electronics

MDPI

Article

Self-Supervised Facial Motion Representation Learning via Contrastive Subclips

Zheng Sun, Shad A. Torrie, Andrew W. Sumsion and Dah-Jye Lee *

Department of Electrical and Computer Engineering, Brigham Young University, Provo, UT 84602, USA; zsun2@student.byu.edu (Z.S.); st392@student.byu.edu (S.A.T.); andreww9@student.byu.edu (A.W.S.)
* Correspondence: djlee@byu.edu

Abstract: Facial motion representation learning has become an exciting research topic, since biometric technologies are becoming more common in our daily lives. One of its applications is identity verification. After recording a dynamic facial motion video for enrollment, the user needs to show a matched facial appearance and make a facial motion the same as the enrollment for authentication. Some recent research papers have discussed the benefits of this new biometric technology and reported promising results for both static and dynamic facial motion verification tasks. Our work extends the existing approaches and introduces compound facial actions, which contain more than one dominant facial action in one utterance. We propose a new self-supervised pretraining method called contrastive subclips that improves the model performance with these more complex and secure facial motions. The experimental results show that the contrastive subclips method improves upon the baseline approaches, and the model performance for test data can reach 89.7% average precision.

Keywords: facial motion; representation learning; self-supervised learning; biometrics

Citation: Sun, Z.; Torrie, S.A.; Sumsion, A.W.; Lee, D.-J. Self-Supervised Facial Motion Representation Learning via Contrastive Subclips. *Electronics* **2023**, *12*, 1369. https://doi.org/10.3390/electronics12061369

Academic Editors: Antonio Fernández-Caballero and Byung-Gyu Kim

Received: 15 February 2023
Revised: 8 March 2023
Accepted: 11 March 2023
Published: 13 March 2023

1. Introduction

Biometric identification technologies like face recognition [1] and fingerprint recognition [2] are becoming increasingly prevalent in our daily lives. These methods utilize users' inherent biological characteristics, making the authentication process easier and more secure than traditional password verification approaches. However, spoofing a biometric identification system using 3D molds or pictures with distinct textures [3] is still possible. Some recent research studies [4,5] have introduced a new concept that uses facial motion as a passcode while confirming the user's regular facial appearance. The corresponding solutions increase the security of identification procedures by requiring the candidates to present the exact biometric trait together with a correct enrolled facial motion.

This new technology's implementation requires only a regular complementary metal oxide semiconductor (CMOS) image sensor and a single-board computer. It can be easily integrated with many existing face recognition systems without adding new sensors. When the user presents their facial action, the main program captures a sequence of consecutive frames, like a live photo, runs a face detection function over all frames, and saves the sequence of face regions in memory. Then, a deep neural network backend runs inference code to generate descriptions of facial appearance and facial motion. Finally, the decision function compares these descriptions with the recorded ones to reject or grant access to the system. Previous work [4,5] has proven that deep neural network models can generate accurate representations for predefined facial expressions or one-off freestyle facial motions. In this work, we studied compound facial actions, which contain more than one significant facial action in one utterance. The advances in this topic can help increase the complexity of password space and security of the authentication system.

Similar to the previous work, we aim to train a deep neural network model that can take a sequence of facial regions as input and generate an embedding vector to represent

the input facial motion. The network design includes a convolutional neural network (CNN) based image feature extractor, a transformer-based sequence encoder, and a linear layer running dimension reduction and computing the final embedding vector as shown in Figure 1.

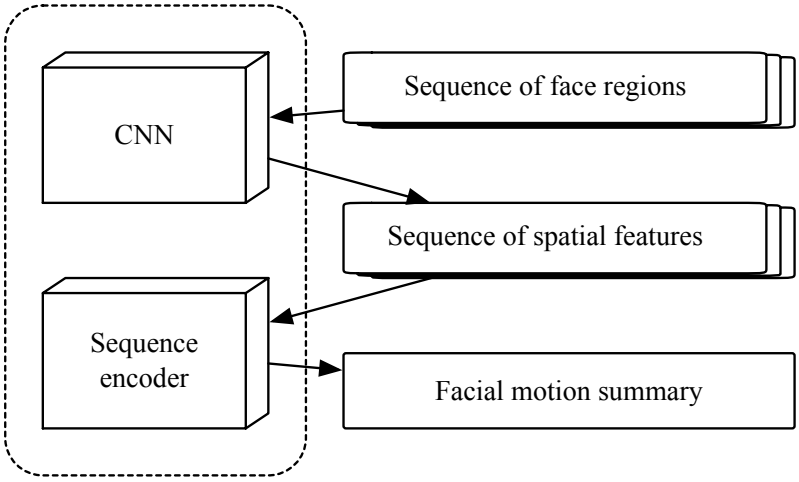


Figure 1. Network architecture. The CNN part uses MobileFaceNet [6] as the backbone and is pre-trained for the facial landmark detection task ahead. The sequence encoder contains two transformer layers and one max-pooling layer (applied to the time axis). The final output of this network is a 1D vector representing the facial motion information in the input sequence.

Because this research focuses on using customized facial motions for identity verification, the clips containing the same subject with the same motion are used to form positive pairs. Clips containing different subjects regardless of the facial motion and the same subject but different facial motions are used to form negative pairs. Our training goal is to increase the embedding similarity for positive clip pairs and lower the similarity for negative clip pairs. In this case, we can compute the pairwise contrastive loss to train the model. This simple strategy works well when all video samples in the dataset contain only one significant facial action. However, it is much harder for the sequence encoder in the network to locate the significant frame or frames for videos with two or more facial actions in a particular order than for those with just one significant facial action. For more secure authentication that requires more complex facial motions containing multiple significant facial actions in one video clip, the model performance using this simple strategy is limited.

Although it is possible to improve the model performance by tuning the hyper-parameters in a larger parameter space, this research explores approaches that lead to more stable performance. Self-supervised learning (SSL) methods have shown their strengths in computer vision tasks like image attribute estimation [7] and video analysis [8]. Unlike the typical supervised learning approaches, SSL does not need labels or curated annotations. It trains the model with unlabeled data in a supervised manner, as the supervision or ground-truth comes from the data itself.

This idea was used in BERT [9] and yielded revolutionary advancements in the language modeling area. They found that self-supervised pretraining on large amounts of data and fine-tuning for a specific task that has limited data greatly improve results. For the image domain, the self-generated labels could be new image views obtained by applying random image transforms to an existing image in the dataset. Even though the dataset used in this research contains labels, we are still interested in SSL's ability to represent facial motions as a pretext task. In this paper, we build our work on these advancements and present an SSL method for the facial motion domain. Our main contributions are as follows:

1. Propose a synthetic facial motion dataset containing compound facial actions;
2. Implement a self-supervised learning approach for facial motion representation learning;
3. Propose a new self-supervised method for facial video analysis, contrastive sub-clips, which improves model performance and outperforms the plain self-supervised pretraining approach.

2. Background

We need a video processing algorithm to parse the facial motion information contained in a video clip. The most related research topic is human action recognition in videos. In the past ten years, deep neural network approaches have dramatically improved video-based computer vision tasks. In addition, many network architectures and learning schemes have emerged to boost action recognition accuracy. These related technologies can help us solve the issues in the facial motion analysis domain.

2.1. Deep Learning and Action Recognition

The major deep architectures for video processing include CNN–RNN [10,11] and 3D CNN [12–14]. They are both uplifted by the outstanding performance of CNNs in image processing [15]. CNN–RNN uses the CNN layers to extract the high-level features in each frame and feeds the generated sequence into stacked RNN cells. Then, it takes the final state of the last cell to infer the output embedding by a linear layer. 3D CNN architectures expand the 2D convolution kernels in established CNN to 3D kernels, and the extra dimension is applied to the time axis. This approach makes it capable of capturing motion information across multiple consecutive frames.

The state-of-the-art method for video-based action recognition is VideoMAE [8]. It uses the vision transformer [16] as the backbone architecture with joint space–time self-attention and employs cube embedding to create video tokens. VideoMAE is also an SSL method, and it demonstrates a masked autoencoder adjusted for video model pretraining. On four standard benchmark datasets of action recognition task, VideoMAE outperformed other existing methods without using any additional data.

2.2. Video-Based Facial Expression Recognition

Another research topic that uses similar tools as our facial motion study is facial expression recognition (FER). It is an essential task in affective computing. The regular models can predict one of seven typical expressions using a face image or video clip. The most recent advancements in this area also use deep neural networks. Fan et al. evaluated both CNN–RNN and 3D CNN for video-based FER accuracy in [17]. They concluded that a parallel hybrid network with these two architectures leads to impressive results. Some later studies [18,19] further proved that the models using the CNN–RNN scheme are practical spatial–temporal feature extractors. In addition, independent 3D CNN architecture and its variants also performed competitively in video-based FER tasks [20,21].

While the datasets used in earlier years in the last decade contain a limited number of videos, three new benchmark datasets [20,22,23] for this topic have emerged in the past two years. They all include more than 10,000 samples collected in real-world scenes. This progress pushes the boundaries of how we design models for video-based FER. Inspired by the success of the transformer unit introduced in [24], Zhao et al. present a network design employing transformer layers for spatial–temporal feature aggregation [25]. The experiments on DFEW [20], one of the three new datasets, show that this new design outperforms all previous architectures. Our model also includes transformer structures.

2.3. Facial Motion and Identity Verification

Our work is not the first demonstration of using facial motion verification to enhance the biometric identification system. In 2018, Yin et al. [4] proposed a method that evaluates facial appearance and expression to verify the user's identity. This work focused on static facial images. Its facial expression branch takes one image as input and predicts one of seven typical facial expressions. If the presented expression matches the enrolled template and the user passes face verification, the decision function will output "pass". The facial expression classifier in this work employs a facial landmark detector and the kernel PCA technique. As cutting-edge facial expression recognition methods use deep neural networks, there is room for improvement.

Our recent work [5] expands this method to verify freestyle facial actions, allowing users to customize their facial motion passwords. It presented a deep neural network model that can process video clips and generate a unified vector representing the facial motion of a video clip. However, only video clips with one significant facial action were investigated. This research work extends it to compound facial actions.

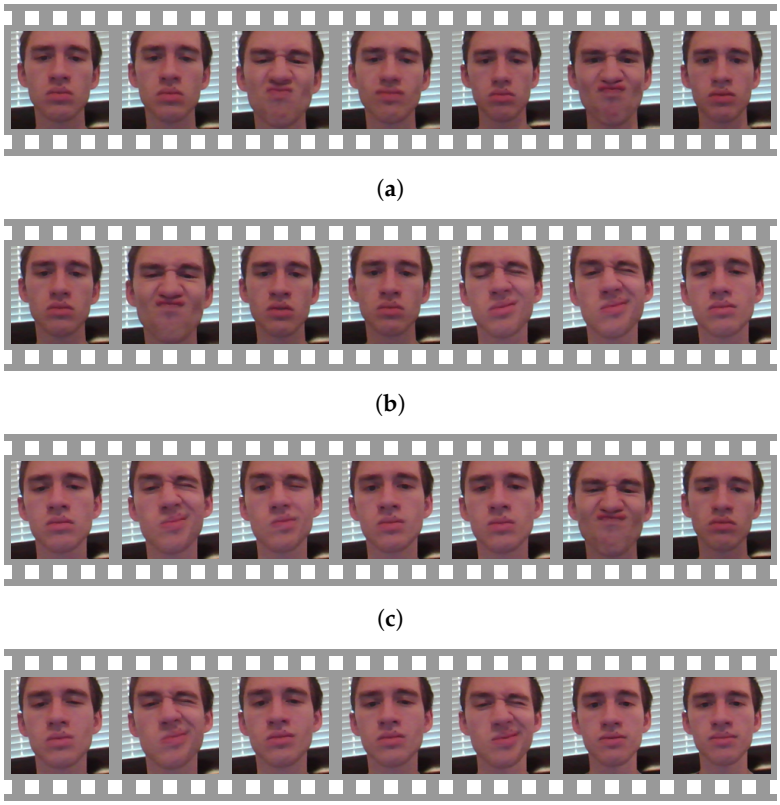
3. Methods

Impressive deep neural network models usually rely on large-volume datasets. This section explains how we collected the data needed for experiments. Also, we discuss the steps to train a baseline model using contrastive loss. Finally, we illustrate the contrastive subclips used in pretraining.

3.1. Compound Facial Actions

The existing public benchmark datasets with facial motion analysis target facial expression recognition. They only include a handful of predefined facial expressions or emotions. This inherent problem limits their usage for exploring the representations of random or user-selected facial motions. Another big challenge in using these datasets for facial motion analysis is the limited number of positive pairs (the same subject making the same facial motion multiple times) for facial motion verification purposes. Therefore, this work uses the original dataset from [5], which contains 59 subjects. All subjects in the dataset made five customized facial motions, and each motion was repeated ten times. Although the recorded user-selected facial motions are not restricted to predefined facial expressions, they are still particular facial actions. One straightforward approach to expand the data for our work is to collect new data with compound motions. Another easier option is to create synthetic compound facial action clips using high-quality samples from the existing datasets. We chose the second method in this work.

For concision, we use D_1 to denote the original dataset containing samples with only one significant facial action and D_2 to represent the synthetic dataset with compound facial actions. If we only look at the facial action intensity of frames in D_1, there are two styles of facial motion. One starts from the neutral face, reaches the apex, then releases and ends in the neutral face. The other also starts from the neutral face but holds the apex until the clip ends. At the forging stage, we work on each subject in D_1 separately. For subject S with N kinds of facial actions, denoted as $\{A_i \mid i \in [1 \dots N]\}$, we can make N^2 different permutations. They correspond to N^2 types of two-way compound facial actions, denoted as $\{C_{ij} \mid i \in [1 \dots N], j \in [1 \dots N]\}$. For each C_{ij}, we create k clips through random padding. All k copies contain the primary motion pieces from clips with A_i and A_j in order, have a random number of neutral frames from A_i padded at the beginning, and a random size of neutral or apex frames from A_j padded at the end. Finally, for subject S, we obtain kN^2 samples with N^2 categories in D_2. Figure 2 shows the clip examples in the final dataset.



(**a**)

(**b**)

(**c**)

(**d**)

Figure 2. Clip examples in our compound facial action dataset. (**a**,**d**) have one type of facial action uttered twice, while (**b**,**c**) both contain two kinds of facial actions. These clips represent four different facial motions in our study.

3.2. Contrastive Learning

We built the baseline model using a contrastive learning scheme. For an input clip pair, the network with tunable parameter set W can generate two embedding vectors, $v_{i,W}$ and $v_{i,W}$. The distance between $v_{i,W}$ and $v_{i,W}$, denoted d_W, is defined as

$$d_W(i,j) = 1 - cos\theta = 1 - \frac{\vec{v_{i,W}} \cdot \vec{v_{j,W}}}{\|\vec{v_{i,W}}\| \|\vec{v_{j,W}}\|}, \tag{1}$$

which represents the cosine similarity of these two vectors. The definitive contrastive loss has the form

$$\mathcal{L}^c(i,j) = y_{i,j} max(d_W(i,j) - m_p,\ 0) + (1 - y_{i,j}) max(m_n - d_W(i,j),\ 0), \tag{2}$$

where $L^c(i,j)$ is a pairwise loss determined by the pair label and distance of two embedding vectors computed by the neural network. All constructed pairs have both clips from the same subject. When two clips show the same compound facial action, it is a positive pair, and $y_{i,j}$ equals 1. The negative pair has clips containing different facial actions, and $y_{i,j}$ is 0. The range of distances that can affect the loss are determined by m_p and m_n.

The data loading program grabs N positive and N negative pairs to form a mini-batch in one feed-forward and back-propagation circle. Different pairs can come from different subjects. The total loss of a mini-batch is

$$\mathcal{L}^c_{total} = \frac{1}{2N} \sum_i \mathcal{L}^c_i. \tag{3}$$

3.3. Contrastive Subclips

SSL methods do not need data annotations in the pretraining phase but depend on data augmentation that creates virtual samples not presented in the dataset. The regular augmentation operators in image study include random crop, color distortion, and Gaussian blur. These operators enable the network to use more significant amounts of data.

In our facial motion study, we must apply the random crop operators carefully, as the network input should contain the whole face region and resize the face to the same scale. We can, nevertheless, use color distortion and Gaussian blur to generate new data instances. With these image transformation methods, we can construct twin clips for any sample in the dataset. The training goal is then to make the network yield identical embedding vectors for twin clips but different vectors for clips made from different samples. Some recent research [26,27] has studied this self-supervised contrastive learning problem and adopted the InfoNCE loss function [28] for pretraining. We implemented our self-supervised pretraining pipeline using InfoNCE as well.

We also noticed a specific characteristic of facial action clips. The noteworthy frames usually appear in the middle of a full clip. If we randomly remove a subclip containing 50% of the frames, the remaining discontinuous frames would probably lose all or partial facial motion information in the original clip. This hypothesis is more assured with compound facial action clips where subjects squeeze two facial actions into a fixed-length time window. By stitching the remaining discontinuous frames, we can obtain a new clip that most likely has different facial motions from the original clip. In a mini-batch of N samples, we can create N more clips using this method. These $2N$ clips are considered distinct in our self-supervised pretraining process. To our knowledge, such a method of combining subclips for facial videos has not been implemented previously for this or similar tasks. We call this novel self-supervised approach contrastive subclips and illustrate its usage in Figure 3.

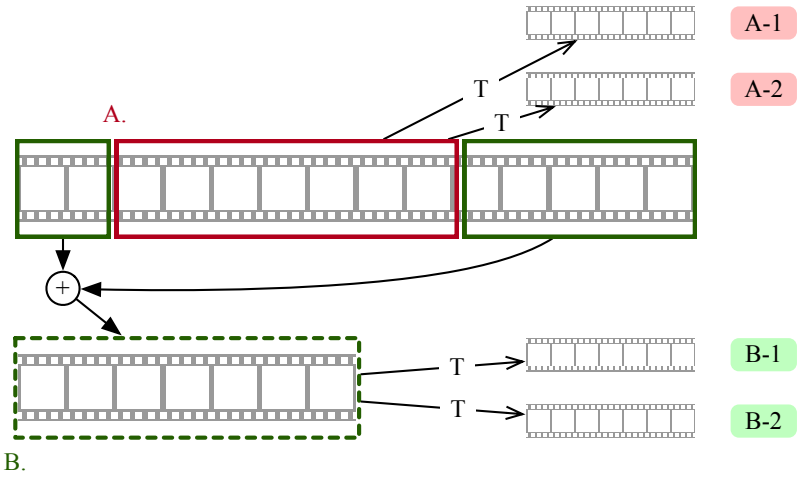


Figure 3. The contrastive subclips. Subclip (**A**) is a random half-length crop of the original sequence, and then we create subclip (**B**) by concatenating the remaining two pieces. T denotes a random instance of the video transform operations containing color distortion, Gaussian blur, and temporal padding. A-1 and A-2 are different subclips but express the same facial motion, and so are B-1 and B-2.

4. Results

After excluding one subject with invalid data, the training set contained 46 subjects, and the validation set included 12. Accordingly, the subjects in the training set were not included in the validation set. The total number of compound motion clips was 11,810. Both pretraining and fine-tuning phases used a stochastic gradient descent (SGD) optimizer with a momentum factor of 0.9 and a cosine annealing scheduler to adjust the learning rate after each gradient update. However, the initial learning rate of fine-tuning phase was one-tenth of that for pretraining.

4.1. Baseline Approach

We obtained the baseline model by training the network from scratch. Besides the hyper-parameters, it is necessary to find a good pair of margins, m_p and m_n, in Equation (2). As no previous research studied the similarities between compound facial actions, we did a grid search in $\{0.1, 0.2, 0.3, 0.5\}$ to look for the best combination. Our program runs a precision-recall (PR) analysis on all possible pairs assembled with clips in the test dataset after each epoch. As shown in Figure 4, we drew the average precision (AP) curve (blue) to indicate the model performance. AP is equivalent to the area under the PR curve, and a more significant AP implies a more robust model. The results show that 0.2 (m_p) and 0.3 (m_n) lead to the highest AP, which is 0.859.

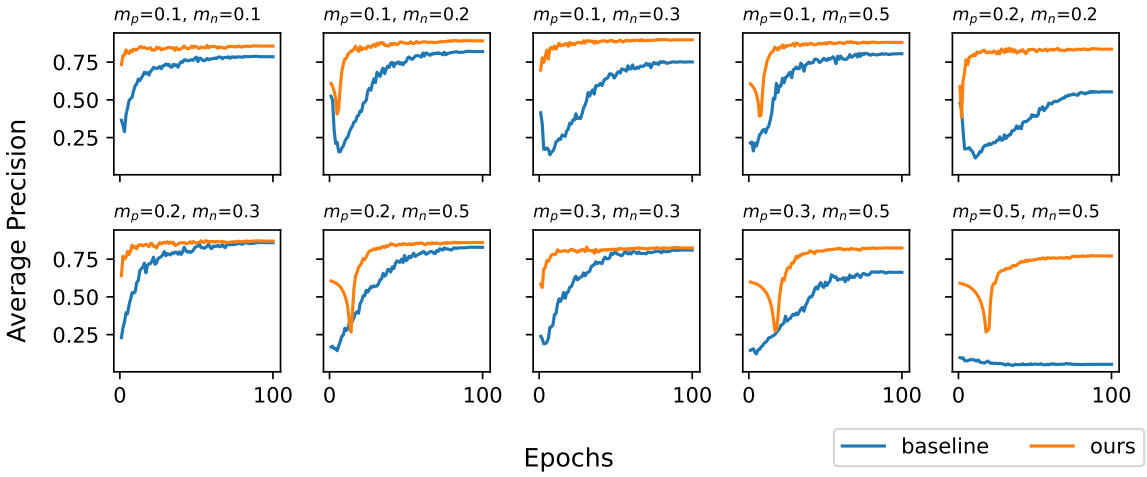


Figure 4. Grid search on distance margins. m_p is the distance margin for positive motion pairs, or the maximum tolerated distance between representation vectors of the same motions. m_n denotes the margin for negative pairs, the minimum distance between two vectors representing different motions without loss penalty. Blue curves denote the baseline model, and orange curves represent our new model. The Y-axis is shared among all subplots.

The other nine blue curves in Figure 4 show that model performance is liable to these two margin parameters. The combination of 0.5 (m_p) and 0.5 (m_n) leads to a model that does not converge. Therefore, if we shift the training to another dataset, we should do this grid search again to find the best combination.

4.2. Self-Supervised Pretraining

During the self-supervised pretraining phase, we did two experiments: one used only image transformations across all frames to create twin clips, and the other manipulated frames to create two new clips in the mini-batch using our contrastive subclips method. When the pretraining phase ended, we used the same model configuration as the baseline and initialized the inside sequence encoder using pretrained parameters. Then, we fine-tuned this model using the same loss function as the baseline.

The results are shown in Figure 5. We used SS to represent the first experiment and SS+ to denote the second experiment with contrastive subclips. We noticed that models with SS and SS+ both converge quickly, and their AP scores reach 0.8 in just ten epochs. However, the final performance of the two models was different. First, SS did not enhance the model's ability, as its curve nearly overlaps with the baseline eventually. While the color distortion or Gaussian blur transform only adds spatial variations that do not change the semantic content of individual frames to the video, it does not involve temporal variations from frame to frame. This theory can explain SS's futility and SS+'s usefulness. SS+ led to an AP equal to 0.897, 3.8% more elevated than the baseline. It indicates that the contrastive subclips method improves upon the plain SSL method for facial videos.

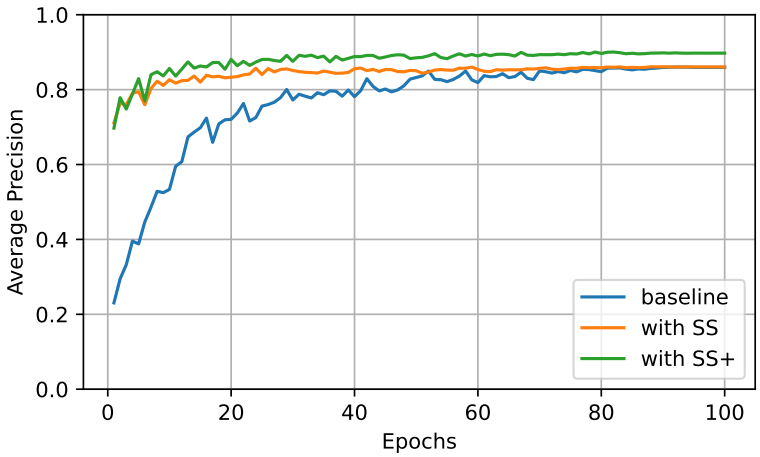


Figure 5. Improvements in self-supervised pretraining. Compared to the baseline approach, the contrastive subclips method (SS+) makes the training process faster to converge and increases the final performance.

We also did a grid search on distance margins for our new approach; see the orange curves in Figure 4. All margin combinations can make the model reach an AP higher than 0.75, meaning that the self-supervised pretraining method makes the fine-tuning process less susceptible to the margin setting. Although combinations like $(0.2, 0.2)$, $(0.3, 0.5)$, and $(0.5, 0.5)$ lead to low AP with the baseline approach, they still show a considerable AP in the new model. These results confirm that our proposed method can improve the model performance and reduce the workload of hyper-parameter tuning.

4.3. Further Study

Because of the improvement obtained by SSL, we plan to bring in more video data during the pretraining phase. While acquiring video clips containing faces from the internet is straightforward, we must carefully deal with the varying head pose in wild videos, which could be largely avoided in our application. The self-supervised method above could fail as the sequence encoder may focus on the head pose changes instead of the facial motions.

There are two possible solutions to this issue. One is taking a head pose estimator [29], which can give the pitch, yaw, and roll combination for a cropped face. When the data loading program grabs video clips for training, it only considers the clips with a stable head pose across all frames. The contrastive learning scheme could learn more about facial motions when the facial appearance and head pose have minimal movement.

The other one is disentangling the facial action features from the head pose on the CNN stage. The CNN model in our approach is truncated from a facial landmark detector, which is associated with the head pose. Its output features differ when input images contain identical facial actions but different head poses. We can replace the landmark detector with an action unit (AU) detector. Most image samples in AU datasets are with frontal faces, so the resultant CNN model could be less accurate with the non-frontal face images. However, one recent study introduced a method that solves this issue using a twin-cycle autoencoder [30]. Evaluating their model for our facial motion study could be a good next step.

5. Conclusions

Our work builds upon previous facial motion representation learning studies. We specifically build upon the work done with the biometric identification technique of distinguishing between users' customized facial motions as an authentication method. As the existing datasets target motions that only include one significant facial action, we need a more complex dataset to expand this technique to motions containing multiple facial actions. We obtained this dataset through a synthetic method, which uses permutations of

different facial actions to create compound facial actions. We then demonstrated that this new dataset confirms that the standard contrastive learning scheme could also handle complicated facial motions, such as those created in our dataset with a combination of multiple facial actions. Our results support the idea that a simple SSL approach for distinguishing between users' customized facial motions does not enhance the model's performance. To overcome this, we propose the novel contrastive subclips method, which enables the model to yield higher-quality motion representations and results in the SSL, performing a more in-depth analysis than was previously attainable in supervised learning technique implementations. We show that this new method improves facial motion representation learning results and increases the accuracy of verifying users based on their customized facial motions.

Author Contributions: Conceptualization, Z.S. and D.-J.L.; Data curation, A.W.S. and S.A.T.; Formal analysis, Z.S., A.W.S. and S.A.T.; Investigation, D.-J.L.; Methodology, Z.S.; Project administration, D.-J.L.; Resources, D.-J.L.; Software, A.W.S. and S.A.T.; Validation, Z.S.; Writing—original draft, Z.S.; Writing—review and editing, A.W.S., S.A.T. and D.-J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets generated during the current study are not publicly available due to their containing information that could compromise the privacy of research participants.

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

References

1. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep Face Recognition. In Proceedings of the British Machine Vision Conference (BMVC 2015), Swansea, UK, 7–10 September 2015; pp. 41.1–41.12.
2. Marasco, E.; Ross, A. A survey on antispoofing schemes for fingerprint recognition systems. *ACM Comput. Surv. (CSUR)* **2014**, *47*, 1–36. [CrossRef]
3. Shmelkin, R.; Friedlander, T.; Wolf, L. Generating master faces for dictionary attacks with a network-assisted latent space evolution. In Proceedings of the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 15–18 December 2021; pp. 01–08.
4. Yin, D.B.M.; Mukhlas, A.A.; Chik, R.Z.W.; Othman, A.T.; Omar, S. A proposed approach for biometric-based authentication using of face and facial expression recognition. In Proceedings of the 2018 IEEE 3rd International Conference on Communication and Information Systems (ICCIS), Singapore, 28–30 December 2018; pp. 28–33.
5. Sun, Z.; Sumsion, A.W.; Torrie, S.A.; Lee, D.J. Learning Facial Motion Representation with a Lightweight Encoder for Identity Verification. *Electronics* **2022**, *11*, 1946. [CrossRef]
6. Chen, S.; Liu, Y.; Gao, X.; Han, Z. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In Proceedings of the Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, 11–12 August 2018; pp. 428–438.
7. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
8. Tong, Z.; Song, Y.; Wang, J.; Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv* **2022**, arXiv:2203.12602.
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
10. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 May 2015; pp. 4694–4702.
11. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 May 2015; pp. 2625–2634.
12. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]
13. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.

14. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–21 June 2018; pp. 6450–6459.
15. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1106–1114.
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
17. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 445–450.
18. Kim, D.H.; Baddar, W.J.; Jang, J.; Ro, Y.M. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Trans. Affect. Comput.* **2017**, *10*, 223–236. [CrossRef]
19. Kuo, C.M.; Lai, S.H.; Sarkis, M. A compact deep learning model for robust facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–21 June 2018; pp. 2121–2129.
20. Jiang, X.; Zong, Y.; Zheng, W.; Tang, C.; Xia, W.; Lu, C.; Liu, J. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2881–2889.
21. Ayral, T.; Pedersoli, M.; Bacon, S.; Granger, E. Temporal stochastic softmax for 3d cnns: An application in facial expression recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 3029–3038.
22. Wang, Y.; Sun, Y.; Huang, Y.; Liu, Z.; Gao, S.; Zhang, W.; Ge, W.; Zhang, W. FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20922–20931.
23. Liu, Y.; Dai, W.; Feng, C.; Wang, W.; Yin, G.; Zeng, J.; Shan, S. MAFW: A Large-Scale, Multi-Modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 24–32.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
25. Zhao, Z.; Liu, Q. Former-dfer: Dynamic facial expression recognition transformer. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 1553–1561.
26. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
27. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 1597–1607.
28. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
29. Hempel, T.; Abdelrahman, A.A.; Al-Hamadi, A. 6d rotation representation for unconstrained head pose estimation. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 2496–2500.
30. Li, Y.; Zeng, J.; Shan, S. Learning representations for facial actions from unlabeled videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 302–317. [CrossRef] [PubMed]