one-sample-out test, since for each testing sequence we have to re-calculate the principled components of the subspace covering the remaining samples in each round of the LOSMPO test. In contrast, low dimension of the proposed representation method avoids such computational consumption.

Moreover, the comparisons with other frame-level or cuboid-level descriptions in Section 5.2.3 also support our assumption that the dynamics of facial movements could not be fully described in a low-dimensional vector spaces. The necessity of representing the subtle dynamics of facial movements on Riemannian manifolds is therefore emphasized.
6 DISCUSSION AND CONCLUSIONS

This article provides a solution for subtle facial movements modeling from both the aspects of motion elucidation and description. First, we propose a consolidated framework, MOTEL, to elucidate hidden movements. MOTEL amplifies subtle facial movements and expands temporal duration in a united manner. It is online and training-free. Moreover, the MOTEL model is computationally economic. (1) Its efficiency originates from the discovery of the underlying linear nature of the MAG and TI modules. (2) It avoids intermediate I/Os and computation of graph Laplacian. (3) Expanding the recursive forms and avoiding decoupling model parameters from the input also contribute to the improvement of speed. (4) The computational costs can be further reduced by using techniques like look-up tables, since the transformation matrix of MOTEL is independent of the input and thus can be computed and stored in advance. Furthermore, it is a bit surprise at first glance to observe that MOTEL also improves accuracy in the experiments. Nonetheless, on closer consideration, it should not be unexpected. The proposed consolidated framework avoids unnecessary separation of modules. Thus, it is less likely to introduce side effects like the diluted problem or accuracy loss due to intermediate processing.

Second, we provide a hierarchical video representation to characterize the dynamics of facial movements. We design the tangent delta descriptor to describe the subtle movements, which shows its superiority in diminishing the influences of extrinsic diversity caused by different individuals and various environments. We then model long-term dynamics by incorporating tangent delta with statistical observation models within the sequential inference approaches. Our solution relaxes the trajectory-smooth assumption made by conventional manifold-based trajectory modeling methods. As a result, the combination of subtle dynamic descriptors with statistical observation models resists the perturbation of points in the trajectory, which is caused by, for example, noise-like extrinsic variations, and thus enhances the robustness.

The experimental results show the potential of the proposed facial motion modeling approach for the tasks of affective computing and human computer interaction. We plan to apply it to other computer vision tasks such as normal facial expression recognition and human action recognition. Moreover, how to combine it with cutting-edge deep-learning methods may also be an interesting route in future work.

ACKNOWLEDGMENTS

We appreciate the insightful discussion with Prof. Anuj Srivastava about the method of TSRVF and Prof. Tong Zhang about parallel transport.

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo. 2011. Human activity analysis: A review. ACM Comput. Surv. 43, 3, Article 16 (Apr. 2011), 16:1–16:43.
[2] Stephan Bloehdorn and Andreas Hotho. 2004. Text classification by boosting weak learners based on terms and concepts. In Proceedings of the 4th IEEE International Conference on Data Mining (ICDM’04). IEEE, 331–334.
[3] Christopher J. C. Burges. 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2, 2 (1998), 121–167.
[4] Peter J. Burt and Edward H. Adelson. 1987. The laplacian pyramid as a compact image code. In Readings in Computer Vision. Elsevier, 671–679.
[5] Yasuko Chikuse. 2006. State space models on special manifolds. J. Multivar. Anal. 97, 6 (2006), 1284–1294.
[6] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR’17). 3444–3453.
[7] Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In Proceedings of the Asian Conference on Computer Vision. Springer, 87–103.
[8] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE.
Characterizing Subtle Facial Movements 94:21

[9] Navneet Dalal, Bill Triggs, and Cordelia Schmid. 2006. Human detection using oriented histograms of flow and appearance. In Proceedings of the European Conference on Computer Vision (ECCV’06).

[10] Adrian K. Davison, Cliff Lanesly, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. 2018. Samm: A spontaneous microfacial movement dataset. IEEE Trans. Affect. Comput. 9, 1 (2018), 116–129.

[11] P. Dollar, V. Rahaud, G. Cottrell, and S. Belongie. 2005. Behavior recognition via sparse spatio-temporal features. In Proceedings of the International Conference on Computer Communications and Networks (ICCCN’05).

[12] Paul Ekman and Erika L. Rosenberg. 1997. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System. Oxford University Press.

[13] Mario Figueiredo and Anil Jain. 2002. Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 24, 3 (2002), 381–396.

[14] M. G. Frank, Malgorzata Herbasz, Kang Sinuk, A. Keller, and Courtney Nolan. 2009. I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In Proceedings of the Annual Meeting of the International Communication Association.

[15] Gene H. Golub and Charles F. Van Loan. 2012. Matrix Computations. Vol. 3. JHU Press.

[16] Yusuf. Hakan Habiboglu, Osman Günay, and A. Enis Çetin. 2012. Covariance matrix-based fire and flame detection method in video. Mach. Vis. Appl. 23, 6 (2012), 1103–1113.

[17] S. L. Happy and Aurobinda Routray. 2019. Fuzzy histogram of optical flow orientations for micro-expression recognition. IEEE Trans. Affect. Comput. 10, 3 (2019), 394–406.

[18] Mehrtash Harandi, Richard Hartley, Chunhua Shen, Brian Lovell, and Conrad Sanderson. 2015. Extrinsic methods for coding and dictionary learning on grassmann manifolds. Int. J. Comput. Vision 114, 2–3 (2015), 113–136.

[19] Mehrtash T. Harandi, Richard Hartley, Brian Lovell, and Conrad Sanderson. 2015. Sparse coding on symmetric positive definite manifolds using bregman divergences. IEEE Trans. Neural Netw. Learn. Syst. 27, 6 (2015), 1294–1306.

[20] Xiaopeng Hong, Yingyue Xu, and Guoying Zhao. 2016. LBP-TOP: A tensor unfolding revisit. In Proceedings of the Asian Conference on Computer Vision. Springer, 513–527.

[21] X. Hong, G. Zhao, M. Pietikäinen, and X. Chen. 2014. Combining LBP difference and feature correlation for texture description. IEEE Trans. Image Process. 23, 6 (June 2014), 2557–2568.

[22] Xiaopeng Hong, Guoying Zhao, Stefanos Zafeiriou, Maja Pantic, and Matti Pietikäinen. 2016. Capturing correlations of local features for image representation. Neurocomputing 184 (2016), 99–106.

[23] Xiaohua Huang, Su-Jing Wang, Xin Liu, Guoying Zhao, Xiaoyi Feng, and Matti Pietikäinen. 2019. Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. IEEE Trans. Affect. Comput. 10, 1 (2019), 32–47.

[24] Xiaohua Huang and Guoying Zhao. 2017. Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern. In Proceedings of the International Conference on the Frontiers and Advances in Data Science (FADS’17). IEEE.

[25] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen. 2016. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing 175 (2016), 564–578.

[26] Yan Ke, Rahul Sukthankar, and Martial Hebert. 2005. Efficient visual event detection using volumetric features. In Proceedings of the International Conference on Computer Vision (ICCV’05).

[27] Hui-Qian Khor, John See, Raphael C.-W. Phan, and Weiyouo Lin. 2018. Enriched long-term recurrent convolutional network for facial micro-expression recognition. In Proceedings of the International Conference on Automatic Face and Gesture Recognition.

[28] Alexander Kläser, Marcin Marszalek, and Cordelia Schmid. 2008. A spatio-temporal descriptor based on 3D-gradients. In Proceedings of the British Machine Vision Conference (BMVC’08).

[29] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou. 2019. Learning modality-consistency feature templates: A robust RGB-infrared tracking system. IEEE Trans. Industr. Electron. (2019), 1–1.

[30] X. Lan, S. Zhang, P. C. Yuen, and R. Chellappa. 2018. Learning common and feature-specific patterns: A novel multiple-sparse-representation-based tracker. IEEE Trans. Image Process. 27, 4 (Apr. 2018), 2022–2037.

[31] Ivan Laptev. 2005. On space-time interest points. Int. J. Comput. Vision 64, 2–3 (2005), 107–123.

[32] Ivan Laptev and Patrick Pérez. 2007. Retrieving actions in movies. Proceedings of the International Conference on Computer Vision (ICCV’07).

[33] Christopher J. Leggetter and Philip C. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Lang. 9, 2 (1995), 171–185.

[34] Shan Li and Weihong Deng. 2019. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Trans. Image Process. 28, 1 (2019), 356–370.

[35] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pflister, Guoying Zhao, and Matti Pietikäinen. 2015. Reading hidden emotions: Spontaneous micro-expression spotting and recognition. Arxiv:1511.00423 (2015).
[36] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen. 2018. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Trans. Affect. Comput. 9, 4 (Oct. 2018), 563–577.

[37] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. 2013. A spontaneous micro-expression database: Inducement, collection and baseline. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition.

[38] Yante Li, Xiaohua Huang, and Guoying Zhao. 2018. Can micro-expression be recognized based on single apex frame?. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP’18). IEEE, 3094–3098.

[39] Ming Liang and Xiaolin Hu. 2015. Recurrent convolutional neural network for object recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3367–3375.

[40] Sze-Teng Liong, Raphael C.-W. Phan, John See, Yee-Hui Oh, and KokSheik Wong. 2014. Optical strain-based recognition of subtle emotions. In Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems. IEEE.

[41] Sze-Teng Liong, John See, Raphael C.-W. Phan, and KokSheik Wong. 2014. Subtle expression recognition using optical strain weighted features. In Proceedings of the Asian Conference on Computer Vision. Springer.

[42] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C.-W. Phan. 2018. Less is more: Micro-expression recognition from video using apex frame. Signal Process.: Image Commun. 62 (2018), 82–92.

[43] Ce Liu, Antonio Torralba, William T. Freeman, Frédo Durand, and Edward H. Adelson. 2005. Motion magnification. ACM Trans. Graph. 24, 3 (2005), 519–526.

[44] M. Liu, S. Shan, R. Wang, and X. Chen. 2016. Learning expressionlets via universal manifold model for dynamic facial expression recognition. IEEE Trans. Image Process. 25, 12 (Dec. 2016), 5920–5932.

[45] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu. 2016. A main directional mean optical flow feature for spontaneous micro-exression recognition. IEEE Trans. Affect. Comput. 7, 4 (2016), 299–310.

[46] Y. Lui. 2012. Advances in matrix manifolds for computer vision. Image Vision Comput. 30, 6–7 (2012), 380–388.

[47] K. Mase and A. Pentland. 1989. Automatic Lipreading by Optical-flow Analysis. Wiley.

[48] Iain Matthews, Tim Cootes, J. Andrew Bangham, Stephen Cox, and Richard Harvey. 2002. Extraction of visual features for lipreading. IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002), 198–213.

[49] Veena Mayya, Radhika Pai, and Manohara Pai. 2016. Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics. IEEE.

[50] AraV. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. 2002. Dynamic Bayesian networks for audio-visual speech recognition. EURASIP J. Appl. Signal Process. 2002, 1 (2002), 1274–1288.

[51] Jacob L. Newman and Stephen J. Cox. 2002. Language identification using visual features. IEEE Audio, Speech, Language Process. 20, 7 (2012), 1936–1947.

[52] JuanCarlos Niebles, Hongcheng Wang, and Li Fei-Fei. 2008. Unsupervised learning of human action categories using spatial-temporal words. Int. J. Comput. Vision 79, 3 (Sept. 2008), 299–318.

[53] Eng-Jon Ong and Richard Bowden. 2011. Learning temporal signatures for lip reading. In Proceedings of the International Conference on Computer Vision (ICCV’11). IEEE, 958–965.

[54] Sungsoo Park and Daejin Kim. 2009. Subtle facial expression recognition using motion magnification. Pattern Recogn. Lett. 30, 7 (2009), 708–716.

[55] Sung Yeong Park, Seung Ho Lee, and Yong Man Ro. 2015. Subtle facial expression recognition using adaptive magnification of discriminative facial motion. In Proceedings of the 23rd ACM International Conference on Multimedia. ACM, 911–914.

[56] Devangini Patel, Xiaopeng Hong, and Guoying Zhao. 2016. Selective deep features for micro-expression recognition. In Proceedings of the 23rd International Conference on Pattern Recognition (ICPR’16). IEEE, 2258–2263.

[57] Min Peng, Chongyang Wang, Tong Chen, Guanguan Liu, and Xiaolan Fu. 2017. Dual temporal scale convolutional neural network for micro-expression recognition. Front. Psychol. 8 (2017), 1745.

[58] Wei Peng, Xiaopeng Hong, Yingyue Xu, and Guoying Zhao. 2019. A boost in revealing subtle facial expressions: A consolidated eulerian framework. In Proceedings of the IEEE International Conference on Automatic face and Gesture Recognition. IEEE.

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 15, No. 3s, Article 94. Publication date: December 2019.
[86] Guoying Zhao, Mark Barnard, and Matti Pietikäinen. 2009. Lipreading with local spatiotemporal descriptors. *IEEE Trans. Multimedia* 11, 7 (2009), 1254–1265.

[87] Guoying Zhao and Matti Pietikäinen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 6 (2007), 915–928.

[88] S. Zhao, G. Ding, Y. Gao, and J. Han. 2017. Approximating discrete probability distribution of image emotions by multi-modal features fusion. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

[89] S. Zhao, G. Ding, Y. Gao, X. Zhao, Y. Tang, J. Han, H. Yao, and Q. Huang. 2018. Discrete probability distribution prediction of image emotions with shared sparse learning. *IEEE Trans. Affect. Comput.* Early Access. DOI: 10.1109/TAFFC.2018.2818685

[90] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W Schuller, and Kurt Keutzer. 2018. Affective image content analysis: A comprehensive survey. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI’18)*. 5534–5541.

[91] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, and Guiguang Ding. 2017. Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Trans. Multimedia* 19, 3 (2017), 632–645.

[92] Z. Zhou, X. Hong, G. Zhao, and M. Pietikäinen. 2014. A compact representation of visual speech data using latent variables. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 1 (Jan. 2014), 1–1.

[93] Ziheng Zhou, Guoying Zhao, Xiaopeng Hong, and Matti Pietikäinen. 2014. A review of recent advances in visual speech decoding. *Image Vision Comput.* 32, 9 (2014), 590–605.

[94] Z. Zhou, G. Zhao, and M. Pietikäinen. 2011. Towards a practical lipreading system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’11)*.

[95] Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. 2011. Towards a practical lipreading system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’11)*. IEEE, 137–144.

[96] Yuan Zong, Xiaohua Huang, Wenming Zheng, Zhen Cui, and Guoying Zhao. 2018. Learning from hierarchical spatiotemporal descriptors for micro-expression recognition. *IEEE Trans. Multimedia* 20, 11 (2018), 3160–3172. DOI: 10.1109/TMM.2018.2820321

Received January 2019; revised June 2019; accepted June 2019