applied sciences

MDPI

Article

Hybrid Approach for Facial Expression Recognition Using Convolutional Neural Networks and SVM

Jin-Chul Kim [1], Min-Hyun Kim [1], Han-Enul Suh [1], Muhammad Tahir Naseem [2] and Chan-Su Lee [1,3,*]

[1] The Department of Automotive Lighting Convergence Engineering, Yeungnam University, Gyeongsan 38541, Korea; happymaker@ynu.ac.kr (J.-C.K.); alsgusdla7@naver.com (M.-H.K.); haneol@yu.ac.kr (H.-E.S.)
[2] Research Institute of Human Ecology, Yeungnam University, Gyeongsan 38541, Korea; nmtahir@yu.ac.kr
[3] The Department of Electronic Engineering, Yeungnam University, Gyeongsan 38541, Korea
* Correspondence: chansu@ynu.ac.kr; Tel.: +82-53-810-3527

Abstract: Facial expression recognition is very useful for effective human–computer interaction, robot interfaces, and emotion-aware smart agent systems. This paper presents a new framework for facial expression recognition by using a hybrid model: a combination of convolutional neural networks (CNNs) and a support vector machine (SVM) classifier using dynamic facial expression data. In order to extract facial motion characteristics, dense facial motion flows and geometry landmark flows of facial expression sequences were used as inputs to the CNN and SVM classifier, respectively. CNN architectures for facial expression recognition from dense facial motion flows were proposed. The optimal weighting combination of the hybrid classifiers provides better facial expression recognition results than individual classifiers. The system has successfully classified seven facial expressions signalling anger, contempt, disgust, fear, happiness, sadness and surprise classes for the CK+ database, and facial expressions of anger, disgust, fear, happiness, sadness and surprise for the BU4D database. The recognition performance of the proposed system is 99.69% for the CK+ database and 94.69% for the BU4D database. The proposed method shows state-of-the-art results for the CK+ database and is proven to be effective for the BU4D database when compared with the previous schemes.

Keywords: facial expression recognition; convolutional neural networks; support vector machine; mixture of classifiers; hybrid model

Citation: Kim, J.-C.; Kim, M.-H.; Suh, H.-E.; Naseem, M.T.; Lee, C.-S. Hybrid Approach for Facial Expression Recognition Using Convolutional Neural Networks and SVM. *Appl. Sci.* **2022**, *12*, 5493. https://doi.org/10.3390/app12115493

Academic Editors: Saho AYABE-Kanamura and Monica Perusquia Hernandez

Received: 15 April 2022
Accepted: 22 May 2022
Published: 28 May 2022

1. Introduction

It is important to understand human emotional states in order to interact with another person effectively to support patients' needs in hospitals and to provide a user-friendly service while using service robots, intelligent agents or smart phones [1]. Facial expressions are frequently used to understand human emotional states. The basic facial expressions, including anger, disgust, fear, happiness, sadness and surprise, are observed in most cultures [2,3] and used in many facial expression recognition systems, even though other evidence shows that facial expressions of emotion are culture-specific [4]. Recently, complex facial expressions, which are combinations of basic facial expressions, have also been analyzed and categorized [5]. A facial expression recognition system may also be useful for performance-based human emotional intelligence assessment [6] by measuring the accurate generation of facial expressions for specific emotions, and for the accurate understanding of expressions generated by other persons. Specific configurations of facial muscle movements appear as if they summarily broadcast or display a person's emotions, which is why they are routinely referred to as emotional expressions and facial expressions [7]. For example, Westerners represent each of the six basic emotions with a distinct set of facial movements common to the group, while Easterners do not [4].

Many facial models have been developed based on 2D and 3D landmarks [8], facial appearances [9], geometry [10] and 2D/3D spatio-temporal representations [11]. A com-

prehensive literature review on this subject can be found in review papers [12–14]. These approaches can be categorized as model-based, image-based and deep learning-based. Most of the image-based approaches are based on engineered features such as Histogram of Oriented Gradients (HOG) [15], Local Binary Pattern Histogram (LBPH) [16] and Gabor filters [17].

Facial landmark points are one of the key features of model-based facial expression analysis. Statistical shape models such as active shape models (ASMs) [8] or appearance models such as active appearance models (AAM) [9] are frequently used with classifiers such as a support vector machine (SVM). A facial expression recognition system using multi-class AdaBoost [18] with pairs of landmarks exhibits high performance in facial expression recognition [19]. Automatic landmark point detection and tracking using multiple differential evolution Markov chain particle filters [20] exhibits improved performance over conventional landmark point tracking based on AAM [9]. An SVM classifier on the displacement of facial landmark points yields high classification accuracy [21].

Recently, convolutional neural networks (CNNs) have been used for face recognition [22,23] and facial expression recognition [24–26]. The face alignment and normalization process using 3D face modeling shows an improvement in face recognition performance in CNN-based face recognition [22]. CNN architectures specialized for facial action unit recognition for supporting region-based multi-label learning have also been developed [27]. Deep CNN provides high-level features from the trained deep model. Features extracted from the CNN are very powerful and are used for numerous computer vision applications [28], including facial expression recognition using SVM [24,29,30] and micro-expression recognition tasks with evolutionary search for features [31]. A hybrid model that integrates CNN and SVM by replacing the fully connected classification layer of the CNN with an SVM classifier has also been developed [32].

Efforts have been made to combine multiple approaches for the improved classification of facial expression. Two-channel CNNs with different kernels are combined in a fully connected layer and exhibit better performance than hand-coded features [33]. The ensemble of multiple deep CNNs is used for static facial expression recognition [34]. Deep belief networks (DBN) are used to extract and combine audio and visual features for emotion recognition from audio–video data [35].

Especially for the recognition of facial expressions, combining geometric features and texture features is important for achieving improved performance by compensating for the limitations of one of the features: the geometry provides global facial motions, whereas the texture provides subtle and detailed variations in expression, such as a wink or eyebrow movement. Texture feature extraction based on discriminative non-negative matrix factorization (DNMF) and geometric displacement from grid tracking was used for distance measurement and fused by the SVM classifier [36].

Recently, learning shape, appearance and dynamics with a combination of a CNN and bi-directional long short-term memory neural network (BLSTM) were used for facial action unit recognition [37]. As an alternative to this approach, the joint fine-tuning of deep neural networks from a deep temporal appearance network (DTAN) and deep temporal geometry network (DTGN) is proposed [38]. Two pre-trained networks and joint loss functions are fine-tuned with additional fully connected layers.

A great deal of work is done by using the hybrid approach, a combination of transfer learning and pre-trained deep convolutional networks [39]. The model classifies the facial expressions into seven classes. For the FER-2013 database, the model gives accuracy of 74.39%. Another facial expression recognition work is also presented using graph-based feature extraction and hybrid classification approach (GFE-HCA) [40]. Feature dimensions from facial parts such as the right eye, left eye, nose and mouth are optimized using a weighted visibility graph, which catalyzes the graph-based features from the edge-based invariant features. The combination of a deep convolutional network and modified joint trilateral filter is also used to efficiently recognize facial emotions [41]. However, a system

that combines a conventional feature-based classifier and deep learning approaches does not exist.

The method discussed in [42] addresses a unique artificial intelligence (AI)-based system for speech emotion recognition (SER) that utilizes hierarchical blocks of the convolutional long short-term memory (ConvLSTM) with sequence learning. In order to extract the local emotional features in a hierarchical correlation, four blocks of ConvLSTM are used. In addition, to extract the spatial cues by utilizing the convolution operations, the ConvLSTM layers are adopted for input-to-state and state-to-state transition. Moreover, in order to extract the global information and adaptively adjust the relevant global feature weights according to the correlation of the input features, a novel sequence learning strategy is also utilized.

In this paper, we have presented the hybrid model: a fusion of CNN-based facial expression recognition and conventional high-performance geometry-based facial expression recognition using SVM classifiers. In order to extract facial motion characteristics efficiently, we applied deep learning algorithms to the dense facial motion flow, which was the extracted optical flow from the initial frame (neutral expression) to the given frame of the sequence. In the case of geometric displacement, the displacement direction and strength of the facial landmark points was extracted and used as a feature vector for the SVM classifier of the geometric facial expression data. The weighted summation of the result of the hybrid classifiers provides state-of-the art results for the Cohn–Kanade (CK+) database [43] and is satisfactory for the BU4D database, which is still comparable with the state-of-the-art schemes.

Limitations of Related Work and Our Contributions

Table 1 summarizes the problems in existing approaches. Previous methods have at least one of the following weakness:

- Models discussed in [44–46] were tested on less diverse databases.
- The work in [47–49] is experimented with a small number of classes.
- Dependence on a fixed set of handcrafted features, which requires extensive knowledge about the image characteristics [50]. They rely on texture analysis, where a limited set of local descriptors computed from an image is fed into a classifier such as random forest, etc. Despite a good level of accuracy in some works, these techniques have limitations in generalization and transfer capabilities in inter-database variability.
- Inefficient algorithms, resulting in higher computational costs and time utilization [38].
- The model in [51] uses a semi-automatic process to select features and, by doing so, important features might be missed that might help in better classification.
- Though the model discussed in [52] gives very good accuracy, it misclassifies two classes: anger and sadness.
- The model discussed in [53] generates too many interest points, which might be crucial when handling large databases.

Contrary to the previous works, the proposed approach does not rely on a semi-automatic process for feature selection but computes features automatically. In addition, the proposed approach correctly classifies the two classes, sadness and anger. Further, the use of a diverse and large number of classes and images in the proposed method makes the model more reliable. In this paper, we present the following major contributions:

- We have proposed the hybrid model: a fusion of CNN and SVM for facial expression data.
- To extract facial motion characteristics efficiently, deep learning models are used for the dense facial motion flow, which was the extracted optical flow from the initial frame (neutral expression) to the given frame of sequence.
- We extracted the displacement direction and strength of the facial landmark points for the geometric displacement, which capture the characteristics of the local displacement of facial expressions efficiently for SVM.

- We have used an optimal weighting combination of the hybrid classifiers, which provides better facial expression recognition results than individual classifiers.
- In order to evaluate the model, the CK+ and BU4D databases were used. We secured 99.69% recognition performance for the CK+ database and 94.69% recognition performance for the BU4D database.

Table 1. Comparison and weaknesses of related work.

Publications	Method	Database	Accuracy	Weakness
Liu et al. [47]	Using DNN	CK+, MMI and SFEW	91.74	Small number of classes
Dapogny et al. [50]	Using WLS-RF	CK+, BU4D and SFEW	92.97	Requires handcrafted features
Happy et al. [52]	Using Salient facial patches	CK+ and JAFEE	94.09	Misclassification between anger and sadness
Jung et al. [38]	Using DTAN + DTGN	CK+ and Oulu-CASIA	96.64	Network complexities
Connie et al. [48]	Using CNN + SIFT	FER-2013 and CK+	99.10	Small number of classes
Xu et al. [44]	Using BoMW	BU4D	63.8	Lack of diverse databases
Sun et al. [51]	Using 2D-HMM	Newly created dynamic 3D facial expression database	68.3	Semi-automatic process to select features at initial stage
Hayat et al. [53]	Using LBP-TOP	BU4D	71.6	Produces large number of interest points, which become difficult to handle for large databases
Sandbach et al. [49]	Using FFD	BU4D	73.4	Small number of classes
Li et al. [45]	Using DGIN	BU4D	92.22	Lack of diverse databases
Xue et al. [54]	Using HOG3D (all frames) + HOG3D (onset frames)	BU4D	82.80 + 96.64	Uses two-phase feature selection, which makes the system complex
Zhen et al. [46]	Using DSF (whole video)	BU4D	94.18	Lack of diverse databases

2. Proposed Facial Expression Recognition System Using the Mixture of CNN and SVM Classifier

2.1. System Overview

The proposed system processes video sequences and recognizes facial expressions. The system consists of three modules: data collection of motion flow and facial landmark flow, CNN-based facial motion flow classification and SVM-based facial geometry flow classification, and expression sequence classification by mixing the CNN and SVM-classifier as shown in Figure 1. To process the localized facial motion from a video sequence, global head translation, in-plane head rotation, and scaling effects are eliminated based on face normalization from two eye locations, as explained in Section 2.2.1. The optical flow from the first frame of each sequence is estimated (Section 2.2.2) and a collection of these optical flows is used for CNN-based facial expression classification (Section 2.3). Landmark flow from model-based landmark tracking is used for the SVM-based facial expression classification (Section 2.4). The weighted summation of the probabilistic estimation of the two classifiers is used for the final facial expression classification for each frame of the test sequence (Section 2.5).

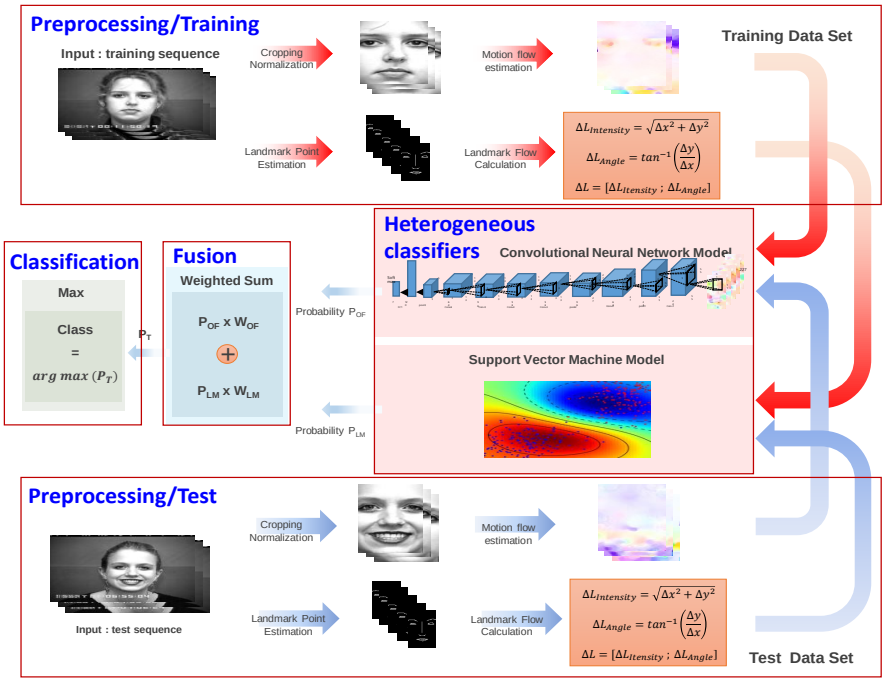


Figure 1. Block diagram of facial expression recognition using mixture of CNN and SVM classifier.

2.2. *Preprocessing*

2.2.1. Face Normalization

To focus on the local facial motion for facial expression recognition, the global head motion is removed by face normalization. When the landmark points are provided in a database such as CK+ [43], we compute the left and right eye centers, and scale and align them such that the eye centers are at a fixed location in the normalized images. In a database with small or no head motion, the eye location is estimated from the first frame and applied to the given sequence with the same global transformation. In a real situation, the use of a face detector at the initial frame can provide the required information for face normalization. Figure 2 shows examples of face normalization from the CK+ database.

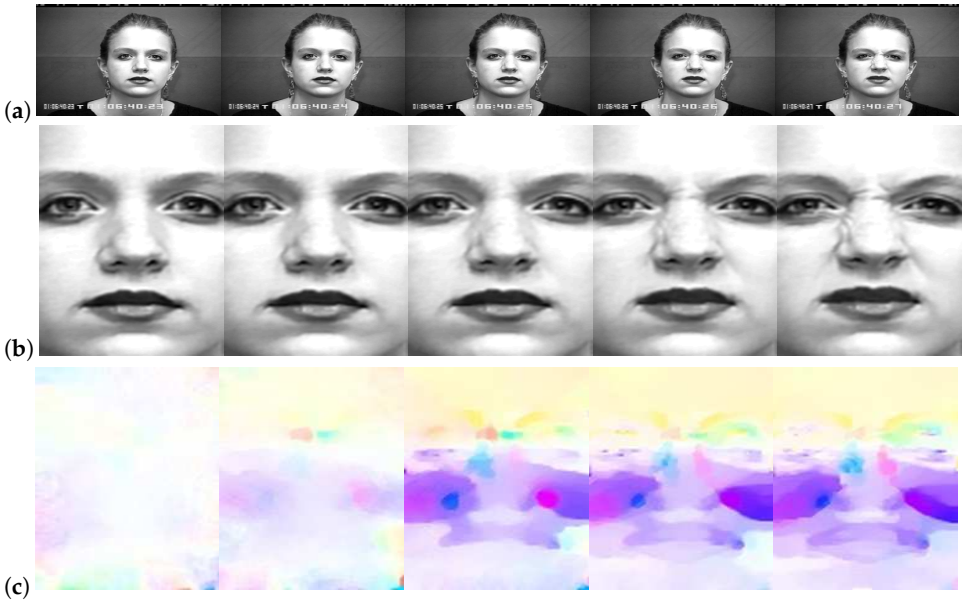


Figure 2. Face normalization and motion flow estimation. (**a**) Original input video sequences. (**b**) Normalized faces. (**c**) Estimated dense optical flow. Optical flow is color-coded by its direction and intensity as in [55].

When a large head motion is encountered, a simple normalization with the initial frame does not eliminate the motion flow due to the global head motion. Recent research has resulted in an advanced landmark detection and localization algorithm [56–58], which can be used to remove large head motion artifacts in real situations.

2.2.2. Motion Flow Estimation

Let $\tilde{I}(x, y, t)$ be an original video sequence, $I(x, y, t)$ be a normalized facial motion video sequence, and $(x(t), y(t))$ be the trajectory of a point in the image plane; then, the brightness constancy assumption states that $I(x(t), y(t), t)$ is constant. Thus,

$$\frac{d}{dt} I(x(t), y(t), t) = 0 \tag{1}$$

when the trajectory of every point is defined by a vector field $\boldsymbol{\mu}(x, y) = (\mu_1(x, y), \mu_2(x, y))$, the vector field $\boldsymbol{u}(x, y)$ satisfies the optical constraints equation [59]:

$$\nabla I \cdot \boldsymbol{u} + \frac{\partial}{\partial t} I = 0 \tag{2}$$

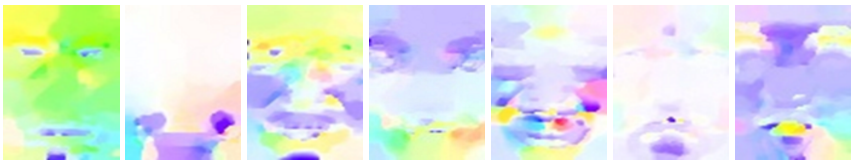
Several solutions have been developed and implemented to solve this under-determined system of equations with additional regularization [55,60–62].

We employed total variation regularization and the robust L_1 norm in the data fidelity term proposed by Zach et al. [63]. To solve the under-determined linear system, the sum of the total variation $\boldsymbol{\mu}$ and L_1 regularization terms are minimized by introducing convex relaxation and applying a duality-based algorithm [63] and point-wise threshold [62].

The collection of $\boldsymbol{\mu}(x, y, t) = (\mu_1(x, y), \mu_2(x, y), t)$ represents the motion flow of the facial expressions from the video sequences. The estimated dense motion flow can be represented by the motion flow direction and its intensity, as shown in Equation (3). This motion flow sequence is robust to variations in the appearance and skin color of the subject. In Figure 2c, the consistent motion flow estimation is shown.

$$\begin{aligned} Intensity(I) &= \sqrt{(\delta x)^2 + (\delta y)^2}, \\ Angle(\theta) &= tan^{-1}\left(\frac{\delta y}{\delta x}\right) \end{aligned} \tag{3}$$

The dense motion flow clearly characterizes the facial expression characteristics of individual expressions. Figure 3 shows the dense motion flow of seven different expressions.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**)

Figure 3. Examples of dense motion flow of different facial expressions. Optical flow is color-coded by its direction and intensity. (**a**) Anger, (**b**) Contempt, (**c**) Disgust, (**d**) Fear, (**e**) Happiness, (**f**) Sadness, (**g**) Surprise.

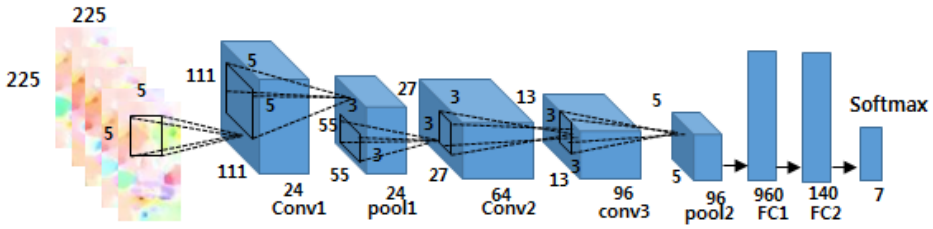
2.3. CNN-Based Facial Motion Flow Classification

The dataset is divided into two parts, one for training and the other for testing, which are collected in advance, and the motion flows and geometric displacement are estimated from the dense optical flow and landmark tracking. In the training dataset, validation data are further separated and are not used in the classifier training.

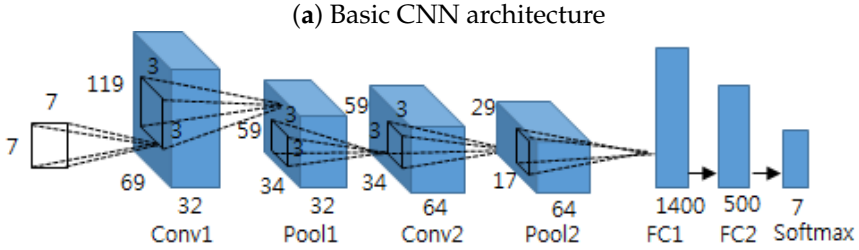
The dense facial motion flow was color-coded, and the color-coded dense motion flows were used as the input to the deep CNN. The flow intensity and its direction were calculated at each pixel for a given optical flow frame. The flow intensity was normalized

by dividing each pixel flow intensity using maximum intensity. After the normalization, the direction of each flow was color-coded by the chronicity values of the pre-defined color wheel, and the normalized intensity of the flow was coded by the saturation value [60]. Thus, the red, blue, and green (RGB) color-coded values of the motion flow provided a continuous representation of the motion flow direction and its intensity at every pixel in each frame.

We have used the basic CNN architecture proposed by Krizhevsky et al. [64] for ImageNet classification. The dense optical flow file was rearranged into $256 \times 256 \times 3$ RGB color as an input to the network. Several different architectures were applied to evaluate the facial expression performance. The selected basic architecture is shown in Figure 4a. Three convolution layers, two pooling layers, and two fully connected layers are used for the facial expression recognition. There is no pooling layer between the second and the third convolution layers. For the activation function, rectified linear units (ReLU) are used as in [64]. The first convolution layer filters the $225 \times 225 \times 3$ input images with a kernel of size $5 \times 5 \times 3$ with a stride of 2 pixels. Overlapping max pooling with windows of size 3×3 is used in each pooling layer. The output layer takes 140 neurons from the last fully connected layer into seven facial expression class labels using the softmax classification. The architecture was further modified to improve the performance of facial expression recognition. The number of convolution layers, the number of pooling layers and their location, and the number of fully connected layers were modified. Based on the validation dataset, we selected another architecture for the facial expression recognition experiment, as shown in Figure 4b.



(**a**) Basic CNN architecture



(**b**) Modified CNN architecture

Figure 4. Basic and modified CNN architectures for facial expression recognition from facial motion flow.

2.4. SVM-Based Facial Geometry Displacement Classification

Facial landmark points are used as input data for facial geometric motion classification. In the case of the CK+ database, the landmark points extracted from the AAMs [9] were provided, and we use the shape data in Equation (4) as the input data for the geometric motion analysis.

$$s(t) = [x_1(t), y_1(t), x_2(t), y_2(t), \cdots, x_i(t), y_i(t), \cdots, x_n(t), y_n(t)], \quad (4)$$

where n is the number of vertices, and t is the sampling time. When the landmark points are not provided, AAM [9] or landmark point detection and tracking algorithms [56–58] can be used to extract the landmark point information from the facial expression sequence.

For facial geometry displacement extraction, we extracted the displacement from neutral facial geometry. Assuming that the facial expression sequence starts from a neutral expression, we extracted the landmark point displacement by subtracting each landmark point sequence by the initial landmark points from discrete sampling frames.

$$\begin{aligned}\Delta_g(t) &= [x_1(t) - x_1(0), y_1(t) - y_1(0), x_2(t) - x_2(0),\\ &\quad y_2(t) - y_2(0), \cdots, x_n(t) - x_n(0), y_n(t) - y_n(0)],\\ &= [\Delta x_1(t), \Delta y_1(t), \Delta x_2(t), \Delta y_2(t), \cdots, \Delta x_n(t), \Delta y_n(t)],\end{aligned} \tag{5}$$

For each landmark point, we further compute the displacement intensity and angle. For example, for the ith landmark point among $[1, \cdots, n]$ at time t, the displacement can be estimated as follows:

$$\Delta L_{Intensity_i}(t) = \sqrt{\Delta x_i(t)^2 + \Delta y_i(t)^2}, \tag{6}$$

$$\Delta L_{Angle_i}(t) = tan^{-1}\left(\frac{\Delta y_i(t)}{\Delta x_i(t)}\right) \tag{7}$$

Therefore, the feature vector for the SVM at time t can be represented by

$$\begin{aligned}\boldsymbol{x}(t) = [&\Delta L_{Intensity_1}(t), \Delta L_{Intensity_2}(t), \cdots, \Delta L_{Intensity_n}(t)\\ &\Delta L_{Angle_1}(t), \Delta L_{Angle_2}(t), \cdots, \Delta L_{Angle_n}(t)],\end{aligned} \tag{8}$$

Figure 5 shows the collection of average landmark flow in different expressions with visualization of facial image patches corresponding to large flow in each expression. The horizontal axis of the graph represents the index of the landmark. The index from 0 to 67 represents δx_i, and the index from 68 to 135 represents δy_i of the landmark points. The average value of each facial expression is plotted in the graph. Relevant facial areas corresponding to their landmark point variations are provided in the graph to show what the average landmark variations represent. The plotted landmark flow shows that landmark flow can be used to characterize local facial action areas such as the mouth, eyes and eye brows in each expression.

For the given feature vectors and class labels, SVMs attempt to find the hyperplane that maximizes the positive and negative observation for a specified class [65]. The decision function can be specified by

$$D(\boldsymbol{x}) = \sum_k y_k \alpha_k^* K(\boldsymbol{x}(k), \boldsymbol{x}) + b, \alpha_k^* \geq 0, \tag{9}$$

$$y_k = \begin{cases} 1 & \text{if } \boldsymbol{x}(k) \in \text{class A} \\ -1 & \text{otherwise} \end{cases} \tag{10}$$

where only support vectors, which are the closest data point to the margin of the classifier, appear in the sum with a nonzero weight, and $y_k = 1$ if they are to be classified as positive samples; otherwise, $y_k = -1$, and the symmetric kernels contain finite or infinite series expansions of the form

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_i \varphi_i(\boldsymbol{x}) \varphi_i(\boldsymbol{x}') \tag{11}$$

The solution for Equation (9) with kernel expansion in Equation (11) can be found by quadratic optimization, and we used the implementation of a library for support vector machines (LIBSVM) [66]. For the kernel map, the radial basis function (RBF) is used in our experiment. As a final result of the SVMs for multiple classes, we can obtain the probabilistic likelihood of each class.

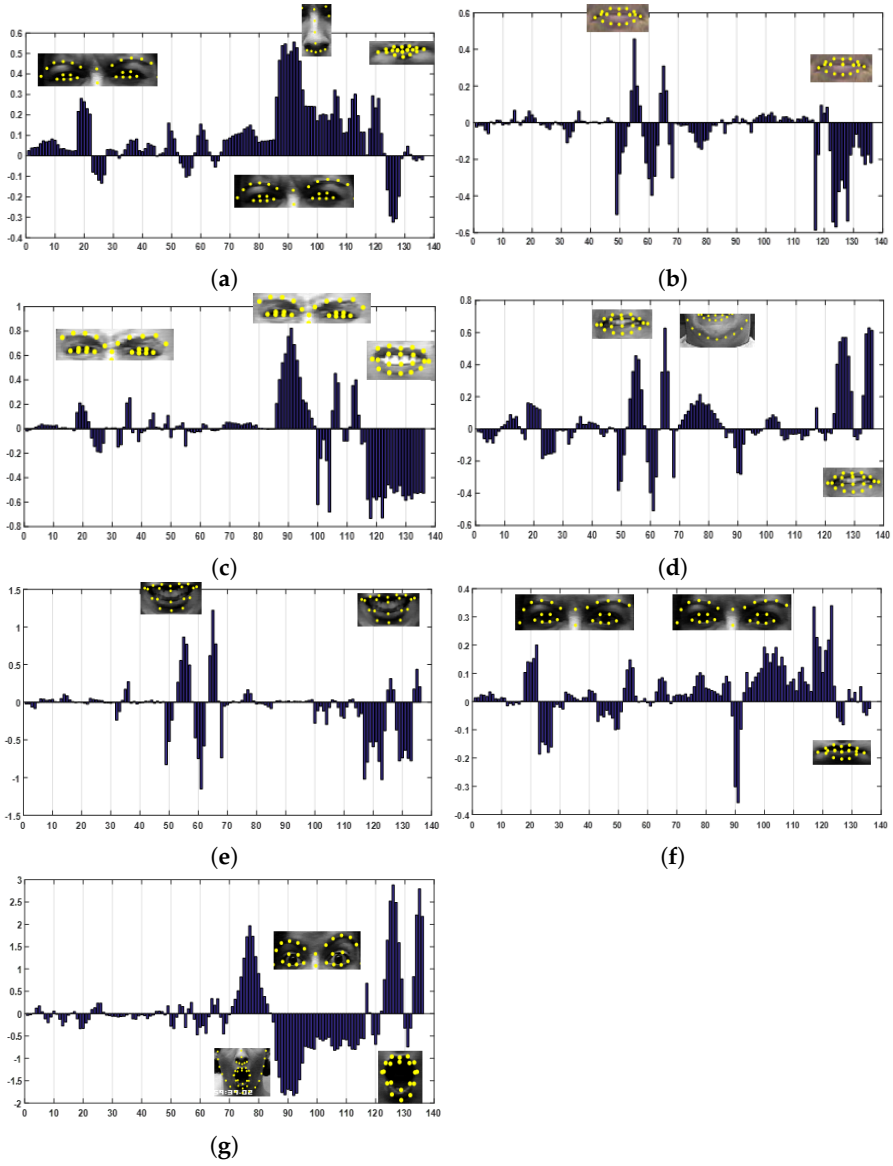


Figure 5. Average landmark flow in different expressions: (**a**) Anger, (**b**) Contempt, (**c**) Disgust, (**d**) Fear, (**e**) Happiness, (**f**) Sadness, (**g**) Surprise. Blue line shows average intensity in each landmark point. Yellow dot shows the landmark flow in the image space.

2.5. Fusion of CNN-Based Facial Motion Classifier and SVM-Based Geometry Displacement Classifier

Both CNN-based classification of facial expressions using facial motion flow and SVM-based facial geometry displacement classification provide the probabilistic likelihood of seven facial expression classes. For the mixture of the CNN and SVM classifier, we applied a weighted summation of the two classifiers, assuming that both of the classifiers provide probabilistic outputs of each class. When P_{cnn} represents the CNN-based facial motion flow classification output, and P_{svm} represents the SVM-based facial geometry displacement classification result, the final output of the mixture classifier can be represented by

$$P_{i_{ALL}}(t) = P_{i_{cnn}}(t) \times w_{cnn} + P_{i_{svm}}(t) \times w_{svm}, \quad (12)$$

where w_{cnn} is the weight for the CNN-based classification, w_{svm} is for the SVM-based classification, and $w_{cnn} + w_{svm} = 1$. For each frame of the facial expression sequence, the

final classification label is selected by determining the maximum probabilistic value after the heterogeneous fusion of the CNN and SVM classifier by weighted summation.

$$\text{Class Label} = argmax_i(P_{i_{ALL}}) \tag{13}$$

For the final classification of the class label for a given facial expression sequence, the majority voting scheme is used. The class with the maximum number of class labels for the given sequence is chosen as the final class label for the given sequence.

The prediction complexity of the RBF kernel SVM is $\mathcal{O}(kd)$, where k is the number of support vectors, with d being the number of input dimensions. If we consider that the matrix computational complexity is $\mathcal{O}(d^3)$ and vector multiplication is $\mathcal{O}(d^2)$, in the CNN model, forward convolution layers are $\mathcal{O}(d^4)$, and fully connected layers are $\mathcal{O}(d^2)$. Therefore, the computational complexity of the CNN model is much higher than the SVM model. The computational complexity of the proposed model is dominated by the CNN model.

3. Experiments

In this section, we evaluate and compare our approach with other state-of-the-art algorithms for facial expression recognition using the CK+ and BU4D databases. We have used 80% data for training and 20% for validation. First, we evaluate the proposed method using CNN-based motion flow. Secondly, the method is tested using the facial geometry displacement-based SVM using the CK+ database. In the last, we evaluate the hybrid model—the mixture of CNN and SVM. Moreover, the proposed method is also evaluated and compared with the state-of-the-art methods using the BU4D database.

3.1. Facial Expression Recognition Using CNN-Based Facial Motion Flow for CK+ Database

The CK+ database was released for the purpose of promoting research into automatically detecting individual facial expressions. Since then, the CK+ database has become one of the most widely used datasets for algorithm development and evaluation as it provides AU coding. Based on the facial action unit criteria, it provides seven facial expression categories: Anger (An), Contempt (Co), Disgust (Di), Fear (Fe), Happiness (He), Sadness (Sa), and Surprise (Su). We used these seven facial expression categories as the class labels of the facial expression recognition system. In all of the following experiments, the CK+ database with seven class labels is used. The number of class samples in the database is different in different expressions.

We evaluated the facial expression recognition performance using four-fold cross-validation for the CK+ database. From the training database, 20% of the data are used for the validation test for the learning of the CNN-based facial expression recognition. For the training of the proposed CNN-based facial expression recognition from facial motion flow, we trained the proposed model with batch size 256 and learning rate starting from 0.0003 and reducing gradually. Figure 6 shows the feature map visualization of the second convolution layer of the model, when we applied training data in different expressions. We selected representative feature maps of each facial expression, which shows feature maps that capture distinguishing characteristics of each facial expression. The visualized feature map of each expression shows different activation areas of each expression, which may be corresponding to the action unit of each expression type. The result implies that the network properly extracts the features of different expression types. Table 2 shows the CK+ facial expression recognition results, with an average recognition rate of 83.3% using CNN-based classification with the network architecture presented in Figure 4a.

Table 2. Confusion matrix: Basic CNN-based facial expression recognition from facial motion flow for CK+.

	An	**Co**	**Di**	**Fe**	**Ha**	**Sa**	**Su**
An	77.8	0.0	6.7	11.1	4.4	0.0	0.0
Co	11.1	22.2	0.0	0.0	66.7	0.0	0.0
Di	1.7	0.0	98.3	0.0	0.0	0.0	0.0
Fe	12.0	0.0	4.0	36.0	16.0	4.0	28.0
Ha	1.4	0.0	0.0	0.0	98.6	0.0	0.0
Sa	21.4	0.0	3.6	10.7	0.0	53.6	10.7
Su	0.0	1.2	2.4	1.2	1.2	1.2	91.6

By the modified network architecture presented in Figure 4b, we achieve higher facial expression recognition performance than the basic network architecture with an average 90.5% recognition rate, as shown in Table 3. We evaluated the facial expression recognition performance in different input data types. Table 4 shows the performance of facial expression recognition for the CK+ database. Three channels are used for the input in the case of Δx, Δy, angle(θ), and Δx, Δy, Intensity(I). The table shows that the proposed two-channel motion flow (Δx, Δy) outperforms other input data types. The performance of the modified CNN-based facial expression recognition using two-channel motion flow (Δx, Δy) is 7.2% better than the basic CNN-based model's performance.

Table 3. Confusion matrix: Modified CNN-based facial motion flow (CK+).

	An	**Co**	**Di**	**Fe**	**Ha**	**Sa**	**Su**
An	89.9	0.4	2.3	1.0	3.9	1.0	1.5
Co	5.6	72.6	0.0	2.2	12.9	0.6	6.1
Di	4.5	0.5	87.7	1.0	4.5	1.5	1.8
Fe	6.5	1.3	1.3	77.3	7.8	0.4	5.4
Ha	7.0	0.4	0.6	1.4	86.7	0.8	3.1
Sa	6.3	1.6	0.9	0.7	4.7	83.6	2.2
Su	2.0	0.5	0.5	1.4	3.5	0.6	91.5

Table 4. Modified CNN-based facial expression recognition performance according to different input data types.

Input Data Type	**Recognition Performance**
Texture image	64.8
Δx	79.51
Δy	83.79
abs($Deltax$, Δy)	77.98
Intensity of motion flow(I)	75.54
Angle of motion flow(θ)	78.29
Intensity(I) and angle(θ) of motion flow	78.29
Δx, Δy, θ	76.76
Δx, Δy, I	88.69
Motion flow(Δx,Δy)	90.52

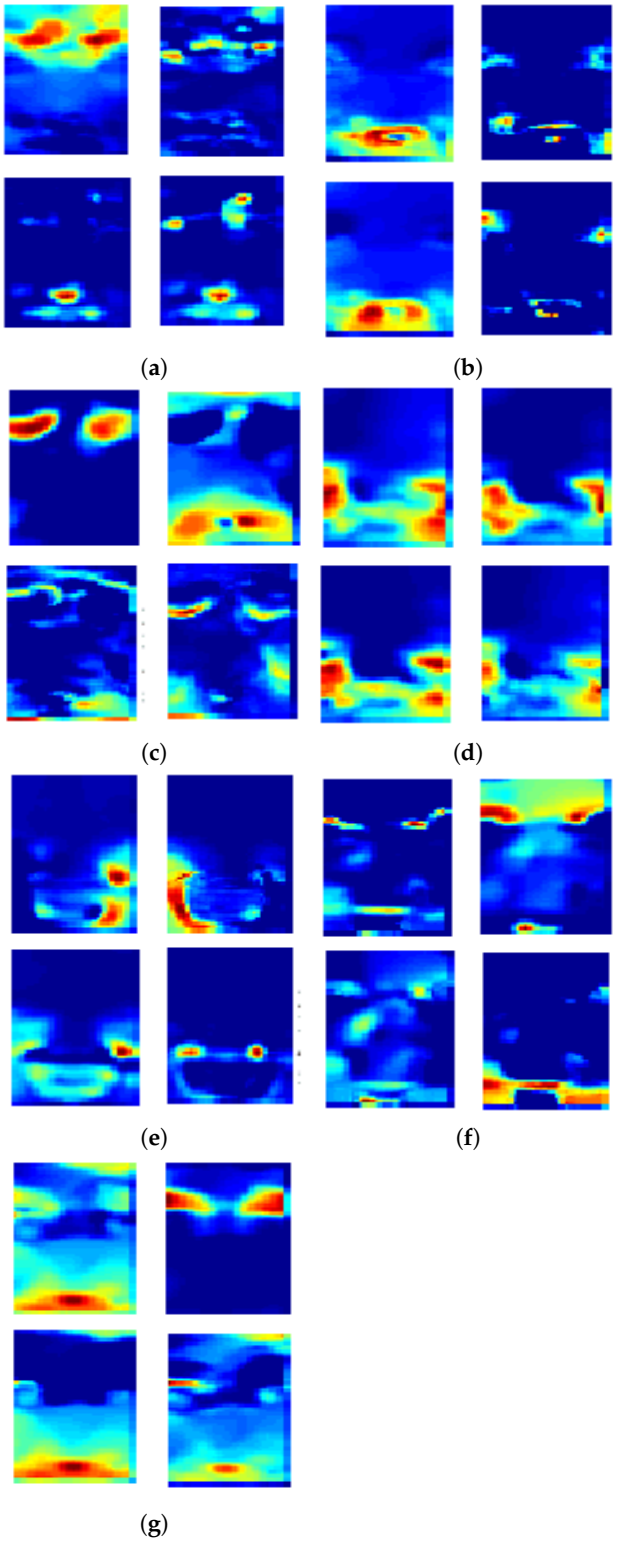


Figure 6. Examples of dense motion flow of different facial expressions. Optical flow is color-coded by its direction and intensity. (**a**) Anger, (**b**) Contempt, (**c**) Disgust, (**d**) Fear, (**e**) Happiness, (**f**) Sadness, (**g**) Surprise.

3.2. Facial Expression Recognition Using Facial Geometry Displacement-Based SVM for CK+ Database

Using the SVM classifier based on facial landmark displacement, we achieved 94.80% facial expression recognition performance for seven facial expression categories. Table 5

shows the confusion matrix of the facial expression recognition performance from the landmark displacement in the sequence. The facial-geometry-based SVM classifier shows superior facial expression recognition performance to the CK+ baseline facial expression recognition by the multi-class SVM using the combination of similarity-normalized shape and canonical appearance features in [43] or the SVM classifier using facial feature point displacement [21].

Table 5. Confusion matrix: SVM-based facial displacement (CK+) from the landmark displacement.

	An	Co	Di	Fe	Ha	Sa	Su
An	95.56	0.0	2.2	0.0	2.2	0.0	0.0
Co	5.55	88.9	0.0	0.0	5.55	0.0	0.0
Di	5.1	0.0	93.2	0.0	1.7	0.0	0.0
Fe	4.0	0.0	0.0	92.0	4.0	0.0	0.0
Ha	2.9	0.0	0.0	0.0	97.1	0.0	0.0
Sa	7.1	0.0	0.0	0.0	3.6	89.3	0.0
Su	1.2	0.0	0.0	0.0	1.2	0.0	97.6

We further investigate the facial expression recognition performance of the SVM with different representations of the facial geometry and its displacement. The landmark point itself, point with normalization by centers of eye location, intensity and angle of the landmark flow, and landmark flow in coordinate (δx, δy) are tested. Table 6 shows the performance of facial expression recognition for the CK+ database according to different input data types of the SVM. In this additional experiment, the SVM classifier using the direct landmark displacement in Equation (5) can achieve 96.64% recognition performance, as shown in Table 7, which is higher than that when using the displacement intensity and angle by 1.83%.

Table 6. SVM-based facial expression recognition performance in different feature data types of facial geometry.

Input Data Type of SVM	Recognition Performance
Landmark point	91.74
Landmark point (normalized)	92.97
Intensity, angle of landmark flow	94.80
Landmark displacement of landmark flow	96.64

Table 7. Confusion matrix: SVM-based facial displacement (CK+) from displacement, intensity and angle.

	An	Co	Di	Fe	Ha	Sa	Su
An	100.0	0.0	0.0	0.0	0.0	0.0	0.0
Co	0.0	100.0	0.0	0.0	0.0	0.0	0.0
Di	1.7	0.0	98.3	0.0	0.0	0.0	0.0
Fe	4.0	0.0	0.0	92.0	4.0	0.0	0.0
Ha	4.3	0.0	0.0	1.4	92.6	0.0	1.5
Sa	3.6	0.0	0.0	0.0	3.6	92.8	0.0
Su	0.0	0.0	0.0	0.0	1.2	0.0	98.8

3.3. Mixture of CNN and SVM Classifier for CK+ Database

For the mixture of the facial expression classifiers using a weighted summation of the probabilistic estimation of the facial expression categories, it is necessary to determine

the weighting factors of the heterogeneous classifiers. The weighting factors can be found based on the validation set that is not used for the training classifiers. Figure 7 shows the variations in the facial expression classification performance according to the weight values when we combine the modified CNN-based facial expression classification and SVM-based classification using the displacement intensity and angle of facial landmarks.

The combined facial expression recognition performance is between 99.69% and 80.7%, which can be higher than the individual classification performance of the two individual classifiers. The performance reaches a peak at a 25% weighting for the CNN-based classifiers and 75% weighting for the SVM-based classifiers. The weighting factors are computed from the validation set in the experiment.

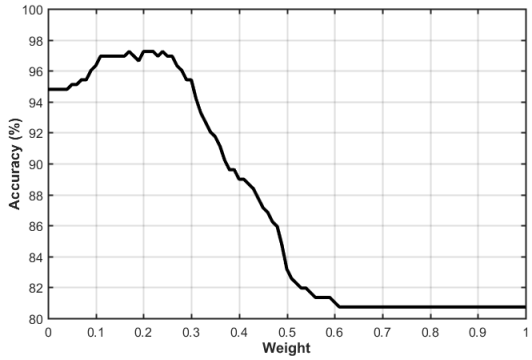


Figure 7. Change in performance according to weight values.

Using the mixture of the basic CNN-based classifier and SVM-based classifier, we achieved 97.25% facial expression recognition performance for the CK+ database. Table 8 presents the confusion matrix of the facial expression recognition performance of the heterogeneous classifier. Figure 8 shows the comparison of the CNN-based classifier, SVM-based classifier, and hybrid CNN-based and SVM-based classifiers. For most of the facial expressions, the fused classifier surpasses or at least matches the peak classification performance of the two heterogeneous classifiers in each facial expression category.

Table 8. Confusion matrix: Weighted summation of basic CNN-based and SVM-based classifiers.

	An	Co	Di	Fe	Ha	Sa	Su
An	97.8	0.0	0.0	0.0	2.2	0.0	0.0
Co	0.0	88.9	0.0	0.0	11.1	0.0	0.0
Di	1.7	0.0	98.3	0.0	0.0	0.0	0.0
Fe	4.0	0.0	0.0	84.0	4.0	0.0	8.0
Ha	0.0	0.0	0.0	0.0	100.0	0.0	0.0
Sa	3.6	0.0	0.0	0.0	0.0	96.4	0.0
Su	0.0	0.0	0.0	0.0	0.0	0.0	100.0

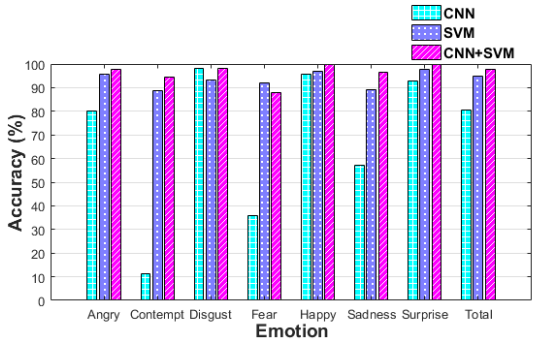


Figure 8. Comparison of facial expression classification performance.

Using the mixture of the modified CNN-based classifier and SVM-based classifier, we achieved 99.69% facial expression recognition performance, which is higher than that of the mixture of the basic CNN-based classifier and SVM-based classifier. The confusion matrix in the case of the fusion of the modified CNN-based classifier and SVM classifier is shown in Table 9.

Table 9. Confusion matrix for CK+ database: Weighted summation of modified CNN-based and SVM-based classifiers.

	An	Co	Di	Fe	Ha	Sa	Su
An	100.0	0.0	0.0	0.0	0.0	0.0	0.0
Co	0.0	94.4	0.0	0.0	5.6	0.0	0.0
Di	0.0	0.0	100.0	0.0	0.0	0.0	0.0
Fe	0.0	0.0	0.0	100.0	0.0	0.0	0.0
Ha	0.0	0.0	0.0	0.0	100.0	0.0	0.0
Sa	0.0	0.0	0.0	0.0	0.0	100.0	0.0
Su	0.0	0.0	0.0	0.0	0.0	0.0	100.0

We compared the performance of the CK+ database with other approaches in the literature. First, we compared the proposed method with AU-Aware Deep Networks (AUDN), in which the researchers proposed a deep learning method for facial expression recognition by incorporating the existing knowledge that the appearance variations caused by expressions can be decomposed into a batch of local facial action units (AUs) [47]. Second, we compared the proposed method with occlusion-robust expression prediction using local subspace random forest (WLS-RF) [50], in which the researchers train random forests upon spatially defined local subspaces of the face. Third, we compared the proposed scheme with the appearance features using salient facial patches [52], in which the researchers suggested a novel foundation for expression identification by extracting a few facial patches, depending upon the location of facial markers. Fourth, we compared the proposed scheme with the joint model of the deep temporal appearance model and deep temporal geometry model (DTAGN) [38], which is stationed on two diverse frameworks. The framework evokes temporal appearance features from the image pattern and temporal geometry features from temporal facial marker spots. Finally, we compared the proposed method with CNN with SIFT (Scale Invariant Feature Transform) [48], in which the researchers suggested CNN for face recognition and attained very good accuracy by using a small portion of training data. Table 10 shows the facial expression recognition performance in different approaches for the CK+ database. The proposed model shows state-of-the-art performance for facial expression recognition on the CK+ database among recent approaches.

Table 10. Facial expression recognition performance in different approaches for CK+ database.

Approach	Details	Recognition Performance
Liu et al. [47]	Using DNN	91.74
Dapogny et al. [50]	Using WLS-RF	92.97
Happy et al. [52]	Using salient facial patches	94.09
Jung et al. [38]	Using DTAN + DTGN	96.64
Connie et al. [48]	Using CNN + SIFT	99.10
Proposed method	Using modified CNN + SVM	99.69

3.4. Facial Expression Recognition for BU4D Database

We have also evaluated the facial expression recognition performance using the BU4D database [67]. BU4D is a high-resolution 3D dynamic facial expression database, which contains facial expression sequences captured from 101 subjects. For each subject, there are

six model sequences showing six prototypic facial expressions (anger, disgust, happiness, fear, sadness and surprise), respectively. Each expression sequence contains approximately 100 frames. Figure 9 shows a sample sequence of a surprise expressions and its landmark points in the BU4D database. Depth data and texture data are provided. We used only texture data for the analysis of facial expressions. Using the mixture of the modified CNN-based classifier and SVM-based classifier, we achieved 94.69% facial expression recognition performance using four-fold cross-validation. Table 11 shows the confusion matrix of facial expression recognition for the BU4D database using the proposed method.

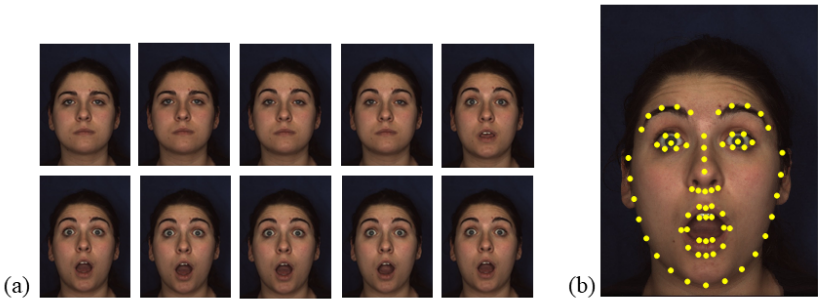


Figure 9. An example sequence of BU4D facial expression database: (**a**) texture images of surprise expression, (**b**) landmark points.

Table 11. Confusion matrix for BU4D database: Weighted summation of modified CNN-based and SVM-based classifiers.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	96.0	1.0	1.0	1.0	0.0	1.0
Disgust	2.0	95.0	0.0	1.0	0.0	2.0
Fear	1.0	0.0	86.0	4.0	1.0	8.0
Happiness	0.0	0.0	0.0	98.0	1.0	1.0
Sadness	3.0	0.0	1.0	1.0	95.0	1.0
Surprise	1.0	0.0	0.0	1.0	0.0	98.0

We also compared the facial expression recognition performance with other facial expression recognition systems using the BU4D database. First, we compared the proposed method with the bags of motion word (BoMW) model [44], which finds facial expressions by using local facial motion descriptors and motion words from clustering its descriptors. Second, we compared the proposed method with the 3D facial surface descriptor, spatial HMM and temporal HMM (2D-HMM) model used for facial expression recognition from 3D facial surfaces [51]. The method uses three classes of HMMs: temporal 1D-HMM, pseudo 2D-HMM (a fusion of a spatial HMM and a temporal HMM) and real 2D-HMM. Third, we compared the proposed method with the Local Binary Pattern Three Orthogonal Plane (LBP-TOP), which is used to extract the spatiotemporal features of facial expressions [53]. The method examines two space–time detectors and four descriptors and uses the bag of features framework for human facial expression recognition. Fourth, we compared the proposed method with 3D motion-based features captured using Free-Form Deformations (FFDs) [49], in which a revealing pattern is geared to contain an onset followed by an apex and an offset. Fifth, we have also compared the proposed method with the Dynamic Geometric Image Network (DGIN), which uses different geometric quantities for dynamic 3D facial expression recognition [45]. A two-stage long-term and short-term sliding window scheme is presented for data augmentation and temporal pooling throughout the training phase. We have also compared the proposed method with the method in [54], which extracts the spatiotemporal features using HOG3D and discriminative feature selection and a hierarchical classifier for interpretable facial expression recognition. They showed the

best performance on the BU4D database by extracting real 4D features in each hierarchical classifier layer from depth sequences around onset frames. However, they showed much lower performance than the proposed method with all frames, the same setting as the proposed model. Finally, the proposed method is compared with the Dense Scalar Fields (DSFs) and their deformation magnification used for 4D facial expression recognition [46]. The method uses two growing but diverse suggestions by evaluating the spatial facial deformations using tools from Riemannian geometry and enhancing them using temporal filtering. The proposed method shows comparable results with state-of-the-art performance on the BU4D database, as shown in Table 12.

Table 12. Facial expression recognition performance in different approaches for BU4D database.

Approach	Details	Recognition Performance
Xu et al. [44]	Using BoMW	63.8
Sun et al. [51]	Using 2D-HMM	68.3
Hayat et al. [53]	Using LBP-TOP	71.6
Sandbach et al. [49]	Using FFD	73.4
Li et al. [45]	Using DGIN	92.22
Xue et al. [54]	Using HOG3D (all frames)	82.80
Zhen et al. [46]	Using DSF (whole video)	94.18
Xue et al. [54]	Using HOG3D (onset frames)	96.64
Xue et al. [54]	Using HOG3D (all frames)	82.80
Proposed method	Using modified CNN + SVM	94.69

4. Conclusions and Future Works

This paper presents the state-of-the-art performance of a facial expression classification system obtained by using a hybrid model: a mixture of CNN-based facial motion flow classification and SVM-based facial geometry displacement classification. The hybrid model exhibits higher facial expression classification performance than each individual facial expression classifier. The facial motion flow, which is the extracted facial motion displacement in every pixel, provides data that can be used to extract motion characteristics without a specific network architecture for motion extraction. A CNN architecture that can model facial motion flow is also presented. In addition, we have compared the proposed model with the state-of-the-art models and shows recognition performance of 99.69% and 94.69% for the CK+ and BU4D database, respectively. For the CK+ database, if we compare our recognition performance with [48], which uses a combination of CNN and SIFT, there is an improvement of 0.59%, and if we compare the proposed model with [38], which uses DTAN and DTGN, there is an improvement of 3%. For the BU4D database, if we compare the recognition performance of the proposed model with [54], which uses HOG3D for onset frames, there is a slight degradation of 1.95%. However, if we compare performance with all frames, our proposed model performs better by 11.84%. In addition, if we compare our proposed model with [46], there is an improvement of .51%, which shows the effectiveness of the proposed approach. In conclusion, the proposed method outperformed the others in the case of the CK+ database and is still comparable for the BU4D database with the previous schemes.

The limitation of proposed system is that we have used only western databases for the recognition of facial expressions, but in the future, we will add an eastern database for facial expression recognition. Even though our system is suitable to detect facial expressions, the detection should be interpreted carefully, because of the limited congruency between the facial expression and the emotion actually experienced by a person. The proposed hybrid model could be improved by improving the CNN-based facial expression classification performance. The weighting factor can be further optimized for better classification by determining the optimal weighting for each expression category or by determining another

nonlinear classifier based on each classifier output. In the future, the proposed model can be used for entertainment applications, such as changing music tracks by identifying the mood of a person by using his/her facial expressions.

Author Contributions: Conceptualization, C.-S.L.; Data curation, H.-E.S.; Methodology, J.-C.K.; Software, J.-C.K., M.-H.K. and H.-E.S.; Supervision, C.-S.L.; Writing—original draft, C.-S.L.; Writing—review & editing, M.T.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A1A03040177).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Song, I.; Kim, H.J.; Jeon, P.B. Deep Learning for Real-Time Robust Facial Expression Recognition on a Smartphone. In Proceedings of the IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 10–13 January 2014; pp. 564–567.
2. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124. [CrossRef] [PubMed]
3. Ekman, P. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychol. Bull.* **1994**, *115*, 268–287. [CrossRef] [PubMed]
4. Jack, R.E.; Garrod, O.G.; Yu, H.; Caldara, R.; Schyns, P.G. Facial expressions of emotion are not culturally universal. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 7241–7244. [CrossRef] [PubMed]
5. Du, S.; Tao, Y.; Matrinez, A.M. Compound Facial Expressions of Emotion. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E1454–E1462. [CrossRef]
6. Mayer, J.D.; Roberts, R.D.; Barsade, S.G. Human Abilities: Emotional Intelligence. *Annu. Rev. Psychol.* **2008**, *59*, 507–536. [CrossRef]
7. Barrett, L.F.; Adolphs, R.; Marsella, S.; Martinez, A.M.; Pollak, S.D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. Public Interest* **2019**, *20*, 1–68. [CrossRef]
8. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active shape models: Their training and applications. *Comput. Vis. Image Underst.* **1995**, *61*, 38–59. [CrossRef]
9. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active Appearance Models. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 1998; Volume 2, pp. 484–498.
10. Kähler, K.; Haber, J.; Seidel, H.P. Geometry-based muscle modeling for facial animation. In Proceedings of the Graphics Interface, Ottawa, ON, Canada, 7–9 June 2001; pp. 37–46.
11. Zhang, L.; Snavely, N.; Curless, B.; Seitz, S.M. Spacetime faces: high resolution capture for modeling and animation. *ACM Trans. Graph.* **2004**, *23*, 548–558. [CrossRef]
12. Fasel, B.; Luettin, J. Automatic facial expression analysis: A survey. *Pattern Recognit.* **2003**, *36*, 259–275. [CrossRef]
13. Sandbach, G.; Zafeiriou, S.; Pantic, M.; Yin, L. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image Vis. Comput.* **2012**, *30*, 683–697. [CrossRef]
14. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
15. Carcagnì, P.; Coco, M.D.; Leo, M.; Distante, C. *Facial Expression Recognition and Histograms of Oriented Gradients: A Comprehensive Study*; SpringerPlus: Berlin/Heidelberg, Germany, 2015; p. 645.
16. Shan, C.; Gritti, T. Learning Discriminative LBP-Histogram Bins for Facial Expression Recognition. In Proceedings of the British Machine Vision Conference (BMVC), Leeds, UK, 7–10 September 2008; pp. 100–109.
17. Lajevardi, S.M.; Lech, M. Averaged Gabor Filter Features for Facial Expression Recognition. In Proceedings of the Digital Image Computing: Techniques and Applications, Washington, DC, USA, 1–3 December 2008; pp. 71–76.
18. Zhu, J.; Zou, H.; Rosset, S.; Hastie, T. Multi-class AdaBoost. *Stat. Its Interface* **2009**, *2*, 349–360.
19. Ghimire, D.; Lee, J. Geometric Feature-Based Facial Expression Recognition in Image Squences Using Multi-Class Adaboost and Support Vector Machines. *Sensors* **2013**, *13*, 7714–7734. [CrossRef] [PubMed]
20. Tie, Y.; Guan, L. Automatic landmark point detection and tracking for human facial expressions. *EURASIP J. Image Video Process.* **2013**, *2013*, 8. [CrossRef]
21. Michel, P.; El Kaliouby, R. Real Time Facial Expression Recognition in Video Using Support Vector Machines. In Proceedings of the International Conference on Multimodal Interfaces (ICMI), Vancouver, BC, Canada, 5–7 November 2003; pp. 258–264.
22. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708.
23. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep Face Recognition. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–10 Septermber 2015.

24. Xu, M.; Cheng, W.; Zhao, Q.; Ma, L.; Xu, F. Facial expression recognition based on transfer learning from deep convolutional networks. In Proceedings of the 2015 11th International Conference on Natural Computation (ICNC), Zhangjiajie, China, 15–17 August 2015; pp. 702–708.
25. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
26. Mousavi, N.; Siqueira, H.; Barros, P.; Fernandes, B.; Wermter, S. Understanding how deep neural networks learn face expressions. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 227–234.
27. Zhao, K.; Chu, W.S.; Zhang, H. Deep Region and Multi-label Learning for Facial Action Unit Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3391–3399.
28. Razavian, A.S.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Columbus, OH, USA, 23– 28 June 2014; pp. 512–519.
29. Vo, D.M.; Le, T.H. Deep generic features and SVM for facial expression recognition. In Proceedings of the National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS), Piscataway, NJ, USA, 14–16 September 2016; pp. 80–84.
30. Baveye, Y.; Dellanrea, E.; Chamaret, C.; Chen, L. Deeep Learning vs. Kernel Methods: Performance for Emotion Prediction in Videos. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an China, 21–24 September 2015; pp. 77–83.
31. Patel, D.; Hong, X.; Zhao, G. Selective deep features for micro-expression recognition. In Proceedings of the International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2258–2263. [CrossRef]
32. Niu, X.X.; Suen, C.Y. A Novel Hybrid CNN-SVM Classifier for Recognizing Handwritten Digits. *Pattern Recogn.* **2012**, *45*, 1318–1325. [CrossRef]
33. Hamester, D.; Barros, P.; Wermter, S. Face expression recognition with a 2-channel Convolutional Neural Network. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; pp. 1–8.
34. Yu, Z.; Zhang, C. Image Based Static Facial Expression Recognition with Multiple Deep Network Learning. In Proceedings of the ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 435–442.
35. Kim, Y.; Lee, H.; Provost, E.M. Deep learning for robust feature generation in audiovisual emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 3687–3691.
36. Kotsia, I.; Nikolaidis, N.; Pitas, I. Fusion of Geometrical and Texture Information for Facial Expression Recognition. In Proceedings of the International Conference on Image Processing, Atlanta, GA, USA, 8–11 October 2006; pp. 2649–2652.
37. Jaiswal, S.; Valstar, M. Deep learning the dynamic appearance and shape of facial action units. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–8.
38. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2983–2991.
39. Abdulhussien, W.R.; El Abbadi, N.K.; Gaber, A.M. Hybrid Deep Neural Network for Facial Expressions Recognition. *Indones. J. Electr. Eng. Inform. (IJEEI)* **2021**, *9*, 993–1007. [CrossRef]
40. Krithika, L.B.; Priya, G.G.L. Graph based feature extraction and hybrid classification approach for facial expression recognition. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *12*, 2131–2147. [CrossRef]
41. Kumari, N.; Bhatia, R. Efficient facial emotion recognition model using deep convolutional neural network and modified joint trilateral filter. In *Soft Computing*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–14.
42. Kwon, S.; Mustaqeem. CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network. *Mathematics* **2020**, *8*, 2133.
43. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J. The Extended Cohn-Kanade Dataset(CK+): A complete dataset for action unit and emotion-specific expression. In Proceedings of the CVPR Workshop, Sam Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
44. Xu, L.; Mordohai, P. Automatic Facial Expression Recognition using Bags of Motion Words. In Proceedings of the British Machine Vision Conference;Aberystwyth, UK, 31 August–3 September 2010; pp. 13.1–13.13.
45. Li, W.; Huang, D.; Li, H.; Wang, Y. Automatic 4D Facial Expression Recognition Using Dynamic Geometrical Image Network. In Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 24–30. [CrossRef]
46. Zhen, Q.; Huang, D.; Drira, H.; Amor, B.B.; Wang, Y.; Daoudi, M. Magnifying Subtle Facial Motions for Effective 4D Expression Recognition. *IEEE Trans. Affect. Comput.* **2019**, *10*, 524–536. [CrossRef]
47. Liu, M.; Li, S.; Shan, S.; Chen, X. AU-aware Deep Networks for facial expression recognition. In Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–6. [CrossRef]

48. Tee, C.; Al-Shabi, M.; Cheah, W.; Ong, M.G.K. Facial Expression Recognition Using a Hybrid CNN-SIFT Aggregator. In Proceedings of the International Workshop on Multi-Displinary Trends in Artificial Intelligence, Gadong, Brunei, 20–22 November 2017; Lecture Notes in Computer Science; Volume 10607, pp. 139–149.
49. Sandbach, G.; Zafeiriou, S.; Pantic, M.; Rueckert, D. A dynamic approach to the recognition of 3D facial expressions and their temporal models. In Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition (FG), Santa Barbara, CA, USA, 21–25 March 2011; pp. 406–413. [CrossRef]
50. Dapogny, A.; Bailly, K.; Dubuisson, S. Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit Detection. *Int. J. Comput. Vision* **2018**, *126*, 255–271. [CrossRef]
51. Sun, Y.; Yin, L. Facial expression recognition based on 3D dynamic range model sequences. In Proceedings of the European Conference on Computer Vision (ECCV), Marseille, France, 12–18 October 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 58–71.
52. Happy, S.L.; Routray, A. Automatic facial expression recognition using features of salient facial patches. *IEEE Trans. Affect. Comput.* **2015**, *6*, 1–12. [CrossRef]
53. Hayat, M.; Bennamoun, M.; El-Sallam, A. Evaluation of Spatiotemporal Detectors and Descriptors for Facial Expression Recognition. In Proceedings of the International Conference on Human System Interactions, Perth, Australia, 6–8 June 2012; pp. 43–47.
54. Xue, M.; Mian, A.; Duan, X.; Liu, W. Learning Interpretable Expression-sensitive Features for 3D Dynamic Facial Expression Recognition. In Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–7. [CrossRef]
55. Sun, D.; Roth, S.; Black, M.J. A Quantitative Analysis of Current Practices in Optical Flow Estimation and the Principles behind Them. *Int. J. Comput. Vis.* **2014**, *106*, 115–137. [CrossRef]
56. Zhu, X.; Ramanan, D. Face detection, pose estimation and landmark localization in the wild. In Proceedings of the CVPR, Provience, RI, USA, 16–21 June 2012; pp. 2879–2886.
57. Yu, X.; Huang, J.; Zhang, S.; Yan, W.; Metaxas, D.N. Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model. In Proceedings of the ICCV, Sydney, Austrlia, 1–8 December 2013; pp. 1944–1951.
58. Burgos-Artizzu, X.P.; Perona, P.; Dollár, P. Robust face landmark estimation under occlusion. In Proceedings of the ICCV, Sydney, Austrlia, 1–8 December 2013; pp. 1513–1520.
59. Horn, B.K.; Schunck, B.G. Determining optical flow: A retrospective. *Artif. Intell.* **1981**, *17*, 185–203. [CrossRef]
60. Baker, S.; Scharstein, D.; Lewis, J.P.; Roth, S.; Black, M.J.; Szeliski, R. A Database and Evaluation Methodology for Optical Flow. *Int. J. Comput. Vis.* **2011**, *92*, 1–31. [CrossRef]
61. Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In Proceedings of the Computer Vision-ECCV 2004, LNCS 3024, Prague, Czech Republic, 11–14 May 2004; pp. 25–36.
62. Pérez, J.S.; Meinhardt-Llopis, E.; Facciolo, G. TV-L1 Optical Flow Estimation. *Image Process. Line* **2013**, *3*, 137–150. [CrossRef]
63. Zach, C.; Pock, T.; Bischof, H. A duality based approach for realtime TV-L1 optical flow. In Proceedings of the DAGM Conference on Pattern Recognition, Bonn, Germany, 28 September–1 October 2007; pp. 214–223.
64. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 25, Lake Tahoe, CA, USA, 3–8 December 2012; pp. 1097–1105.
65. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A Training Algorithm for Optimal Margin Classifiers. In Proceedings of the Annual Workshop on Computational Learning Theory (COLT), Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
66. Chang, C.C.; Lin, C.J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
67. Yin, L.; Chen, X.; Sun, Y.; Worm, T.; Reale, M. A high-resolution 3D dynamic facial expression database. In Proceedings of the 2008 8th IEEE International Conference on Automatic Face Gesture Recognition, Amsterdam, The Netherlands, 17–19 September 2008; pp. 1–6. [CrossRef]