sensors

MDPI

Letter

A Novel Bilinear Feature and Multi-Layer Fused Convolutional Neural Network for Tactile Shape Recognition

Jie Chu, Jueping Cai *, He Song, Yuxin Zhang and Linyu Wei

School of Microelectronics, Xidian University, Xi'an 710071, China; jiechu@stu.xidian.edu.cn (J.C.); songhe@stu.xidian.edu.cn (H.S.); yxzhang_77@stu.xidian.edu.cn (Y.Z.); lywei@stu.xidian.edu.cn (L.W.)

* Correspondence: jpcai@mail.xidian.edu.cn; Tel.: +86-029-8820-2505

Received: 1 September 2020; Accepted: 4 October 2020; Published: 15 October 2020

Abstract: Convolutional neural networks (CNNs) can automatically learn features from pressure information, and some studies have applied CNNs for tactile shape recognition. However, the limited density of the sensor and its flexibility requirement lead the obtained tactile images to have a low-resolution and blurred. To address this issue, we propose a bilinear feature and multi-layer fused convolutional neural network (BMF-CNN). The bilinear calculation of the feature improves the feature extraction capability of the network. Meanwhile, the multi-layer fusion strategy exploits the complementarity of different layers to enhance the feature utilization efficiency. To validate the proposed method, a 26 class letter-shape tactile image dataset with complex edges was constructed. The BMF-CNN model achieved a 98.64% average accuracy of tactile shape. The results show that BMF-CNN can deal with tactile shapes more effectively than traditional CNN and artificial feature methods.

Keywords: tactile shape; convolutional neural network; bilinear feature; multi-layer fusion

1. Introduction

Tactile perception is relayed to bionic machines through an electronic skin by transforming external stimuli into sensitive component output changes, thus improving human–computer interaction intelligence [1]. The tactile shape is one of the major physical properties of objects and has been widely used as an important recognizable target in numerous and different fields, including human–computer interaction, intelligent medical treatment, and wearable devices [2,3].

In shape recognition tasks, the pressure information collected by the sensor array is treated as an image for subsequent processing [4–6]. Thus, many artificial design approaches that achieve outperformance in the visual domain have been transferred to tactile shape perception. Lou et al. established a dataset consisting of 18 real objects including combs, balls, cups, etc., and proposed a tactile-SIFT method to extract features, which achieved 91.33% accuracy [7]. Gandarias et al. developed a tactile-SUFR descriptor and trained a support vector machine (SVM) classifier to recognize 5-class objects and 3-class human body parts, and achieved 80% accuracy and fast processing time [8]. Khasnboish et al. applied chain codes to describe the tactile shape of eight woody geometries and four irregular shaped household objects, and a compositive SVM was employed to classify with an average classification accuracy of 93.46% [9]. Through a literature review, it can be found that these artificial local feature-based methods are capable of accomplishing tactile recognition. However, such feature design processes are labor-intensive because the description of features requires researchers to analyze the raw information and consider various invariants, and then manually design them for a specific application. To improve this situation, end-to-end semantic feature approaches

have been proposed that automatically extract features and complete the classification [10]. One of the most prominent methods is the use of convolutional neural networks (CNNs). Due to their excellent 2D and 3D information extraction capabilities, CNNs are suitable for extracting tactile features [11–14]. Gandarias et al. applied several CNN models, including eight transfer learning-based models and three custom-made models to identify 22-class contacted household objects using a high-density sensor array, and achieved over 95% accuracy [15]. Pastor et al. acquired pressure images at different grip forces during active perception in order to form a 3D tactile tensor, which was then fed to 3DCNN. A recognition rate of 98% was obtained on a self-constructed dataset consisting of 24 categories of objects [16]. Cao et al. proposed a residual orthogonal tiling and pyramid convolution ensemble method for tactile recognition based on a robotic arm, and achieved 97.5% accuracy on the HCs-10 dataset [17]. Tsuji et al. constructed a pen-type tactile system, inputting time and amplitude sequences as 2D images, and obtained a roughness recognition accuracy of 82.1% for a single user [18]. Hui et al. developed a fusion model of CNN and long short term memory (LSTM) network. A self-built 14-category of common objects was tested for 14 classifications and four classifications, whose accuracy rates were 94.2% and 95%, respectively [19]. Funabashi et al. used morphology-specific convolutional neural networks for tactile object recognition with a multi-fingered hand and seven different combinations of variations were evaluated; an object recognition rate of over 95% with 20 objects was achieved [20]. Alameh et al. transformed 3D tensorial tactile data into 400 × 400 RGB images, and used multiple mainstream deep CNNs for 3-class touch modalities recognition through transfer learning methods, in which the inception-resnet method achieved the highest recognition rate of 76.9% [21].

However, tactile shape perception remains challenging, which can be attributed to the fact that the raw signal from the sensor inevitably suffers from the following non-ideal effects. Firstly, the density of sensors cannot be comparable to vision, at only ~$10^2/cm^2$ [22]. The resulting tactile images have a low resolution. Thus, the raw shape mapping is inadequate. Additionally, current mainstream neural networks are based on visual images and enhanced learning ability by deepening and widening the structure, and CNNs are no exception [23–25]. However, the tiny size and low resolution of tactile images limit the depth of configurable CNNs. Secondly, due to the requirement of flexibility, sensors are made of a low modulus elastic material [26]. Because of the unavoidable elastic coupling, the unpressed pixels around the pressed pixels are prone to deform during the touch action, generating pseudo outputs [27]. This leads to edge blurring, which is especially intensified in shapes with complex contours. Thus, the raw information is mapped inaccuracy. As a result, low resolution and edge-blurred tactile images are learned inadequately, leading to their misrecognition.

We explore how to improve the performance of tactile shape recognition using CNNs, aiming to handle the inadequate and inaccurate raw information mapping. Firstly, the bilinear feature was developed for tactile shapes to improve the feature extraction capabilities of CNN. The bilinear calculation of features is similar to the quadratic expansion of kernels [28,29]. The bilinear feature contains second-order characteristics of the tactile image that are sensitive to edge and texture information. Here, we use a single-mode CNN to generate bilinear features that allow the potential of the network itself to be fully exploited. Secondly, the multi-layer feature fusion strategy is applied to promote the feature utilization efficiency of CNN without increasing the network depth. The fused multi-layer features incorporate the advantages of both local and global information. Since the essence of CNN is a series of extractors stacked from low to high, and the obtained hierarchical features are the encoding of different layers of information [30], high-layer features describe semantic information but with fewer details, while low-layer features describe more detailed information but suffer from semantic ambiguity [31,32]. In this work, we use a skip connection to fuse the high-layer and low-layer features. The dimensionality of low-layer features is not reduced in the skip connection, and thus more complete details remain.

Therefore, a novel bilinear feature and multi-layer fused convolutional neural network (BMF-CNN) is proposed for tactile shape recognition. Letter shapes were chosen for performance verification because their contour is more complex than household objects and their non-ideal effects are more

obvious. Tactile data were collected through a high-density sensor array. By comparing with the baseline traditional CNN and other manual design methods, the superiority of BMF-CNN was verified. We achieved an average recognition accuracy of 98.64%. This paper focuses on improving the recognition of tactile recognition limited by non-ideal effects. These non-ideal effects of density limitation and edge blurring caused by elastic coupling are unavoidable, and thus our method is equally suitable for other tactile perception tasks. Furthermore, the proposed framework can also further assist readers who intend to use CNNs for the task of inadequate mapping of raw information.

2. The Proposed BMF-CNN

2.1. Introduction to CNNs

The CNNs consist of alternately configured convolutional layers and pooling layers for feature extraction and fully connected layers for forming semantic features. A traditional two-dimensional CNN and feature extraction process is shown in Figure 1. For convenience, we use **C** and **P** to represent the convolutional layer and pooling layer, respectively, and **F-C** for fully-connected layer. The convolution kernels are a set of filters, which generate feature maps by sweeping over the input. Mathematically, the lth output feature maps are calculated as:

$$f^l = \psi(w^l \otimes f^{l-1} + b^l) \tag{1}$$

where w and b represent the convolutional kernels and bias, respectively, and ψ (•) denotes the activation function, usually *ReLU* [33]. *ReLU* is described as:

$$ReLU(x) = \begin{cases} x, \ if \ x > 0 \\ 0, \ if \ x \leq 0 \end{cases} \tag{2}$$

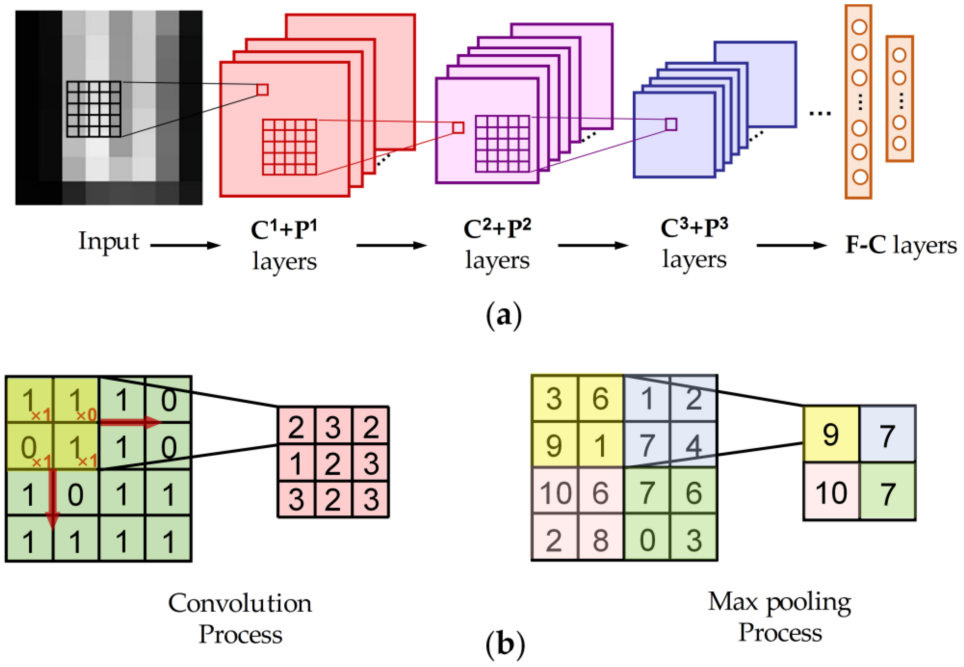


Figure 1. (**a**) The structure of a traditional convolutional neural network (CNN), (**b**) the process of convolution (kernel size: 2 × 2) and max pooling (pooling size: 2, stride: 2).

Because of the constant derivation (0 or 1), the gradient disappearance can be avoided and sparse activation is encouraged. The max pooling layer $\mathbf{P_{max}}$ follows the **C** to perform the down-sampling operation to reduce the feature dimension, which is conducted by:

$$\max pool(m,n) = \max(\alpha(s \times m, s \times n) : \alpha(s \times m + k \times m, s \times n + k \times n)) \quad (3)$$

where s is the pooling stride. Concretely, the process is presented in Figure 1b. Maximum pooling preserves the maximum value within the window and therefore tends to preserve the highlighted textures. Finally, the distributed feature maps are represented to the label space by the **F-C**.

2.2. Single-Mode Bilinear Feature

The single-mode bilinear feature is constructed based on the matrix outer product strategy of the feature map extracted by the same convolutional neural network (red dotted box in Figure 2). The purpose is to represent the second-order characteristics of the tactile shape through expansion of the feature map, in a way similar to a quadratic kernel. These characteristics are very sensitive to edges and textures. Therefore, it is introduced in recognition of the low resolution of tactile shapes accompanied by elastic coupling, aiming to enhance the feature extraction capability of the network without increasing the depth.

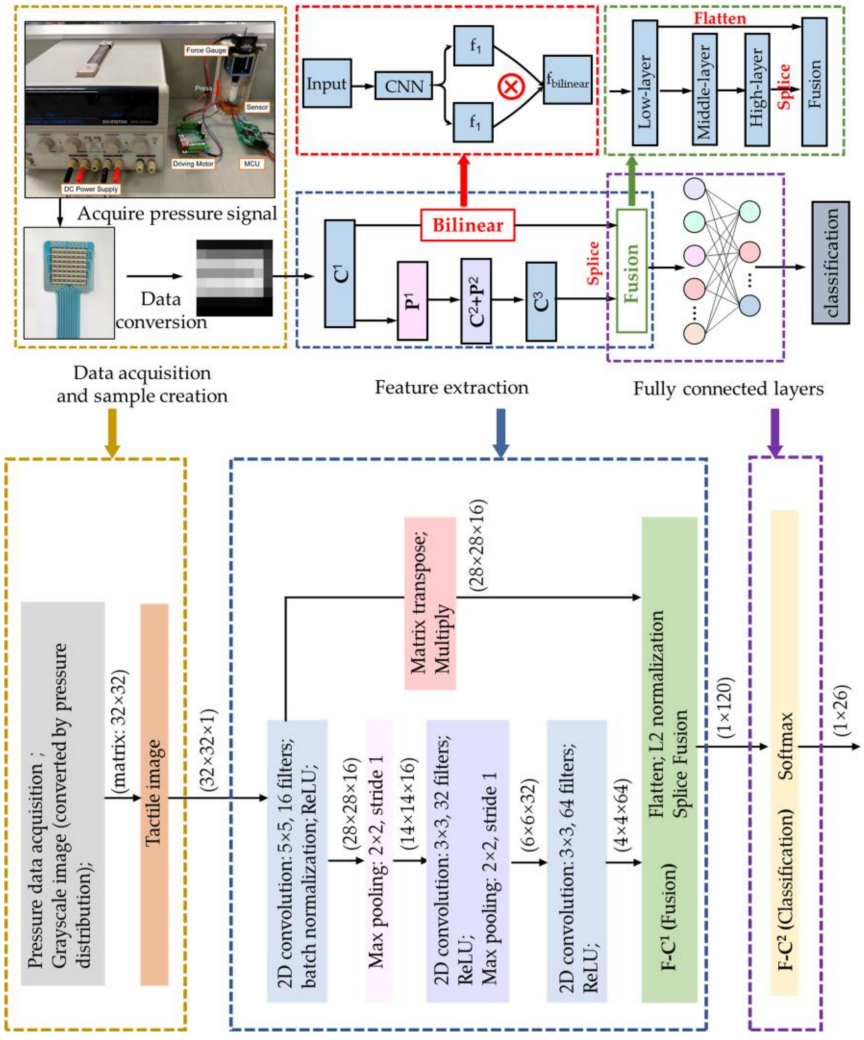


Figure 2. The framework of the proposed bilinear feature and multi-layer fused convolutional neural network (BMF-CNN).

The features extracted by CNN are represented as f, and the corresponding bilinear feature Φ is calculated as:

$$\Phi = bilinear(f^\top f) = \frac{1}{n}(\sum_{i=1}^{n} x_1 x_1^\top) + \varepsilon I \tag{4}$$

This generates a covariance matrix that captures the pairwise interaction between the two features. Since the two features are the same, the obtained matrix is a symmetric positive semi-definite matrix. To further improve the bilinear feature representation, the above result was normalized using an elementwise signed square-root and followed by L2 norm regularization [29]:

$$Z(f, f) = \frac{sign(\Phi) \times \sqrt{\Phi}}{\left\| sign(\Phi) \times \sqrt{\Phi} \right\|_2} \tag{5}$$

Due to the use of second-order features, the performance of single-mode bilinear CNN is better than the strategy of using first-order features for texture and edge-sensitive recognitions.

2.3. Multi-Layer Feature Fusion

Due to the blurring and low resolution of tactile images, the number and quality of the extracted features are limited. Hierarchical features are extracted by alternating convolution and pooling layers. Higher layer features contain mainly semantic information with less local detail, while lower layer features are more concerned with detailed information, but suffer from semantic ambiguity. How can the performance of the network be improved using limited features without increasing network depth? We can deal with this issue with an effective idea: the complementary strengths of features from different layers can be used. We suggest: (1) a high-utilization network should contain different information layers; (2) the combination of bilinear low-layer information and high-layer semantic information can better represent the characteristics of the tactile shape. Therefore, we developed a feature fusion strategy to feed multi-layer features into the network to improve feature utilization.

We apply skip connections to splice low layer features and high layer features for fusion (green dotted box in Figure 2). Feature maps of different layers have different sizes and channels, which cannot be directly fused. It is necessary to flatten the feature map into a 1D feature vector before fusion. If $\{h_1 \times h_1/m\}$ denotes the dimension of low-layer features and $\{h_2 \times h_2/n\}$ denotes the dimension of high-layer features, they are flattened into vectors of length $h_1 \times h_1 \times m$ and $h_2 \times h_2 \times n$, respectively. The feature vectors are then spliced in the fusion layer as:

$$f_{fusion}(f_{low}, f_{high}) = \psi([f_{low}, f_{high}]) \tag{6}$$

whose lengths are $h_1 \times h_1 \times m + h_2 \times h_2 \times n$. Activation values of features from multi-layers vary widely, so the L_2 normalization is used before the fusion to avoid features being overwritten. Since the splice operation does not require feature map dimensional matching, there is no dimension reduction of low-layer features. Hence, more complete local information is retained.

2.4. The BMF-CNN Framework

The proposed BMF-CNN framework is shown in Figure 2. First, the pressure data are converted into grayscale tactile images. Herein, the input size is $32 \times 32 \times 1$. The BMF-CNN consists of three convolutional layers ($\mathbf{C^1}$, $\mathbf{C^2}$, $\mathbf{C^3}$), two max pooling layers ($\mathbf{P^1_{max}}$, $\mathbf{P^2_{max}}$) and two fully connected layers ($\mathbf{F\text{-}C^1}$, $\mathbf{F\text{-}C^2}$). The convolutional layers are used for feature extraction, and the kernels size here is $\{5 \times 5/16\}$, $\{3 \times 3/32\}$, and $\{3 \times 3/64\}$ with a stride of 1. Batch normalization [34] is added after $\mathbf{C^1}$ to improve the convergence speed. The max pooling layer is applied when the convolution, batch normalization, and *ReLU* activation operations are complete, aiming to reduce the data dimensionality and further guarantee feature invariants, including translation and rotation. The pooling size here is 2×2 with a stride of 2. Two features are used for splice fusion: the low-layer features are bilinear

computations of the features extracted by $\mathbf{C^1}$ (bilinear $\mathbf{C^1}$), and the high-layer features are features extracted by $\mathbf{C^3}$. $\mathbf{F\text{-}C^1}$ is used for feature splicing to obtain fused features ($\mathbf{f_{fused}}$) as an input to the classifier. A total of 120 neurons are set in $\mathbf{F\text{-}C^1}$, that is, the 1×1 convolutional kernel is applied to reduce the $\mathbf{f_{fused}}$ to 1×120. The softmax classifier was chosen, which can directly classify multiple categories and select the category with the highest confidence score as:

$$softmax(f)_i = \frac{e^{f_i}}{\sum_j e^{f^j}} = P_i \tag{7}$$

This is consistent with the cross-entropy function in the backpropagation training, which is expressed by:

$$Loss = -\sum_i f_i \log P_i \tag{8}$$

The details of the BMF-CNNs are shown in Table 1. The details of a baseline traditional CNN are also listed.

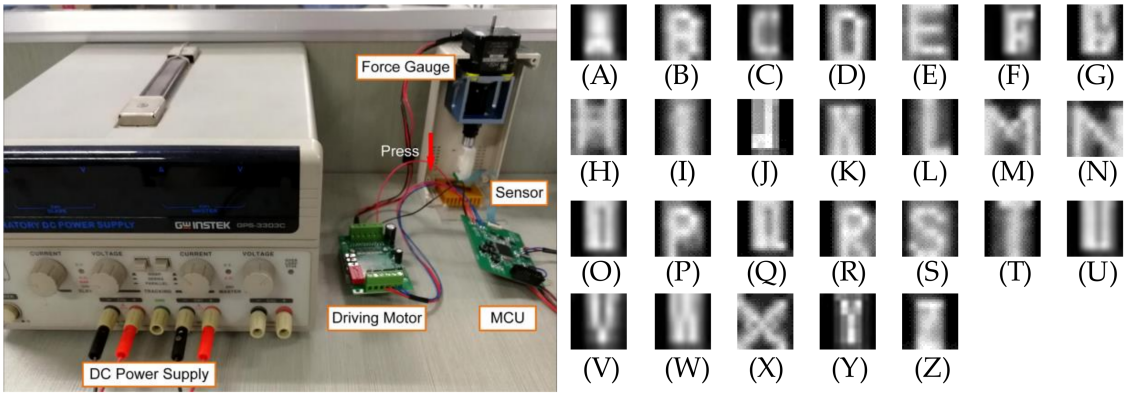
Table 1. Detailed parameters of BMF-CNN and baseline CNN (traditional).

Layer	BMF-CNN	CNN (Traditional)
Input	$32 \times 32 \times 1$	$32 \times 32 \times 1$
$\mathbf{C^1}$	$5 \times 5/16$	$5 \times 5/16$
Batch normalization (BN)	BN	BN
Bilinear	**28 × 28/16**	/
$\mathbf{P^1}$	2×2, stride 2	2×2, stride 2
$\mathbf{C^2}$	$3 \times 3/32$	$3 \times 3/32$
$\mathbf{P^2}$	2×2, stride 2	2×2, stride 2
$\mathbf{C^3}$	$3 \times 3/64$	$3 \times 3/64$
$\mathbf{F\text{-}C^1}$	120 **(Fusion)**	120 **(No Fusion)**
$\mathbf{F\text{-}C^2}$	26	26

3. Experiment and Results

3.1. Dataset and BMF-CNN Training

As shown in Figure 3a, a self-made data acquisition system consisting of a sensor array, a force meter, a drive module, and a microcontroller unit was used for the experiment. A detailed description of the sensor array is shown in Table 2. The letter shapes were chosen because their shapes are more complex than those of household objects, which helps to verify the superiority of our proposed method. A stamp with letter-shaped apophysis was fixed to the force gauge and pressed on the sensor array. The output of the sensor array is a readout with a 20 kHz scanning frequency quantized and stored as a 32×32 matrix. The matrix is converted into a tactile grayscale image. A total of 500 samples per letter are collected within a random pressure range of 30–75 kPa, along with angle and position. To avoid overfitting, data augmentation is applied to expand the sample quantity to 1500 per letter category, including translation, random inversion, and contrast enhancement. These tactile images are divided into 26 categories marked A–Z according to the shape of the stamp. Figure 3b shows the data samples for each letter shape. Five-fold cross validation is used to divide the dataset for training and validation in the learning process. The BMF-CNN weights are initialized by the Xaiver initialization scheme and optimized by the Adam algorithm with hyperparameter parameters 0.9, 0.999, and $\varepsilon = 10^{-8}$ [35,36]. The learning rate was set to 0.001, batch size to 50, and the training epoch to 200. The experiment was implemented in the MatConvNet framework on a PC with an i5-5820 CPU@2.8 GHz.



(**a**) (**b**)

Figure 3. Self-made experimental system. (**a**) Pressure data acquisition system; (**b**) letter shape tactile image samples collected by this system.

Table 2. Parameter details of the sensor array applied in the experiment.

Parameter	Value
Pressure range	0–100 kPa
Temperature range	−25 to + 60 °C
Density	64 pixels/cm^2
Sensitive pixel size	0.7 × 0.7 mm
Sensitive pixel repeatability	−2% to +6%
Number of signal lines	64
Sensor array height	5 mm
Sensor array height	5 mm
Sensor array thickness	0.1 mm

3.2. *Performance of the BMF-CNN*

In the training processes, the losses and gradients of the BMF-CNNs are recorded in Figure 4a. The baseline traditional CNN (details in Table 1) was applied to the same dataset and the results are shown in Figure 4b. The losses of the network dropped rapidly before the 60th epoch, and no overfitting occurred. The gradient vanishing problem did not exist in the training processes of the BMF-CNN and traditional CNN, indicating that our dataset and model are reasonable and matched. It can be observed that the training losses and test losses of the BMF-CNN were lower than those of a traditional CNN. This demonstrates that our model is easy to train.

Figure 5a shows the test accuracies using the BMF-CNN and the baseline CNN during 200 epochs. It can be observed that the BMF-CNN obtained the highest recognition accuracy of 98.66%, which is 5.45% higher than the 93.21% recognition accuracy of the traditional CNN. In addition to the traditional CNN, the SURF-SVM, and SIFT-SVM described in [7,8] were applied as a comparison benchmark to reveal the enhancement of the bilinearity and fusion strategies to the conventional CNN and manual design methods. The recognition accuracy of each method is illustrated in Figure 5b. In order to avoid randomness from affecting the experiment results, five trails were conducted for each method. The BMF-CNN achieved the highest accuracy of 98.64% with a 0.78% standard deviation, the convolutional CNN achieved 93.18% with a 0.73% standard deviation, and the two manual design methods of SURF-SVM and SIFT-SVM achieved 82.93% with a 1.31% standard deviation and 87.71% with a 1.23% standard deviation, respectively. The proposed method thus demonstrated excellent recognition performance and good stability. The recognition time of an image was computed to reveal the complexity of the network. Here we used the test processing time of a single tactile image as an indicator, which only includes the time of the previous propagation and classification of the network,

excluding the processing time of back propagation and network parameter update. As shown in Figure 5b, SURF-SVM achieved the fastest recognition speed of 0.037 ms at the expense of recognition performance, while SIFT-SVM was the slowest at 3.2 ms, and BMF-CNN and the traditional CNN were 0.47 and 0.44 ms, respectively. The fast recognition speed shows that the performance of our BMF-CNN improved without entailing additional computational overhead, which is very important for practical applications. These results verify the effectiveness of the BMF-CNN to improve feature extraction ability and utilization efficiency.

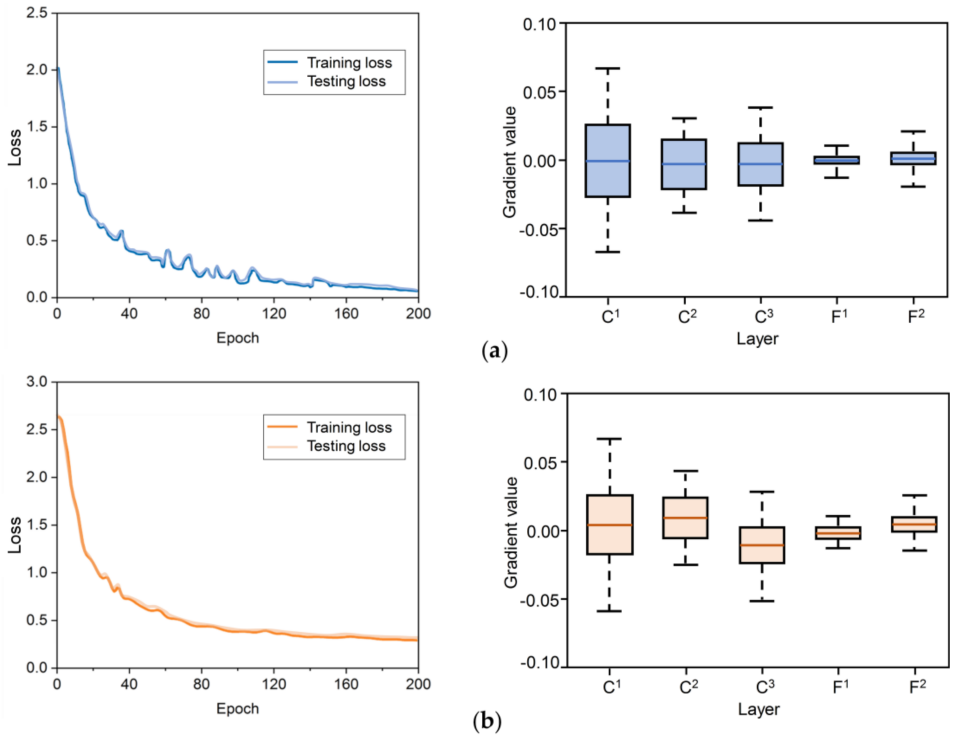


Figure 4. The training processes of the BMF-CNN and baseline CNN (traditional). (**a**) The losses of BMF-CNN and box plot of the gradients of the BMF-CNN; (**b**) the losses of traditional CNN and box plot of the gradients of the traditional CNN.

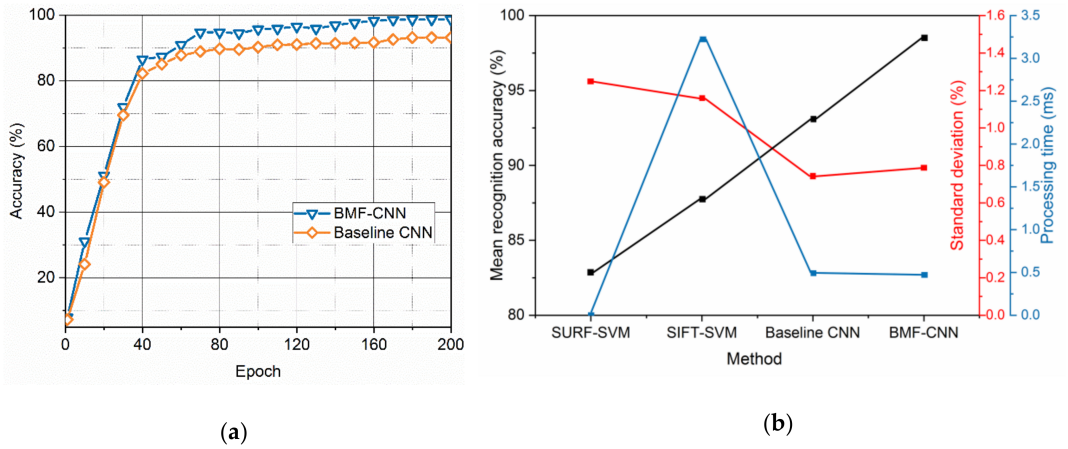


Figure 5. Recognition accuracy. (**a**) The test accuracy of the BMF-CNN and the baseline CNN on the same dataset; (**b**) performance comparison of SURF-SVM, SIFT-SVM, traditional CNN and BMF-CNN (mean recognition accuracy, standard deviation, and processing time).

To unravel the classification results of each tactile shape in detail, a confusion matrix with 26 classifications was drawn, as shown in Figure 6. It can be observed that letter shapes with simple outlines, such as C, I, O, and T, were easier to recognize and obtained higher accuracy. In particular, C and I both reached 100% accuracy. By contrast, samples with complex contours, such as G, R, Q, and B, were more likely to be misclassified due to low pixels and blurred edges. Some Q-shaped samples were erroneously classified as P, R, G, or D. Thus, the accuracy was only 93.2%. This can be attributed to the complex edges and spurious output caused by elastic coupling. Because the density of the sensor array is limited, the tactile image is low-resolution, which leads to insufficient and inaccurate mapping of the original information.

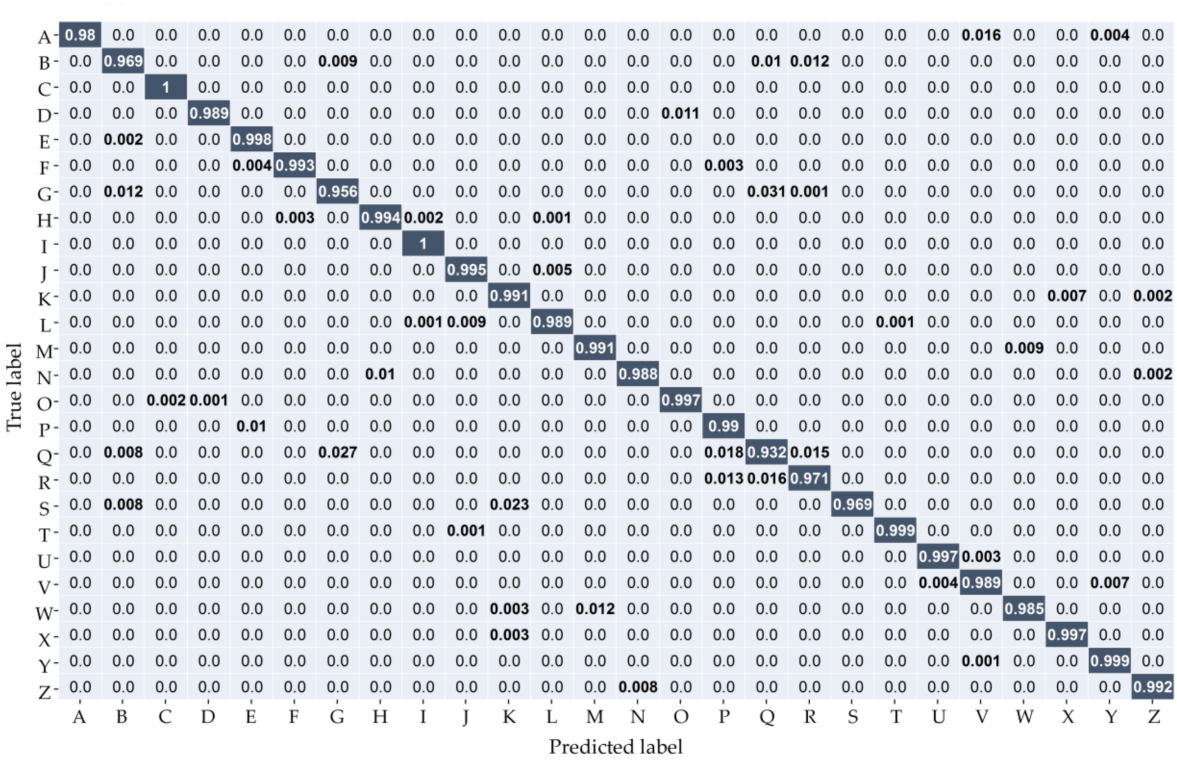


Figure 6. Confusion matrices of BMF-CNN.

3.3. The Contribution of the Bilinear Features and Multi-Layer Fusion Strategies

In order to illustrate the contribution of bilinear features and fusion strategies to network performance in detail, we compared the recognition accuracy of different layers. For easy comparison, Table 3 reports the single-layer extracted feature accuracy. Obviously, the bilinear calculation significantly improved the separability of features. The accuracy increased by 16.5% from 45.6%. This result verifies that the bilinear features promote the feature extraction ability of the CNN. This is attributed to the fact that the bilinear calculation is similar to the feature expansion in the quadratic kernel. Due to the utilization of second-order characteristics, the single-mode BMF-CNN performed better on texture and edge-sensitive recognition tasks. To further reveal the bilinear features in detail, we chose to visualize the nine feature maps with max activation values obtained in $\mathbf{C^1}$ after activating *ReLu* and corresponding bilinear feature maps. For a random sample (Y), Figure 7a shows the top nine feature maps in $\mathbf{C^1}$ and Figure 7b shows the bilinear feature maps. It can be clearly observed that the bilinear features had significant symmetry. In addition, the difference between the feature maps was more obvious, which is of great significance to improve the separability of fusion features.

Table 3. The accuracy of different single layers.

Method	C^1	Bilinear (C^1)	C^2	C^3	F-C^1
CNN (traditional)	0.457	/	0.646	0.807	0.908
BMF-CNN	0.456	0.637	0.647	0.807	0.953

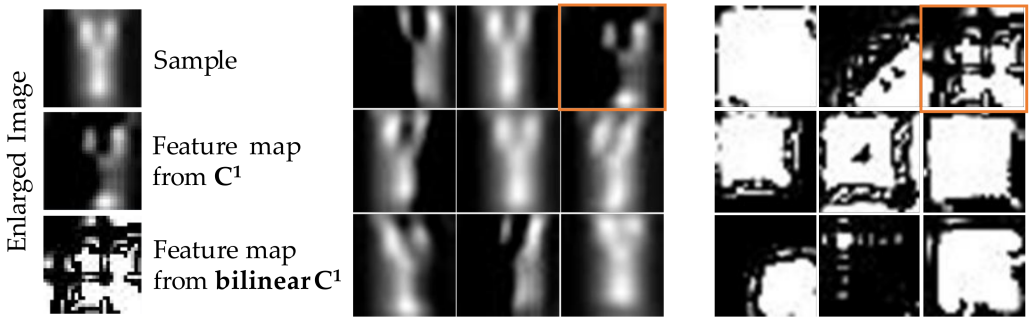


Figure 7. Feature map visualization. (**a**)the top-9 activating feature maps in $\mathbf{C^1}$. (**b**) the corresponding bilinear feature maps of $\mathbf{C^1}$. The enlarged images are marked by orange box.

Our method aims to utilize the complementarity of high semantics and low local details. We explored the combination of different layers to reveal the effectiveness of the fusion strategy. The fusion operation was unified as follows: firstly, features from different layers are normalized by L2, then flattened, and finally spliced in $\mathbf{F\text{-}C^1}$. Here we tested the accuracy of $\mathbf{F\text{-}C^1}$. We used the accuracy of $\mathbf{F\text{-}C^1}$ without fusion, obtained only by $\mathbf{C^3}$, as a benchmark for comparison. In addition, we also calculated the bilinear features extracted from $\mathbf{C^2}$. The results are reported in Table 4. Compared with no fusion, the multi-layer fusion strategy achieved better performance. The accuracy was promoted significantly, with a maximum increase of 4.5%. The improvement of the combination of direct fusion was less than the improvement of the combination of bilinear feature fusion. Moreover, the fusion combinations containing $\mathbf{C^1}$ outperformed the corresponding fusion combinations containing $\mathbf{C^2}$. This is ascribed to the lower layer features describing more sufficient local information. Therefore, the fused features were more distinguishable.

Table 4. The accuracy of different combinations of two layers.

Layers	C^3	Bilinear (C^1) + C^3	C^1 + C^3	Bilinear (C^2) + C^3	C^2 + C^3
Accuracy	0.908	**0.953**	0.927	0.931	0.916

To explain the superiority of bilinear features and multi-layer fusion more qualitatively, t-distributed random neighbor embedding (t-SNE) is used to visualize the learned features [37]. We used t-SNE to reduce the learned features in $\mathbf{F\text{-}C^1}$ to two-dimensions. The mapped features for four methods are shown in Figure 8 (Five categories are randomly displayed, each containing 40 samples). In Figure 8, the shape features of different categories based on two artificially designs overlapped and did not cluster well, which matches the lower mean recognition accuracies shown in Figure 5b. For the traditional CNN, features of the same shape category were clustered, but slight overlap still occurred between different categories. By contrast, for BMF-CNN, the features within each category were tightly clustered and the features of different categories were widely spaced. This means that the features extracted by BMF-CNN have better separability, which is more intuitive for explaining the superiority of our proposed method.

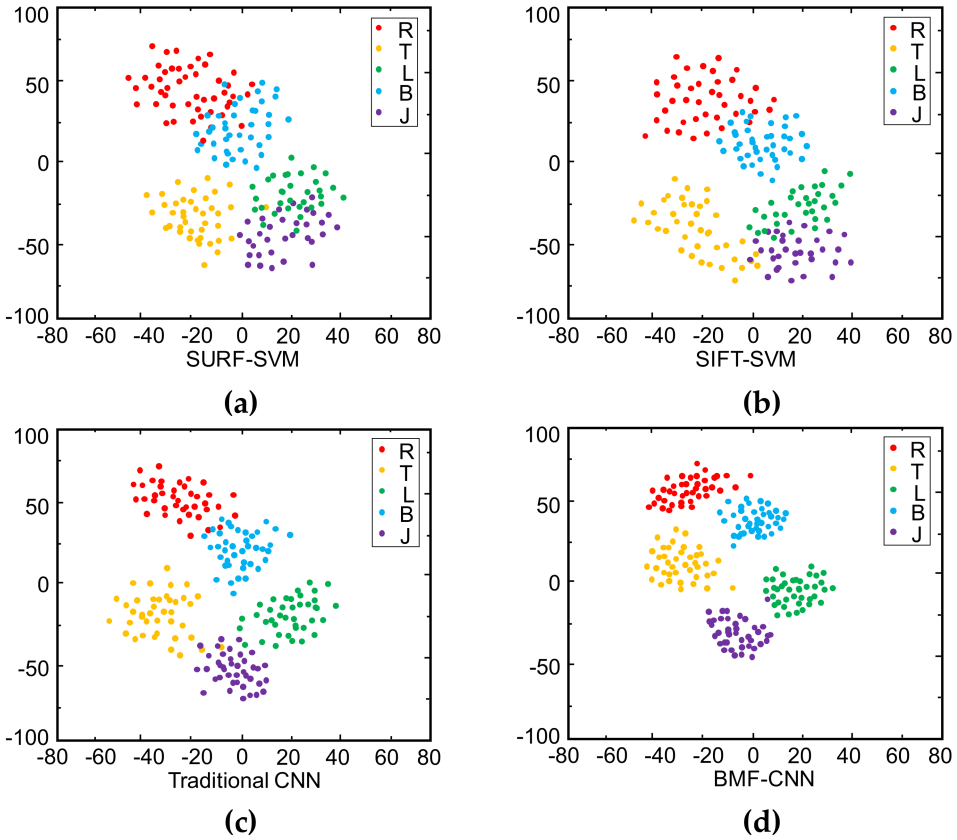


Figure 8. 2D feature visualization using t-distributed random neighbor embedding (t-SNE): (**a**) SURF-SVM, (**b**) SIFT-SVM, (**c**) traditional CNN, (**d**) BMF-CNN.

4. Discussion

In this paper, a BMF-CNN using bilinear features and multi-layer fusion was proposed for low resolution and edge blurring tactile shape recognition. The bilinear feature is able to improve the ability of feature extraction so that the BMF-CNN can learn second-order characteristics better. Multi-layer fusion promotes feature utilization efficiency. For instance, the recognition accuracy is increased without deepening the network, as verified in Tables 3 and 4. Due to non-ideal raw data, environmental interference, and other factors, very small and low-resolution images are common in practical recognition tasks. How to select the architecture of a CNN is still an open issue, and these problems are similar to the non-ideal effects of haptic images. Thus, we attempted to validate our framework on an additional dataset: *MNIST* [38], and achieved 99.21% accuracy (details are in the Supplementary Material). Our framework may enlighten us to handle other low-resolution tasks and we will focus on this topic in future work.

5. Conclusions

A framework called BMF-CNN was proposed for low-resolution and blurred edges tactile shape recognition. In this framework, the bilinear feature is developed to improve the neuron network feature extraction capability and the feature fusion strategy is applied to promoted the feature utilization capability to deal with low resolution and blurred edges. A 26 classes letter shape with complex edges dataset was used to verify the proposed BMF-CNNs. The recognition accuracy was 98.64%, which is much higher than that of traditional CNN and artificial design methods. Furthermore, we analyzed the contribution of bilinear features and fusion strategy to performance in detail through layer-by-layer accuracy verification and feature map visualization.

Supplementary Materials: The following are available online at http://www.mdpi.com/1424-8220/20/20/5822/s1, Figure S1: Examples of MNIST, Figure S2: Accuracy of BMF-CNN on MNIST, Table S1: Performance comparison on MNIST.

Author Contributions: Conceptualization, J.C. (Jie Chu) and J.C. (Jueping Cai); methodology, J.C. (Jie Chu); data acquisition system, H.S.; validation, J.C. (Jie Chu), L.W., Y.Z.; writing (original draft preparation), J.C. (Jie Chu); writing (review and editing), J.C. (Jueping Cai) All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Innovation Wisdom Base for Wide Bandgap Semiconductor and Micro-Nano Electronics of China (B12026), Wuhu and Xidian University special fund for Industry–University–Research Cooperation (012019005), Strategic International Scientific and Technological Innovation Cooperation Key Projects (2016YFE0207000), and Shaanxi Natural Science Basic Research Project (2019JQ-553).

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

References

1. Choi, S.; Lee, H.; Ghaffari, R.; Hyeon, T.; Kim, D.H. Recent Advances in Flexible and Stretchable Bio-Electronic Devices Integrated with Nanomaterials. *Adv. Mater.* **2016**, *28*, 4203–4218. [CrossRef]
2. Yang, J.C.; Mun, J.; Kwon, S.Y.; Park, S.; Bao, Z.N.; Park, S. Electronic Skin: Recent Progress and Future Prospects for Skin-Attachable Devices for Health Monitoring, Robotics, and Prosthetics. *Adv. Mater.* **2019**, *31*, 1904765. [CrossRef]
3. Chortos, A.; Liu, J.; Bao, Z.A. Pursuing prosthetic electronic skin. *Nat. Mater.* **2016**, *15*, 937–950. [CrossRef]
4. Pezzementi, Z.; Plaku, E.; Reyda, C.; Hager, G.D. Tactile-Object Recognition from Appearance Information. *IEEE Trans. Robot.* **2011**, *27*, 473–487. [CrossRef]
5. Suto, S.; Watanabe, T.; Shibusawa, S.; Kamada, M. Multi-Touch Tabletop System Using Infrared Image Recognition for User Position Identification. *Sensors* **2018**, *18*, 1559. [CrossRef]
6. Gastaldo, P.; Pinna, L.; Seminara, L.; Valle, M.; Zunino, R. A Tensor-Based Pattern-Recognition Framework for the Interpretation of Touch Modality in Artificial Skin Systems. *IEEE Sens. J.* **2014**, *14*, 2216–2225. [CrossRef]
7. Luo, S.; Mou, W.X.; Althoefer, K.; Liu, H.B. Novel Tactile-SIFT Descriptor for Object Shape Recognition. *IEEE Sens. J.* **2015**, *15*, 5001–5009. [CrossRef]
8. Gandarias, J.M.; Gomez-de-Gabriel, J.M.; Garcia-Cerezo, A. Human and Object Recognition with a High-Resolution Tactile Sensor. In Proceedings of the 2017 IEEE Sensor, Glasgow, UK, 29 October–1 November 2017; pp. 981–983.
9. Khasnobish, A.; Jati, A.; Singh, G.; Konar, A.; Tibarewala, D.N. Object-Shape Recognition by Tactile Image Analysis Using Support Vector Machine. *Int. J. Pattern. Recogn.* **2014**, *28*, 1450011. [CrossRef]
10. Luo, S.; Bimbo, J.; Dahiya, R.; Liu, H.B. Robotic tactile perception of object properties: A review. *Mechatronics* **2017**, *48*, 54–67. [CrossRef]
11. Li, P.X.; Wang, D.; Wang, L.J.; Lu, H.C. Deep visual tracking: Review and experimental comparison. *Pattern Recogn.* **2018**, *76*, 323–338. [CrossRef]
12. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intel. Neurosc.* **2018**, *2018*, 7068349. [CrossRef]
13. Qi, C.R.; Su, H.; Mo, K.C.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 77–85.
14. Zhu, Z.T.; Wang, X.G.; Bai, S.; Yao, C.; Bai, X. Deep Learning Representation using Autoencoder for 3D Shape Retrieval. *Neurocomputing* **2016**, *204*, 41–50. [CrossRef]
15. Gandarias, J.M.; Garcia-Cerezo, A.J.; Gomez-de-Gabriel, J.M. CNN-Based Methods for Object Recognition With High-Resolution Tactile Sensors. *IEEE Sens. J.* **2019**, *19*, 6872–6882. [CrossRef]
16. Pastor, F.; Gandarias, J.M.; Garcia-Cerezo, A.J.; Gomez-de-Gabriel, J.M. Using 3D Convolutional Neural Networks for Tactile Object Recognition with Robotic Palpation. *Sensors* **2019**, *19*, 5356. [CrossRef]

17. Cao, L.L.; Sun, F.C.; Liu, X.L.; Huang, W.B.; Kotagiri, R.; Li, H.B. End-to-End ConvNet for Tactile Recognition Using Residual Orthogonal Tiling and Pyramid Convolution Ensemble. *Cogn. Comput.* **2018**, *10*, 718–736. [CrossRef]
18. Tsuji, S.; Kohama, T. Using a convolutional neural network to construct a pen-type tactilesensor system for roughness recognition. *Sens. Actuators A Phys.* **2019**, *291*, 7–12. [CrossRef]
19. Hui, W.; Li, H.; Chen, M.; Song, A. Robotic tactile recognition and adaptive grasping control based on CNN-LSTM. *Chin. J. Entific. Instrum.* **2019**, *40*, 211–218.
20. Funabashi, S.; Yan, G.; Geier, A.; Schmitz, A.; Ogata, T.; Sugano, S. Morphology-Specific Convolutional Neural Networks for Tactile Object Recognition with a Multi-Fingered Hand. In Proceedings of the 2019 IEEE Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; p. 18904260.
21. Alameh, M.; Valle, M.; Ibrahim, A.; Moser, G. DCNN for Tactile Sensory Data Classification based on Transfer Learning. In Proceedings of the 2019 15th Conference on Ph.D Research in Microelectronics and Electronics (PRIME), Lausanne, Switzerland, 15–18 July 2019; pp. 237–240.
22. Wang, S.H.; Xu, J.; Wang, W.C.; Wang, G.J.N.; Rastak, R.; Molina-Lopez, F.; Chung, J.W.; Niu, S.M.; Feig, V.R.; Lopez, J.; et al. Skin electronics from scalable fabrication of an intrinsically stretchable transistor array. *Nature* **2018**, *555*, 83–90. [CrossRef]
23. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 21–27 June 2016; pp. 770–778.
24. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
25. Brahimi, S.; Ben Aoun, N.; Ben Amar, C. Improved Very Deep Recurrent Convolutional Neural Network for Object Recognition. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 7–10 October 2018; pp. 2497–2502.
26. Chu, J.; Cai, J.P. Flexible pressure sensors with a highly pressure- and strain-sensitive layer based on nitroxyl radical-grafted hollow carbon spheres. *Nanoscale* **2020**, *12*, 9375–9384. [CrossRef]
27. Li, F.Y.; Akiyama, Y.; Wan, X.L.; Okamoto, S.; Yamada, Y. Measurement of Shear Strain Field in a Soft Material Using a Sensor System Consisting of Distributed Piezoelectric Polymer Film. *Sensors* **2020**, *20*, 3484. [CrossRef]
28. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear CNN Models for Fine-grained Visual Recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1449–1457.
29. Zhang, W.X.; Ma, K.D.; Yan, J.; Deng, D.X.; Wang, Z. Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network. *IEEE Trans. Circ. Syst. Vid.* **2020**, *30*, 36–47. [CrossRef]
30. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Hypercolumns for Object Segmentation and Fine-grained Localization. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 447–456.
31. Yu, W.; Yang, K.Y.; Yao, H.X.; Sun, X.S.; Xu, P.F. Exploiting the complementary strengths of multi-layer CNN features for image retrieval. *Neurocomputing* **2017**, *237*, 235–241. [CrossRef]
32. Sermanet, P.; LeCun, Y. Traffic Sign Recognition with Multi-Scale Convolutional Networks. In Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN), San Jose, CA, USA, 31 July–5 August 2011; pp. 2809–2813.
33. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), Ft. Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
34. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015.
35. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
36. Glorot, X.; Bordes, A. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010; pp. 249–256.

37. Laurens, V.D.M.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
38. Lecun, Y. MNIST Handwritten Digit Database. Available online: http://yann.lecun.com/exdb/mnist/ (accessed on 22 September 2020).