Intelligent Recognition of Spontaneous Expression Using Motion Magnification of Spatio-temporal Data

B.M.S. Bahar Talukder[1], Brinta Chowdhury[1], Tamanna Howlader[2],
and S.M. Mahbubur Rahman[1](✉)

[1] Department of Electrical and Electronic Engineering,
Bangladesh University of Engineering and Technology, Dhaka 1205, Bangladesh
sabquatfast@gmail.com, brinta.buet.09.eee@gmail.com,
mahbubur@eee.buet.ac.bd

[2] Institute of Statistical Research and Training,
University of Dhaka, Dhaka 1000, Bangladesh
tamanna@isrt.ac.bd

Abstract. The challenges of recognition of spontaneous expressions from spatio-temporal data include the characterization of subtle changes of facial textures, which in many cases occur for a very brief duration. In this context, the paper presents an intelligent approach for spontaneous expression recognition algorithm, wherein adaptive magnification of motion of spatio-temporal data is applied prior to the extraction of features of expression. The proposed magnification enhances the low-intensity facial activities without introducing notable artifacts for the high-intensity activities. The local binary patterns extracted from three-orthogonal planes of the Eulerian magnified spatio-temporal data are used as features of spontaneous expressions. The extracted features are classified using the well-known support vector machine classifier. Experiments are conducted on commonly-referred spatio-temporal databases such as the SMIC and MMI that have spontaneous expressions representing the micro- and meso-level facial activities, respectively. Experimental results reveal that the proposed approach of motion magnification prior to feature extraction significantly improves the detection and classification accuracy at the expense of acceptable robustness.

Keywords: Eulerian motion magnification · Expression features · Local binary patterns · Spatio-temporal data · Spontaneous expression

1 Introduction

In the recent years, the understanding of emotional state of humans from physiological traits has been gaining increasing research interest, especially in the area of security aware applications. This is mainly due to the fact that estimating the meaningful emotional state can be very useful for wide-deployment of interactions between humans and machines as well as for automatic analysis of

M. Chau et al. (Eds.): PAISI 2016, LNCS 9650, pp. 114–128, 2016.
DOI: 10.1007/978-3-319-31863-9_9

social behavior of people. The emotional state can be continuous in three independent spaces, namely, valance, arousal and dominance [1]. However, Ekman and Friesan [2] have shown that discrete emotional level can be categorized in six basic expressions, viz., Happy, Sad, Anger, Surprise, Fear, and Disgust, which are independent of cultures. The physiological traits that are used for recognition of expressions include the measurements of activities of faces via spatio-temporal imaging modalities [3], activities of tones via voice modalities [4], or even that of bio-signals, e.g., electromyography and electroencephalography [5]. Nevertheless, the spatio-temporal affective analysis has remained in the top-priority due to the fact that such modalities capture a significant amount of information in a non-intrusive manner.

Spatio-temporal expression classification methods can be broadly classified in three categories, namely, the model or geometric-based, holistic or appearance-based, and combination of these or hybrid approach [6]. In the geometric approach, certain key points on the face images are identified and the features of expressions are obtained from these fiducial points. For example, the inter distance among these points, the change of textures of the neighboring region of these points over the frames that is often referred to as the facial action units (FAUs) are considered as measures of facial activities. The main problem of the geometric approach lies in the selection of fiducial points, which often requires manual intervention even in controlled environments. Further, the accuracy of classification is highly sensitive to the localization of the fiducial points. At the same time, the geometric approach often requires computationally expensive algorithms such as the elastic bunch graph matching method to obtain the features of facial activities [7].

In the appearance-based approach, the entire facial region is considered for extracting features of expression instead of certain fiducial points. Conventional holistic approaches that have been used for extracting facial features include the principal component analysis (PCA), independent component analysis, Fisher discriminant analysis (FDA) with Asymmetry-Face, kernel PCA-FDA, non-negative matrix factorization, and mixture covariance analysis applied to the whole face. In the hybrid approach, features of expressions are extracted from the local neighboring regions of facial parts called patches, which are partitioned uniformly or selectively. Manifold learning of patches have been shown to be effective for classification of expressions in the cases of significant distortions of faces such as those due to occlusions [8]. The texture-based features of the uniform or selective patches of facial images that were used for expression classification include the scale invariant feature transform (SIFT), histogram of oriented gradients (HOGs), local binary patterns (LBPs), local directional number pattern, local directional patterns variance, multiscale Gaussian derivatives, Gabor, log-Gabor features, and the geometric orthogonal moments. Densely sampled facial features have also been used to determine the expressions in-the-wild [9]. The expressions in question are determined from the chosen features using well-known classification techniques such as the support vector machine (SVM), Bayesian dynamic network, and neural network. Suitable feature selection strategies including the

AdaBoost and bagging have also been applied to minimize the redundancy and maximize the relevancy to improve the classification performance.

In order to accommodate the depth information of facial parts, the 3D face images have also been used in addition to the 2D intensity image for expression analysis. For example, the SIFT, HOG, and LBP-based features have been fused together to classify expressions from 3D face images [10]. It is to be pointed out that due to the constrained settings of 3D imaging, the practical facial expression analysis still depends very much on the spatio-temporal information available in 2D image sequence. Ji and Idrissi [11] have recommended that LBPs obtained from three orthogonal planes (TOPs) of spatio-temporal data can effectively represent features of facial expressions. Recent surveys on spatio-temporal analysis for facial affective analysis can be found in [12,13].

In practice, there exist two major types of expressions, viz., the posed- and spontaneous-type, based on the exaggeration of facial activities available in spatio-temporal data. The involvement of professional guidance are involved while disposing the posed-type expressions, while touches of real-life such as subtle changes of facial textures exist in the spontaneous-type expressions. Studies reveal that spontaneous-type expressions are significantly different from posed-type expressions, and in general, the slightest change of facial activities in spatio-temporal data can be more important in the former than that in the latter. In our day-to-day life, the subtle facial activities may last for couple of seconds representing a meso-level spontaneous expression. On the other hand, in micro-level spontaneous expressions, when people try to conceal their emotions, the extraction of facial activities is even more challenging. This is mainly due to the fact that in such cases the facial changes occur in fraction of a second [14]. The recognition of micro-expressions serves as important clue for detecting lies that usually occur in high-stake situations when people know about serious consequences of lying or cheating. In the literature, there exists few number of research studies that focus on the automatic recognition of micro-expressions. For example, Shreve et al. [15] used strain patterns as a feature descriptor for spotting posed-type micro-expressions in spatio-temporal data. Polikovsky et al. [16] have used the HOGs as a descriptor for micro-expression recognition in posed scenario. An initial research carried out by Li et al. [14] has shown that the features obtained in terms of the LBP-TOP of spatio-temporal data can perform well for classifying the micro-expression even in the case of spontaneous scenario. In [17], class-specific pass bands of temporal filters have been prescribed for magnification of spatio-temporal data. This method detects the micro-expressions by recognizing that the low, mid-range, and high frequency temporal data correspond to three types of movements, viz., broad head, lip/brow, and eye/pupil. Discriminative learning of the bands of temporal filter is proposed for recognizing subtle facial expressions in [18]. Dynamics of depth information and dense motion field of faces while uttering certain vocabularies are also used to determine micro-expressions [19]. In [20], a set of arbitrarily chosen magnification factors for the region specific motion vectors of geometric face features was used to enhance the recognition performance of subtle expressions. In a recent report, Li et al. [21] use peak contrast of feature difference to spot the micro-expressions in spatio-temporal data. In this algorithm, heuristically chosen ten discrete levels are applied to magnify the Eulerian motion of spatio-temporal data prior to the extraction of LBP or HOG features for recognizing the micro-expressions.

In this paper, we argue that instead of applying fixed-level of magnification of motion, certain adaptive magnification should be employed to enhance the subtle activities of spatio-temporal data for the purpose of recognition of spontaneous expressions. The objectives of adaptation of motion magnification is two-fold: (i) to amplify the low intensity motions in micro-expressions and (ii) to reduce excessive artifacts induced in the meso-level expressions due to magnification. In particular, we promote the use of simple yet effective mean-adaptive Eulerian motion magnification in conjunction with the LBP-TOP features for the representation of micro- or meso-level facial activities in the spontaneous expressions. Experimental results obtained from databases having micro-expressions, namely, spontaneous micro-expression (SMIC) and meso-level expressions, namely, M&M initiative (MMI) reveal that the proposed approach of adaptive motion magnification improves the performance of expression classification significantly.

The paper is organized as follows. Section 2 presents the proposed approach of adaptation for the Eulerian motion magnification. The feature extraction and classification of expressions are detailed in Sects. 3 and 4, respectively. The experimental results showing the significance of proposed approach for improving the classification performance is given in Sect. 5. Finally, Sect. 6 provides the conclusions.

2 Motion Magnification

Let $I(x, y; t)$ denote the spatial intensity at position (x, y) and time t in a spatio-temporal data of size (X, Y, T). Let the initial intensity in a given frame $I(x, y; 0)$ be denoted as $f(x, y)$. If the translational displacement called motion vector in time t is $\delta^{xy}(t) = \delta^{x}(t) + j\delta^{y}(t)$, where j is a complex operator, then $I(x, y; t) = f(x, y; \delta^{xy}(t))$. The Eulerian motion magnification of the data by a fixed-level α refers to synthesizing a signal given by [22]

$$I_m(x, y; t) = f(x, y; (1 + \alpha)\delta^{xy}(t)) \tag{1}$$

According to first-order Taylor series expansion around (x, y), the motion magnified signal can be approximated as

$$\tilde{I}_m(x, y; t) \approx f(x, y) + (1 + \alpha)\sqrt{\left(\delta^{x}(t) f_x\right)^2 + \left(\delta^{y}(t) f_y\right)^2} \tag{2}$$

where $f_x \equiv \partial f(x, y)/\partial x$ and $f_y \equiv \partial f(x, y)/\partial y$. In a general case, the selective-band temporal filter is used so that a good approximation of motion magnified signal can be attained [17]. Let $\delta^{xy}_{\omega}(t) = \delta^{x}_{\omega}(t) + j\delta^{y}_{\omega}(t)$ represent the different spectral components of $\delta^{xy}(t)$ in a continuous variable of temporal frequency ω. Let the frequency dependent motion magnification factor be α_{ω}. In such a case, the resultant motion magnified intensity is given by

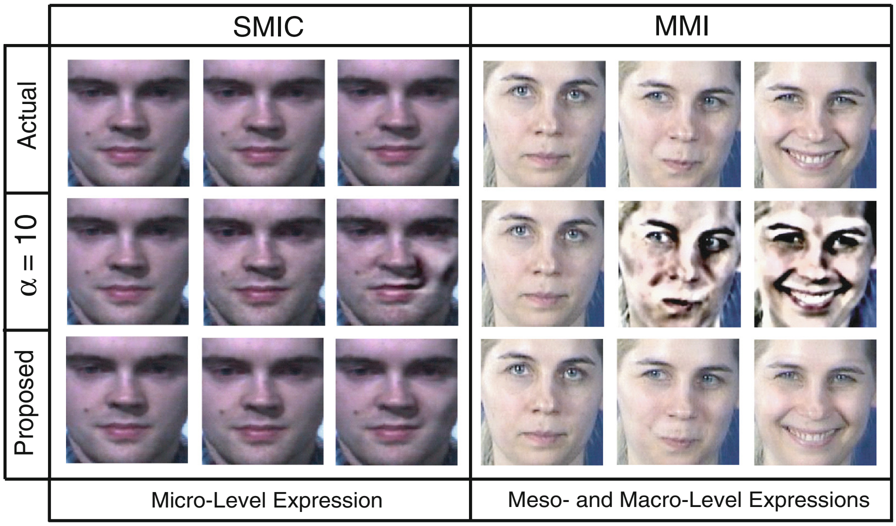


Fig. 1. Comparison of outputs of video frames due to motion magnification in micro-, meso-, and macro-level expressions. The top row of frames show the actual frames. The middle row shows the motion magnified frames when $\alpha = 10$. The bottom row shows the motion magnified frames as per the proposed approach.

$$\hat{I}_m(x, y; t) \approx f(x, y) + \int_{\omega} (1 + \alpha_\omega) \sqrt{\left(\delta_\omega^x(t) f_x\right)^2 + \left(\delta_\omega^y(t) f_y\right)^2} \mathrm{d}\omega \tag{3}$$

The first-order Taylor series expansion may introduce significant artifacts when spatial frequency is considerably high such as for noticeable changes in $f(x, y)$. In order to restrict the over magnification due to high spatial frequencies, the magnification factor α_ω is constrained according to the recommendation given in [22] as

$$(1 + \alpha_\omega)\delta_\omega^x(t) < \frac{\lambda_u}{8} \tag{4}$$

$$(1 + \alpha_\omega)\delta_\omega^y(t) < \frac{\lambda_v}{8} \tag{5}$$

where $\lambda_u = 2\pi/u_x$ and $\lambda_v = 2\pi/v_y$ are the wavelengths of spectral components of $f(x, y)$ that are expressed in terms of the continuous variables of spatial frequencies u_x and v_y, respectively. These restrictions may not work well for the meso-level spontaneous expressions and macro-level posed expressions, when the facial activities are non-trivial. Hence, an adaptive magnification of spatio-temporal data is required, in which the magnification level can be selected according to the overall motions available in the data. In particular, the adaptive magnification level can be inversely proportional to the mean of magnitude of displacement vectors available in the entire data given by

$$\hat{\alpha}_{\omega} = \eta XYT \left[\int_x \int_y \int_t \sqrt{\left(\delta^x_{\omega}(t)\right)^2 + \left(\delta^y_{\omega}(t)\right)^2} \mathrm{d}x \mathrm{d}y \mathrm{d}t \right]^{-1} \tag{6}$$

where η ($\eta \geq 1$) is a proportional constant. A value of η close to unity is preferable to avoid excessive magnification and resulting artifacts in the frames. In practice, any pixel-based motion estimation algorithm can be used to find the adaptive motion magnification factor $\hat{\alpha}_{\omega}$. In the proposed method, the block-matching and optical flow-based subpixel motion estimation technique is recommended due its fast implementation [23]. Figure 1 shows a typical comparison of the motion magnified frames of spatio-temporal data having micro-, meso-, and macro-level expressions that are available in the SMIC and MMI databases. It is seen from this figure that a heuristic choice of magnification factor such as $\alpha = 10$ can produce significant artifacts in spatial textures of the frames, which in fact severely affects the expression recognition performance. As expected, the artifacts caused due to improper magnification is increasingly pronounced in the case of micro-, meso-, and macro-level expressions. However, if the proposed adaptation of scaling of magnification is considered, then the artifacts are reduced significantly. It may be mentioned that the perceptual quality of motion magnification and reduced amount of artifacts provided by the proposed method appears to be even better than that shown in Fig. 1, when the entire frames of the spatio-temporal data are viewed rather than just a few number of frames.

3 Features for Expressions

In the proposed recognition algorithm, the features of expressions are extracted from the magnified spatio-temporal data using the LBP-TOP algorithm [24]. This descriptor is obtained by concatenating LBP on three orthogonal planes: XY, XT, and YT, and considering only the co-occurrence statistics in these three directions. Let us consider that a set of dynamic textures in terms of LBP of size $X_d \times Y_d \times T_d$ ($x_c \in \{1, 2, \cdots, X_d\}, y_c \in \{1, 2, \cdots, Y_d\}, t_c \in \{1, 2, \cdots, T_d\}$) are estimated by considering only the center part of the neighborhood in the magnified data. The histogram of the dynamic feature is estimated as

$$H_{i\ell} = \sum_{x,y,t} I\{\varPhi^c_{\ell}(x, y, t) = i\} \quad i = 1, 2, \cdots, n_{\ell} \quad \ell = 1, 2, 3 \tag{7}$$

where $\varPhi^c_{\ell}(x, y, t)$ expresses the uniform LBP code of central pixel (x_c, y_c, t_c), n_{ℓ} is the number of different labels produced by the LBP operator in the ℓth plane ($\ell = 1 : \text{XY}, 2 : \text{XT}, 3 : \text{YT}$) and

$$I\{A\} = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases} \tag{8}$$

In order to get a coherent description, the histograms obtained are normalized as

$$N_{i\ell} = \frac{H_{i\ell}}{\sum_{k=1}^{n_{\ell}} H_{k\ell}} \quad i = 1, 2, \cdots, n_{\ell} \quad \ell = 1, 2, 3 \tag{9}$$

The labels from the XY-plane contain information about the appearance, whereas that in the XT- and YT-planes represent the co-occurrence statistics of motion in horizontal and vertical directions. The three histograms are concatenated to build a global description F that represents dynamic feature with the spatial and temporal characteristics of the data, which is often referred to as the LBP-TOP.

4 Feature Classification

In the proposed method, the kernel-based SVM is employed to classify the histograms of LBP-TOP by acknowledging that it is a well established statistical learning theory applied successfully in many classification tasks in computer vision. The kernel SVM implicitly maps the LBP-based features into a higher dimensional feature space to find a linear hyperplane, wherein the expressions can be categorized with a maximal margin. Given a training set of $\varGamma$ labeled expressions $\varTheta_{tr} = \{(F_\gamma, d_\gamma) | \gamma = 1, 2, \cdots, \varGamma\}$, where $F_\gamma \in \mathbb{Z}^{n_\ell}$ and $d_\gamma \in \{-1, 1\}$, the test feature F is classified using the decision function

$$\mathcal{D}(F) = \text{sign}\left(\sum_{\gamma=1}^{\varGamma} \beta_\gamma d_\gamma \varPsi\left(F_\gamma, F\right) + b\right) \tag{10}$$

where β_γ are the Lagrange multipliers of a dual optimization problem that describe the separating hyperplane, $\varPsi\left(F_\gamma, F\right)$ is a kernel function, and b is the weight of bias. The training samples F_γ with $\beta_\gamma > 0$ are called the support vectors. The SVM finds the separating hyperplane that maximizes the margin with respect to these support vectors. In order to map the LBP-based histogram into the higher dimensional feature space for classification, the most frequently used kernel functions such as the linear, polynomial, and radial basis function can be used.

Admitting that the SVM provides a binary decision, the multiclass decisions can be obtained by adopting the several two-class or one-against-rest problems. In the proposed method, one-against-rest problems are chosen, and hence ultimate expression class is obtained by $\varGamma$ number of binary learners. With a view to select the parameters of the SVM, a grid-search on the hyper-parameters is used by adopting a cross-validation scheme. The parameter settings that produce the best cross-validation accuracy are used for obtaining the decision on the LBP-TOP feature under test.

5 Experimental Results

The experiments presented in this paper mainly focus on the effect of proposed motion magnification on the recognition of spontaneous expressions. The experiments are conducted on both the micro- and meso-level spontaneous expressions. In order to present representative results, only the findings of the recognition

Table 1. Results of Expression Classification Accuracy of SMIC-HS Dataset

Method	Class	Actual		Magnified	
		Overall (%)	Best (%)	Overall (%)	Best (%)
Detection	Micro	63.65±0.0016	69.05	65.34±0.1111	70.63
Recognition	Positive	37.46±0.0068	52.38	40.63±0.0175	57.14
	Negative	40.63±0.0233	71.43	44.13±0.0199	66.67
	Surprise	52.06±0.0199	76.19	53.97±0.0145	76.19

Table 2. Results of Expression Classification Accuracy of SMIC-VIS Dataset

Method	Class	Actual		Magnified	
		Overall (%)	Best (%)	Overall (%)	Best (%)
Detection	Micro	52.00±0.0039	61.67	54.88±0.0049	65.00
Recognition	Positive	59.33±0.0135	90.00	60.00±0.0243	100.00
	Negative	39.33±0.0107	50.00	49.33±0.0278	80.00
	Surprise	44.00±0.0126	60.00	48.00±0.0246	80.00

performance of micro-expressions of SMIC database and that of the meso-level expressions of the MMI database are presented. In this section, first we provide a brief description of the two datasets, which is followed by the experimental setup and the performance comparisons of expression recognition with and without magnification of motions of data.

5.1 Datasets

The SMIC is a spontaneous micro-expression database having 164 video clips, in which the involuntary emotions were induced by displaying audiovisual films and the required level of inhibition in expressing emotions were strictly maintained by imposing enough pressure to 16 participants [14]. The facial activities were captured using three types of cameras, viz., high speed (HS), normal visual (VIS) and near infra-red (NIR), all having pixel resolution of 640 × 480. The frame rate of HS camera is 100 fps and that of the rest two is 25 fps. The 71 video clips captured from the VIS and NIR cameras yield data similar to standard web cameras, including their limitations such as motion blurs. The micro-expressions in this dataset is classified into three classes (i) Surprise, (ii) Positive representing the emotion Happy and (iii) Negative representing any of the emotions Sad, Fear, or Disgust. The dataset did not elicit any micro-expressions for the emotion Anger. There is an extra class of video clips called non-micro which display no emotion though they have facial movements.

The MMI database has 197 video sequences of faces displaying mostly for the spontaneous-type facial expressions of one of the six basic emotions [25]. The video clips are collected from 75 subjects with a standard camera of frame rate 24

Table 3. Results of Expression Classification Accuracy of SMIC-NIR Dataset

Method	Class	Actual		Magnified	
		Overall (%)	Best (%)	Overall (%)	Best (%)
Detection	Micro	55.44±0.0039	66.67	57.00±0.0044	66.67
Recognition	Positive	58.67±0.0270	80.00	62.00±0.0246	90.00
	Negative	53.33±0.0152	70.00	57.33±0.0135	80.00
	Surprise	62.67±0.0421	90.00	67.33±0.0121	100.00

Table 4. Results of Expression Classification Accuracy of MMI Dataset

Method	Class	Actual		Magnified	
		Overall (%)	Best (%)	Overall (%)	Best (%)
Recognition	Happy	29.33±0.0292	50.00	31.33±0.0327	60.00
	Sad	32.00±0.0160	50.00	39.33±0.0235	70.00
	Surprise	48.67±0.0184	70.00	52.67±0.0121	80.00

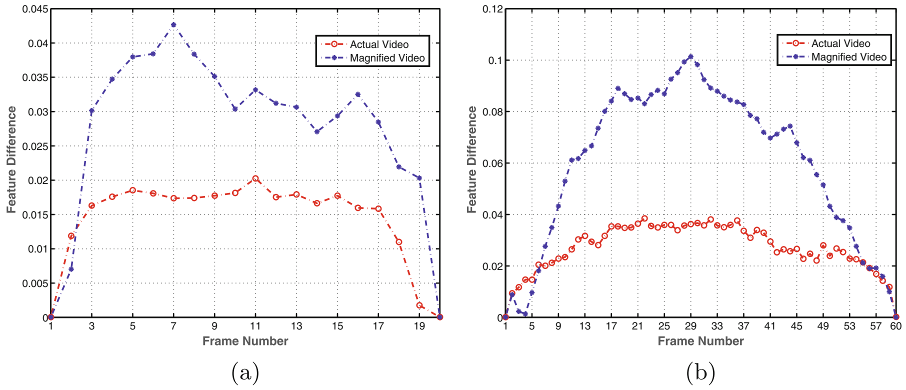
fps and pixel resolution of 720 × 576. The sequences in the data corpus are fully annotated for the presence of single or multiple FAUs in the video. In order to generate a generic dataset of meso-level spontaneous expressions, we choose 25, 21, and 22 number of spatio-temporal clips from this database representing the Happy (i.e., Positive), Sad (i.e., Negative), and Surprise expressions, respectively. The face region is detected first using the Viola-Jones algorithm [26] and by using the coordinates of eye pair. Each spatio-temporal data representing the facial activities is cropped and resized to a spatial resolution of 180 × 240 using the bi-cubic interpolation.

5.2 Setups

To extract the LBP-TOP features from datasets, the frame lengths of clips of all expressions are normalized. In particular, the linear interpolation is used to normalize the frame length 20 for the clips of SMIC and 60 for the MMI datasets. The XY-plane of each of the clips is partitioned to 5×5, and the LBP-TOP features are calculated for all the partitions using the codes available in the website[1] maintained by the developers of LBP. These features are concatenated to obtain the ultimate features of expressions. In order to magnify the motion of a video, the parameter η is chosen to be close to unity. In the experiment, it is found that $\eta = 1.2$ works very well for both the SMIC and MMI datasets. The mean values of motions of spatio-temporal clips of datasets are obtained from the codes available in the website[2] of one of the authors of [23]. The motions are estimated using a block-size of 8×8 and a search limit of 10. The codes available

[1] http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab.

[2] http://scholar.harvard.edu/stanleychan/software/.



(a) (b)

Fig. 2. Amplified feature intensities due to the proposed motion magnification of spatio-temporal clips of the datasets (a) SMIC and (b) MMI.

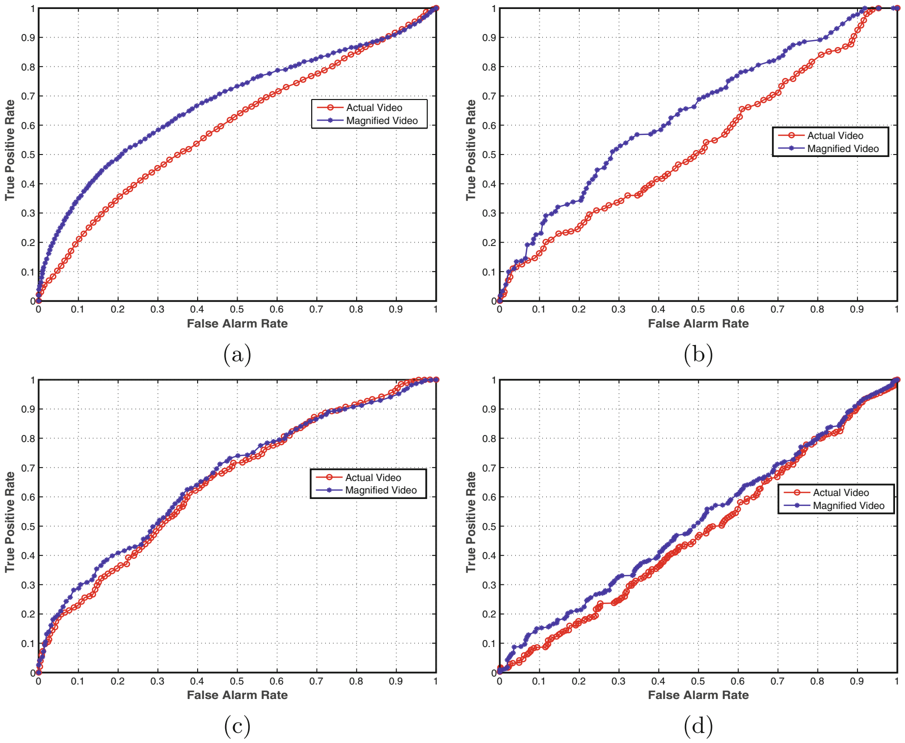
in the website[3] of one of the authors of [22] are employed for the Eulerian video magnification. The Laplacian pyramid with cutoff wavelength 16 is used for spatial filtering during the magnification. The temporal filtering is performed by using two infinite impulse response low-pass filters that have weights 0.4 and 0.05. The chrome attenuation factor is chosen as 0.1. These parameters are set as default in the codes for motion magnification.

The proposed expression recognition algorithm uses supervised learning technique by employing the kernel SVM. In particular, randomly chosen 50 % of the spatio-temporal clips for each of the expressions are treated as the training set, and the rest as the probe set. The expression recognition accuracies are estimated from the correctly classified clips in the probe set. The overall classification performance are reported in terms of the mean and standard deviation of the accuracies obtained from 15 independent randomly chosen training-probe sets. The hyper parameters of the kernel-based SVM classifier are estimated from the cross validation scheme applied only on the training sets. In the experiments, it has been found that the linear kernel function performs the best for the LBP-TOP-based expression classification.

5.3 Results

We first evaluate the changes of magnitude of the LBP-based features due to the proposed motion magnification. In particular, the LBP features are calculated for the XY-plane only, and the frame-by-frame feature differences are obtained from the last frame representing neutral expression of each of the spatio-temporal clips. Figure 2 shows typical frame-by-frame feature differences for the expression Surprise in the SMIC and MMI datasets for both the actual and motion magnified clips according to proposed approach. Since both the datasets have

[3] http://people.csail.mit.edu/mrub/evm/.



(a)

(b)

(c)

(d)

Fig. 3. Results concerning the ROC curves showing improved expression detection due to the proposed motion magnification of spatio-temporal clips. The curves are obtained for detection of (a) micro expressions from SMIC-HS clips, (b) Surprise expression from SMIC-VIS clips, (c) Positive expression from SMIC-NIR clips and (d) Happy expression from MMI clips.

the neutral expression at both ends of the clips, the feature distances slowly increase from the starting frame, reach a peak and then fall to the off-set. What is notable here in this figure is that due to the magnification of motion, the feature differences significantly increase from the actual ones. In other words, the proposed magnification of motion is capable of providing amplification of facial activities, which in turn is expected to provide improved performance for recognizing facial expressions.

Tables 1, 2 and 3 present comparisons of detection accuracy of micro expression and recognition accuracy of the Positive, Negative and Surprise expressions when the clips are magnified according to proposed approach and remain as actual in [14] for the three resolutions of SMIC dataset, viz., HS, VIS and NIR, respectively. It is seen from these tables that the detection accuracy of micro-level expressions is the highest for the HS camera, and it can be increased further by introducing the proposed motion magnification with a very slight compromise in the robustness. This is expected because, the motion magnification can slightly increase the chance of false detection, if the magnification factor is not scaled appropriately. In the case of recognition of types of expressions, the magnification of motion increases the accuracy by more than 3.9 % on average and shows relatively better performance improvement for the VIS and NIR clips. A negligible decrease in the robustness is seen for the HS and VIS clips, but noticeable improvement of robustness is observed for the low-resolution NIR clips especially for recognizing the Surprise expression. In all cases, the best detection or recognition accuracy of magnified video is never less than that without magnification. Table 4 presents the improvement of recognition accuracy for three expressions, namely, Happy, Sad, and Surprise in two scenarios when the spatio-temporal clips of MMI dataset are magnified or remain untouched. As can be seen from this table, even for the meso-level expressions, the improvement of average accuracy is more than 4.25 %, and the highest improvement is seen for the case of Sad expression. The robustness slightly decreases for the Happy and Sad expressions, and increases for the Surprise expression. But the improvements in the mean accuracy or that in the best accuracy due to the introduction of proposed motion magnification of spatio-temporal data actually surpass the slight sacrifice in the robustness.

The effect of magnification of spatio-temporal clips is also evaluated by estimating the receiver operating characteristics (ROC) curves for detection of micro-level expressions as well as for detecting a certain-type of expression from the pool of clips of a dataset. Figure 3 shows comparisons of ROC curves when the clips are magnified or remain as actual for typical scenarios in the SMIC and MMI datsets. In particular, Fig. 3(a) shows that the true positive rate for the detection of micro-expression in the HS clips of SMIC dataset significantly improves especially in the case of low-level false alarm rate, when the clips undergo the proposed motion magnification. Similar improvements are also observed for the recognition of Surprise and Positive expressions in the SMIC dataset as shown in Fig. 3(b) and (c), when the clips are captured by the VIS and NIR cameras, respectively. Due to the poor resolution, the improvement in the NIR clips due to motion magnification is observed to be marginal as compared to that in the HS and VIS clips of the SMIC dataset. Nevertheless, the motion magnification provides significantly higher values of true positive rate for a given false alarm rate for the MMI dataset, which is primarily dominated by the meso-level spontaneous expressions. A typical example of such improvement in detection accuracy of the expression Surprise is shown in Fig. 3(d). Thus, the proposed magnification invariably improves the detection or recognition performance both for the micro- or meso-level spontaneous expressions.

6 Conclusion

The most challenging aspect of detection and recognition of spontaneous expressions is the low-level of the facial activities in the captured spatio-temporal clips. Not only the duration of these facial activities is very brief in period, but also the trivial dynamics of the textures pose notable challenge to extract effective features of expressions. In order to overcome such problems, motion magnification prior to the extraction of features was recommended for detecting and classifying the micro-expressions. However, existing methods use heuristic choice to set the level of motion magnification. In such cases, artifacts are introduced for the facial activities, especially when there remains meso-level facial activities in the spontaneous expressions let alone the macro-level activities in posed expressions. Thus, the proposed paper has introduced a mean-adaptive level of motion magnification, so that small-scale dynamics of faces can be magnified without causing any significant artifacts. The features of the expressions can be extracted from the magnified clips with minimum error and thereby increasing the detection and classification accuracy.

In the proposed method, subpixel-based block-matching algorithm has been used for the motion estimation and the Eulerian technique for motion magnification. The LBP-TOP method has been adopted to extract the features for expression. It has been shown that the proposed magnification of spatio-temporal data enhances the feature intensities as a function of facial activities. The kernel SVM-based classification of features shows that the proposed method can significantly improve the mean accuracy of detection of micro-expression and that of classification of expressions for the SMIC dataset. Such improvements have been observed invariably for the three resolutions, namely, HS, VIS, and NIR of the spatio-temporal clips of SMIC dataset. The increased average classification has also been observed for the MMI dataset, which has meso-level activities representing the spontaneous expressions. The improvements of performance of detection of expressions have also been verified by constructing the ROC curves, which shows that the true positive rate for a given false alarm rate increases on average due to the motion magnification of the clips. We claim that such improvements have resulted due to the proposed adaptation of the level of motion magnification. The only challenge of the proposed magnification is the very tiny scale decay of the robustness of the detection and classification accuracies. But the results of overall classification accuracy, best accuracy, and the true positive rate clearly reveal that the use of proposed adaptation of motion magnification is worthy in detection or classification of spontaneous expressions. We expect that the proposed method can play a significant role for the next generation affective computing in the area of security and surveillance applications.

References

1. Wundt, W.M.: Grundzüge de physiologischen Psychologie. Engelman, Leipzig (1905)
2. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124–129 (1971)
3. Wang, Z., Wang, S., Ji, Q.: Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, pp. 3422–3429 (2013)

4. Bartlett, M.S., Littlewort, G.C., Frank, M.G., Lainscsek, C., Fasel, I.R., Movellan, J.R.: Automatic recognition of facial actions in spontaneous expressions. J. Multimedia **1**(6), 22–35 (2006)
5. Koelstra, S., Patras, I.: Fusion of facial expressions and EEG for implicit affective tagging. Image Vis. Comput. **31**(2), 164–174 (2013)
6. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1424–1445 (2000)
7. Ghimire, D., Lee, J., Li, Z.N., Jeong, S., Park, S.H., Choi, H.S.: Recognition of facial expressions based on tracking and selection of discriminative geometric features. Int. J. Multimedia Ubiquitous Eng. **10**(3), 35–44 (2015)
8. Ptucha, R., Tsagkatakis, G., Savakis, A.: Manifold based sparse representation for robust expression recognition without neutral subtraction. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, Barcelona, Spain, pp. 2136–2143 (2011)
9. Kahou, S.E., Froumenty, P., Pal, C.: Facial expression analysis based on high dimensional binary features. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8926, pp. 135–147. Springer, Heidelberg (2015)
10. Hu, Y., Zeng, Z., Yin, L., Wei, X., Zhou, X., Huang, T.S.: Multi-view facial expression recognition. In: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, Netherlands, pp. 1–6 (2008)
11. Ji, Y., Idrissi, K.: Automatic facial expression recognition based on spatiotemporal descriptors. Pattern Recogn. Lett. **33**(10), 1373–1380 (2012)
12. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1113–1133 (2015)
13. Wang, S., Ji, Q.: Video affective content analysis: a survey of state-of-the-art methods. IEEE Trans. Affect. Comput. **6**(4), 410–430 (2015)
14. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikäinen, M.: A spontaneous micro-expression database: inducement, collection and baseline. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, Shanghai, China, pp. 1–6 (2013)
15. Shreve, M., Godavarthy, S., Goldgof, D., Sarkar, S.: Macro- and micro-expression spotting in long videos using spatio-temporal strain. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, Santa Barbara, pp. 51–56 (2011)
16. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. In: Proceedings of the IET IEEE International Conference on Crime Detection and Prevention, London, UK, pp. 1–6 (2009)
17. Gogia, S., Liu, R.: Motion magnification of facial micro-expressions. Technical report 4, Massachusetts Institute of Technology (2014). http://runpeng.mit.edu/project#research
18. Park, S.Y., Lee, S.H., Ro, Y.M.: Subtle facial expression recognition using adaptive magnification of discriminative facial motion. In: Proceedings of the ACM IEEE International Conference on Multimedia, pp. 911–914 (2015)
19. Akagi, Y., Kawasaki, H.: A method of micro facial expression recognition based on dense facial motion data. In: Proceedings of the IEEE International Conference on Central Europeon Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic, pp. 39–44 (2014)
20. Park, S., Kim, D.: Subtle facial expression recognition using motion magnification. Pattern Recogn. Lett. **30**(7), 708–716 (2009)

21. Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., Pietikäinen, M.: Reading hidden emotions: spontaneous micro-expression spotting and recognition. Technical report 1511.00423v1, Cornell University, arXiv e-prints (2015)
22. Wu, H.Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., Freeman, W.: Eulerian video magnification for revealing subtle changes in the world. ACM Trans. Graph. **31**(4), 1–8 (2012)
23. Chan, S.H., Vo, D.T., Nguyen, T.Q.: Subpixel motion estimation without interpolation. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, Dallas, TX, pp. 722–725 (2010)
24. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
25. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: Proceedings of the IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, pp. 1–5 (2005)
26. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)