Evaluating Brush Movements for Chinese Calligraphy: A Computer Vision Based Approach

Pengfei Xu[1], **Lei Wang**[1], **Ziyu Guan**[1], **Xia Zheng**[2*], **Xiaojiang Chen**[1], **Zhanyong Tang**[1], **Dingyi Fang**[1], **Xiaoqing Gong**[1] and **Zheng Wang**[3]
[1] School of Information Science and Technology, Northwest University, China
[2] Department of Cultural Heritage and Museology, Zhejiang University, China
[3] MetaLab, School of Computing and Communications, Lancaster University, U. K.
(pfxu, ziyuguan, xjchen, zytang, dyf, gxq) @nwu.edu.cn,
wangleiworks@foxmail.com, z.wang@lancaster.ac.uk.

Abstract

Chinese calligraphy is a popular, highly esteemed art form in the Chinese cultural sphere and worldwide. Ink brushes are the traditional writing tool for Chinese calligraphy and the subtle nuances of brush movements have a great impact on the aesthetics of the written characters. However, mastering the brush movement is a challenging task for many calligraphy learners as it requires many years' practice and expert supervision. This paper presents a novel approach to help Chinese calligraphy learners to quantify the quality of brush movements without expert involvement. Our approach extracts the brush trajectories from a video stream; it then compares them with example templates of reputed calligraphers to produce a score for the writing quality. We achieve this by first developing a novel neural network to extract the spatial and temporal movement features from the video stream. We then employ methods developed in the computer vision and signal processing domains to track the brush movement trajectory and calculate the score. We conducted extensive experiments and user studies to evaluate our approach. Experimental results show that our approach is highly accurate in identifying brush movements, yielding an average accuracy of 90%, and the generated score is within 3% of errors when compared to the one given by human experts.

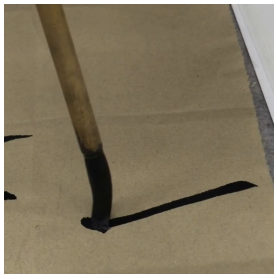
1 Introduction

Chinese calligraphy is a popular visual art form in the Chinese cultural sphere, and is one of the most important culture aspects in countries like China, Japan, Korea and Vietnam. The study of Chinese calligraphy is composed of imitating exemplary works from reputed calligraphers, which requires many years of practices and expert supervision. Correct character strokes and structures, balance and rhythm are essential to the beauty of calligraphy, but all these depend on how well the calligrapher can control the movement of the writing tool – a hair ink brush. Given that competency in Chinese calligraphy requires lengthy practice, many learners do not have access to expert supervision at all time. If we can have an automatic system that tells users how close a written character is to the exemplary one, we can help them to adjust the way the brush is used to improve the writing without expert involvement.

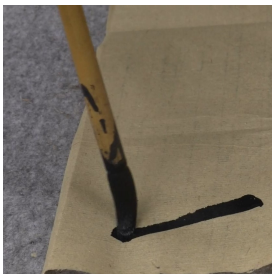
This paper aims to develop such an automatic system to help Chinese calligraphy learners to master brush movements. Our intuition is that when the learners master the correct way of writing, their brush moving trajectories will be similar to those of the exemplary ones. Our intuition is illustrated in Figure 1. In this figure, the instantaneous states of the brush holding by two calligraphy writers, who are considered to have mastered the correct way of writing, are very similar. After watching the recorded videos of how these two users were writing, we discovered that their brush movements are also highly similar. Based on this observation, we have developed a novel approach to use video recording to capture the target user's brush movements during writing. Our key insight is that if we can extract the brush movement trajectories from the video footage, we can then compare them with exemplary writing patterns (i.e., the examples given by the calligraphy teacher in this paper) to quantify the quality of each written character stroke. With this quantified method in place, a calligraphy learner can continuously evaluate his/her writing and improve the brush movement and control with little expert supervision.

Transforming this high-level idea into a practical system is non-trivial. One of the key challenges is to identify when the user started and finished writing a stroke. A naïve solution would be to measure the distance between the brush tip and the paper to infer if the brush was lifted or put down. However, we found that this strategy leads to poor recognition precision. This is because the brush hairs (after dipping ink) are in the same black color as the written character, which in combination of the changing camera shooting angle make it difficult to accurately identify the brush movements. Our solution is to exploit the spatial and temporal movements during writing to infer the brush state – whether it is the start, end or half way

* Corresponding author : zhengxia@zju.edu.cn



(a) The instantaneous states of the brush by user 1



(b) The instantaneous states of the brush by user 2

Figure 1: The brush similarity states when writing a horizontal stroke for two learners. Users who have mastered Chinese calligraphy tend to follow similar patterns when writing the same stroke.

for writing a stroke. Using the spatial information, we can identify which of the basic strokes was likely to be written; and using the temporal information (i.e., consecutive video frames), we can further refine when the brush was lifted or put down by utilizing the structure information of the stroke.

Since we need to recognize the common fine-grained, subtle brush states, it is important to take full advantage of the motion information. While classical machine learning and convolutional neural network (CNN) based methods can be applied to detect the brush movement at the frame level, they only utilize part of the available information. As a result, these approaches fail to detect the fine-grained brush motions, and lead to poor recognition performance. In this paper, we show that to successfully recognize the brush states requires one to leverage not only the spatial information within a single frame but also the temporal information across consecutive video frames. To achieve this requires learning a global description of the video's temporal evolution in addition to the spatial information at the frame level.

This paper presents a novel deep neural network to detect the brush state from video footage. Instead of just using CNNs, we employ the Long Short Term Memory (LSTM) network [Greff *et al.*, 2017] as it has been shown to be useful in classifying and predicting time series. To capture the spatial information within a video frame, we exploit the CNN structure, which is proven to be effective in extracting features and information from images. We show that the combination of multiple convolutional layers and the LSTM (termed MCNN-LSTM in this work) enables us to accurately recognize the brush states. In addition to the MCNN-LSTM model, we also leverage a sophisticated tracking algorithm called TLD. We then use a linear regression model to take in the tracking information to produce a writing score by comparing the detected brush movements to the exemplary writing patterns.

We evaluate our approach by applying it to six calligraphy beginners from our institution. Experimental results show that our approach is highly accurate in evaluating the writing quality, where the given score is with 3% of errors when compared to the one given by human experts (termed *artificial scores*). The key contribution of this work is a novel learning and vision based approach for identifying brush movement states from video footage.

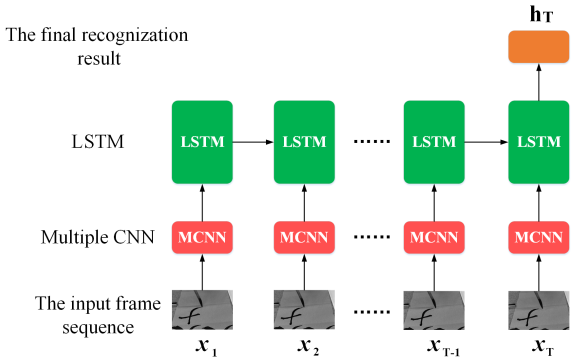


Figure 2: The network structure of brush state recognition.

2 Identifying Brush States

Our MCNN-LSTM model is depicted in Figure 2, where $X = (x_1, x_2, \cdots, x_i, \cdots, x_T)$ represents the input frame sequence, and each element x_i is a frame image. h_T is the final recolonization results. The key of this network structure lies in the design of the convolutional layers.

By analyzing the videos of calligraphy writing, we find these videos have some differences from the common videos of human actions. On the one hand, the movements of the brush and the change of the ink are very imperceptible in several continuous neighboring frames. Therefore, unlike common human behavior recognition, we need to identify the fine-grained actions. On the other hand, we need to handle the videos with a high resolution due to the subtle writing action, but the convolutional layers used to extract the spatial features from these high resolution frames would have too many parameters and too high computational cost. For a traditional CNN, a convolutional layer has a larger or smaller convolution kernel to aware the areas with different sizes. The convolutional layer with a larger convolution kernel can extract the features of bigger movements, but it has limited representation for the subtle movements, which leads to the network cannot focus on the detail features of the moving tip of a writing brush. However, the convolutional layer with a smaller convolution kernel has the exact opposite ability. All these factors make it difficult for a vanilla CNN to extract the spatial features of writing movement accurately.

To solve these problems, we utilize several convolutional layers with smaller convolution kernels and stack them to replace the original, larger convolution kernel. Doing so can reduce the network parameters and result in a deeper network to improve the fitting expression ability. Therefore, we can simultaneously extract the motion features of bigger and subtle movements. This new MCNN consists of four convolutional layer groups, three fully connected layers and the final softmax layer. Figure 3 describes the overall structure of our CNN, and gives the sizes of the convolution kernels, the numbers of convolutional layers, the pooling types adopted in the pooling layer and other important information about our MCNN.

We note that each convolutional layer group of our MCNN has two or three convolutional layers with the convolution kernel size 3×3. The number of the convolution kernels in each layer starts from 64, and doubles at each following layers until it reaches 512. The MCNN increases the

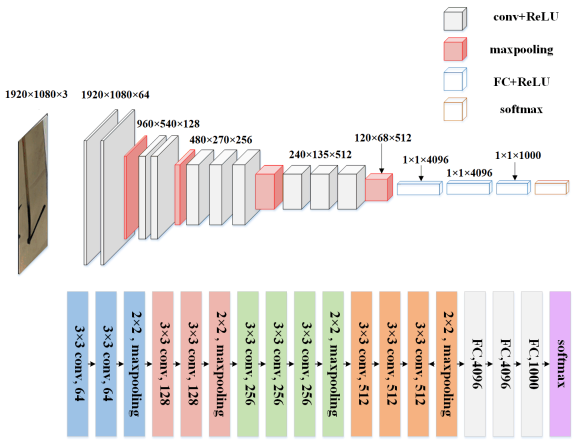


Figure 3: Multiple convolutional neural networks

number of layers of the network to achieve better non-linear fitting, thus leading to more discriminative decision functions. Besides, the MCNN has fewer parameters than a traditional CNN, which makes the processing of high-resolution videos more efficient. We can use a group of frame sequence $X = (x_1, x_2, \cdots, x_t, \cdots, x_T)$ as the input data, and a feature sequence $X' = (x'_1, x'_2, \cdots, x'_t, \cdots, x'_T)$ can be obtained by multiple CNNs, where x'_t is a feature vector of a frame, and can be used as the input of the LSTM to recognize the brush states.

In the whole recognition process, the number T of the input consecutive frames directly affects the recognition accuracy. If we use a short consecutive frame sequence, then the point-in-time of setting the brush to the paper (`SBTP`) and lifting the brush from the paper (`LBFP`) can be determined more exactly. However, the brush movements are very slight, so we need more consecutive frames to extract enough motion features to ensure high recognition accuracy. Besides, different people writing the same stroke may have different writing speed, which also affects the selection of T. In this paper, we set T to 5 to balance the recognition exactitude and accuracy. This threshold is determined via our initial experiments.

3 Evaluating Calligraphy Imitating

The key of calligraphy imitating is to master the movements of the brush, which is directly reflected in the trajectory of the moving brush. In order to extract the brush trajectories in writing each stroke, we utilize the fast RCNN [Ren *et al.*, 2017a; Gao *et al.*, 2017] to replace the detector in traditional TLD [Kalal *et al.*, 2012] to improve its detection performance. Further, a calligraphy learner should also need to have a great control on the structure of a Chinese character, which includes the character sizes, the relative positions and the master-subordinate relationships between strokes. Therefore, we evaluate calligraphy imitating by analyzing the brush trajectories of writing the whole character and its strokes.

As an outcome of the evaluation process, we should give an evaluation score for calligraphy imitating. This is achieved by applying a linear regression model to the obtained brush trajectories. The challenge for this task is that the coordinates of the tracked brush trajectories cannot be directly used as the feature information for the regression model. To solve this problem, we connect the corresponding brush trajectories by calligraphy teachers as the exemplary templates, and then use the Dynamic Time Warping (DTW) [Rasouli and Attaran, 2012] to calculate the similarity between the brush trajectory of the target users and the exemplary patterns. The similarity information is then used as the features for the regression model to produce an evaluation score.

Assuming that a Chinese character has N strokes, the trajectory sequence of calligraphy imitating is $\{P^{(1)}, P^{(2)}, \cdots, P^{(N)}, P^{(N+1)}\}$, and the trajectory sequence of calligraphy templates is $\{Q^{(1)}, Q^{(2)}, \cdots Q^{(N)}, Q^{(N+1)}\}$. In every trajectory sequence, the first N elements represent the trajectories of each stroke, and the $N+1$ element represents the trajectory of the overall Chinese character. Where, the trajectory of the k stroke are $P^{(k)} = \{p_1^{(k)}, p_2^{(k)}, p_3^{(k)}, \cdots, p_t^{(k)}, \cdots, p_n^{(k)}\}$ and $Q^{(k)} = \{q_1^{(k)}, q_2^{(k)}, q_3^{(k)}, \cdots, q_t^{(k)}, \cdots, q_m^{(k)}\}$, $p_t^{(k)} = (x_t^{(k)}, y_t^{(k)})$ and $q_t^{(k)} = (x_t^{(k)}, y_t^{(k)})$ represent the coordinates of the tracked points in frame t, and n and m are the lengths of these two trajectory sequences, then the similarity between them can be calculated by the following equation

$$s_i = DTW(P^{(k)}, Q^{(k)}) = f(n, m) \tag{1}$$

where

$$f(i,j) = \left\| p_i^{(k)} - q_j^{(k)} \right\| + \min \begin{cases} f(i, j-1) \\ f(i-1, j) \\ f(i-1, j-1) \end{cases} \tag{2}$$

$$\begin{aligned} &f(0,0) = 0, f(i,0) = f(0,j) = \infty \\ &(i = 1, \cdots, n; j = 1, \cdots, m) \end{aligned} \tag{3}$$

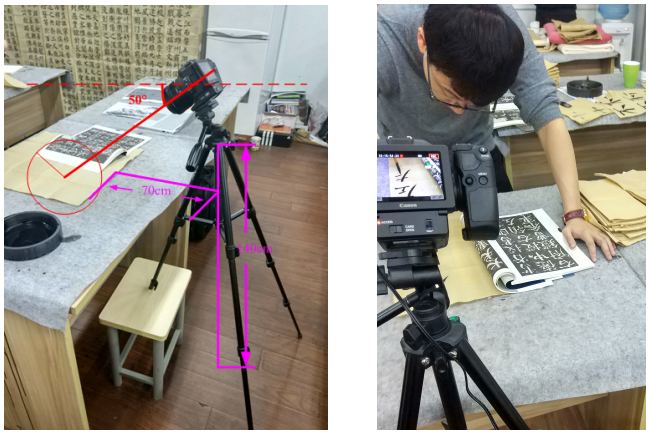
By utilizing DTW, we can obtain a similarity sequences, $S = \{s_1, s_2 \cdots s_N, s_{N+1}\}$, as the features for the regression model, where, the former N elements are the similarity information of each stroke, and the $N+1$ element is the similarity of the overall Chinese character. Then an evaluation score for calligraphy imitating is given by:

$$score = W \cdot S \tag{4}$$

where W is the weight matrix of regression model, and the mean square error between the predicted score and the artificial score, which is given by the calligraphy teachers, is set as the objective function, then we can solve it by gradient descent.

4 Experimental Setup

In order to verify the effectiveness of our approach and its superiority over other methods, we build a video database of Chinese calligraphy writing, and some existing methods and ours are tested on this video database. These high definition videos mainly record the movements of the brush controlled by a calligraphy teacher and his six students in our institution, as shown in Feature 4. We have some parameters for the placement of the camera, and minor changes to these camera parameters have negligible impact on the accuracy, although a large deviation in the parameters would require a calibration of our algorithm, as shown in Figure 4 (a). In this process,



(a) Camera setting (b) Recording scene

Figure 4: Collecting calligraphy writing videos.

these seven calligraphy writers imitated the regular script of Yan Zhenqing style of eleven Chinese characters "永，十，古，大，夫，内，上，不，兄，左，子" for twenty times. Therefore, we have 1540(11 characters × 20 times × 7 person) videos recorded the writing process of each characters, and the videos have resolution of 1920×1080 and frame rate of 50 frames per second. Our approach runs on a workstation with an Intel E5 CPU, a Nvidia GTX 1080 GPU and 64GB of RAM. We compare our approach against a CNN-based method [Yoo, 2017] and a CNN and LSTM mixed approach (term CNN-LSTM) [Ng *et al.*, 2015]. We also compare the writing quality score produced by approach to the one given by human expert.

5 Experimental Results

5.1 Performance of Video Segmentation

In the first stage of the experiment, we need to segment one whole video of writing a character into several sub-videos of writing each stroke. Assuming that the process of writing a stroke is between the process of SBTP and LBFP, so we can recognize the action of SBTP or LBFP by analyzing the states of the brush movements. Then, MCNN-LSTM, CNN and CNN-LSTM are utilized to identify the transformation of writing each stroke. We randomly select 300 videos from the whole set as the training data by labeling the frame sequences of SBTP, writing a stroke and LBFP, and the remaining videos are used as the test data.This process is repeated 10 times. The structure of convolution layers in MCNN-LSTM is design as mentioned above, batch size is 50， and learning rate is set as 0.001. The experiment results are shown in Table 1. CNN (1/1) means the identification accuracy of all the tested frames, CNN (5/5), CNN (4/5) and CNN (3/5) show the identification accuracy of the frame sequence, which is calculated on the tested frames by manual countercheck. CNN (5/5) means that all five frames are identified correctly in a frame sequence with five frames, CNN (4/5) means that there are four frames identified correctly in a frame sequence with five frames, and CNN (3/5) has the similar situation.

From the data shown in Table.1, we can see that CNN [Ng *et al.*, 2015] is difficult to achieve better results for segmenting the videos with temporal information, only the results of CNN (3/5) have the similar accuracy with those of CNN-LSTM. Compared with the identification of a frame sequence by identifying each frame, CNN-LSTM has a better performance. The features obtained by CNN from frame sequences are used as the input of LSTM, which has excellent property for signal processing with temporal characteristics. However, these methods cannot handle the slight movements in the videos well due to their big or small convolution kernel in CNN. By contrast, MCNN has several superimposed convolution layers with smaller convolution kernels, which makes MCNN extract more effective movement characteristics of subtle motion. Further, the features extracted by MCNN from five successive frames are used as the input of LSTM to identify the movement state of the brush. Therefore, MCNN-LSTM has the best performance for the identification of frame sequences from calligraphy writing videos. In Table 1, MCNN-LSTM has average 3.66% higher accuracy than CNN-LSTM, 2.53% than CNN (3/5), 29.71% than CNN (4/5), 64.23% than CNN (5/5), and 17.83% than CNN (1/1). And although the differences between the accuracy of MCNN-LSTM and those of other methods are uneven, MCNN-LSTM is generally superior to these comparison methods, as shown in Figure 5.

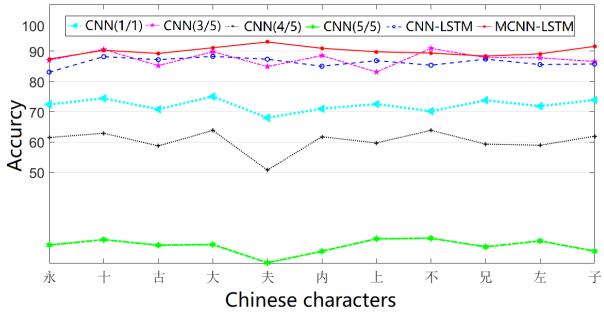


Figure 5: The identification accuracy (%) of different methods.

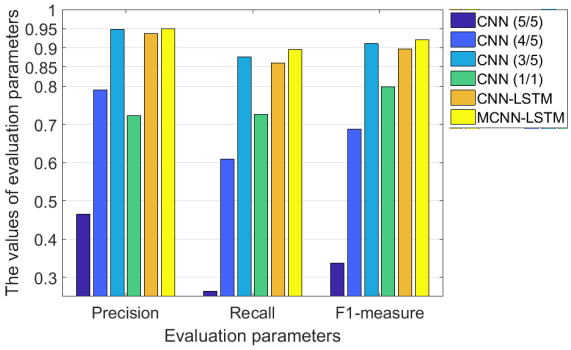


Figure 6: Evaluation parameters of different methods

Table 2 gives precision, recall and F1-score by applying all approaches to all the test videos. MCNN-LSTM outperform all comparative approaches for every single metric. On average, MCNN-LSTM has 1.33% and 22.63% higher precision, a 3.45% and 16.94% higher recall, and a 2.46% and 12.3% better F1-score over CNN-LSTM and CNN (1/1) respectively. The advantages of MCNN-LSTM are also confirmed in Figure 6 that shows the histogram of these evaluation metrics.

Methods	永	十	古	大	夫	内	上	不	兄	左	子
CNN (1/1)	72.43	74.46	70.85	75.05	68.02	71.04	72.55	70.21	73.81	71.86	73.90
CNN (5/5)	26.07	27.86	26.02	26.24	20.24	24.02	28.14	28.31	25.49	27.42	24.06
CNN (4/5)	61.52	62.93	58.77	63.87	50.79	61.77	59.74	63.89	59.36	58.98	61.95
CNN (3/5)	87.03	**90.58**	85.24	89.85	84.92	88.55	83.11	**90.92**	87.91	87.82	86.59
CNN-LSTM	83.09	88.19	87.20	88.29	87.34	85.03	86.86	85.37	87.38	85.58	85.78
MCNN-LSTM	**87.42**	**90.29**	**89.26**	**91.13**	**93.04**	**90.94**	**89.79**	**89.38**	**88.40**	**89.12**	**91.58**

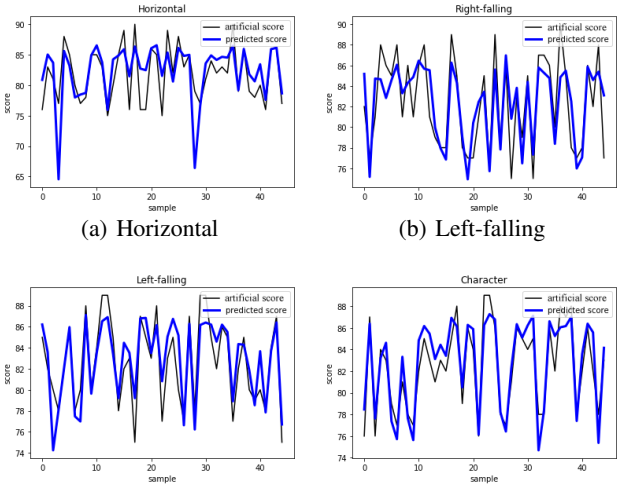
Table 1: The identification accuracy for the brush movements by different methods (%).

Methods	Precision	Recall	F1-score
CNN (1/1)	0.7233	0.7256	0.7984
CNN (5/5)	0.4643	0.2643	0.3368
CNN (4/5)	0.7895	0.6087	0.6874
CNN (3/5)	0.9480	0.8763	0.9107
CNN-LSTM	0.9363	0.8605	0.8968
MCNN-LSTM	**0.9496**	**0.8950**	**0.9214**

Table 2: Precision, recall and F1-score of different methods

5.2 Evaluating Brush Movements

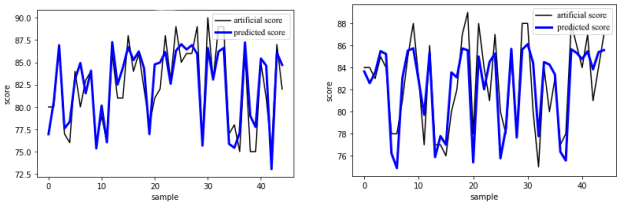
There are subjective and objective evaluations for Chinese calligraphy writing. Firstly, the calligraphy teacher makes a subjective evaluation for the writing works by each student. Then the faster RCNN-TLD is used to track the trajectories of the moving brush for writing each stroke. The accuracy of the brush tracking is 96.7%, and it costs 21 seconds to run on 1000 frames. Finally, DTW is utilized to transform the trajectory information into the motion features, which are used to give predicted scores for Chinese calligraphy imitators. Table 3 shows the predicted scores (PS) and artificial scores (AS) of each strokes and the Chinese character "大" writing by six students. Figure 7 is the curve comparison of the PS and AS for the experiment results of 45 tested videos, which are randomly selected from 120 imitating videos of writing the Chinese character "大", and Figure 8 shows the similarity of the PS and AS of other two characters "不" and "永". From the experimental results, we can see that the scores given by our method largely match the score given by the calligraphy teacher. We also observe similar results for other eight characters which are not shown due to space constraints.



(a) Horizontal (b) Left-falling

(c) Right-falling (d) The whole character

Figure 7: The curve comparison of the PS and AS of each strokes and the whole character of the Chinese character "大".



(a) "不" (b) "永"

Figure 8: The curve comparison of the PS and AS of the Chinese characters "不" and "永".

6 Related Work

Research of Chinese calligraphy digitization to date mainly focuses on two aspects: static and dynamic digitization. Static calligraphy digitization is mainly to analyze and process the static images, including image denoising [Shi *et al.*, 2016], image segmentation [Zheng *et al.*, 2016] and information extraction [Xu *et al.*, 2016] for the tablet and writing works. In 2016, Zheng et al. [Zheng *et al.*, 2016] began to pay attention to the extraction of the spirit information from Chinese calligraphy images for the first time, and proposed a method based on multi channels and guided filters to extract both the form and spirit information simultaneously. There also are some works to solve the deeper issues based on the calligraphy image data bases, such as Chinese calligraphic character recognition [Lin *et al.*, 2013], Chinese calligraphic style recognition [Xia *et al.*, 2013; Zhang and Nagy, 2012], calligraphic writer identification [Brink *et al.*, 2012], evaluating the beauty of calligraphy [Xu *et al.*, 2012; Rasouli and Attaran, 2012], and so on. These methods accomplish their recognition tasks by utilizing the form information of the writing Chinese characters. Besides, the trajectories of calligraphic characters can be used to evaluate and synthesize Chinese calligraphy [Li *et al.*, 2014], and also to match and recognize Chinese calligraphic characters [Zhang and Zhuang, 2012]. Further, the researches of dynamic calligraphy digitization are mainly about virtual brush [Lu *et al.*, 2011] and calligraphy robots [Yao *et al.*, 2004]. While the state, pressure and angle of the brush are estimated by analyzing the form of the written Chinese characters to control the moving of the virtual brush or the brush held by calligraphy robots. Therefore, all these methods are in applicable to our problem.

Chinese gradually attach importance to their own traditional culture, and more and more people begin to learn Chinese cal-

Learner	Horizontal		Left-falling		Right-falling		The character	
	PS	AS	PS	AS	PS	AS	PS	AS
S1	**85**	86	**87**	88	**84**	86	**85**	88
S2	**84**	87	**87**	86	**85**	87	**86**	87
S3	**83**	82	**82**	84	**81**	82	**84**	83
S4	**82**	83	**85**	82	**84**	81	**83**	81
S5	**77**	77	**75**	76	**78**	79	**78**	77
S6	**75**	78	**79**	77	**77**	77	**78**	78

Table 3: Comparing our generated scores with expert-given scores.

ligraphy writing. Due to the serious shortage of professional calligraphy teachers, it is difficult to carry out the extensive of calligraphy teaching. If we can develop a guidance system for calligraphy writing based on videos of calligraphers' writing, which would be very helpful for learners. However, there are few works on this subject. To accomplish our task, we take references from several videos processing methods, including video classification [Ng *et al.*, 2015; Karpathy *et al.*, 2014], human action recognition [Qin *et al.*, 2015; Simonyan and Zisserman, 2014], event detection [Chang *et al.*, 2017; Wang *et al.*, 2017b], image and video captioning [Ren *et al.*, 2017b; Wang *et al.*, 2017a; 2017c], and so on. In particular, the extensive application of deep neural network in recent years significantly improved the performances of video processing methods. For instance, deep CNN can automatically and effectively extract the abstract features from the images, and has achieved good performances in image classification [Krizhevsky *et al.*, 2012]. RNN can extract the temporal information better, and have the property of information memory, which make RNN and its related methods have good performances for sequence signal processing [Donahue *et al.*, 2015; Yoo, 2017]. Most of these works are applied to recognize actions with high activity levels. However, the brush movements are relative smaller and more imperceptible, unlike the standard movements seen in prior action recognition and event detection tasks. Hence, these existing methods would misjudge the motions in Calligraphy videos. Moreover, high-resolution videos are required in our problem to well discern motions with fine-grained distinctions. The existing methods are not suitable for our context since they usually have many parameters and would incur high computational cost.

7 Conclusions and Future Work

This paper has presented a novel approach for helping Chinese calligraphy learners to understand the quality of their ink brush movements. Our approach takes in a video footage that captures the writing process to produce a score to quantify the writing quality. Central to our approach is a novel deep neural network that combines multiple CNNs and a LSTM to exploit the spatial and temporal information of the video footage to detect the brush states. Our network enables one to effectively track the brush movements by using a classical video tracking method. The detected brush movements are then compared with exemplary movements to evaluate the writing quality of each character stroke. We evaluate our approach by applying it to six calligraphy learners, showing that our approach is highly effective in quantifying the writing quality. Future work will extend the proposed approach to other writing styles and evaluate it on an extensive set of Chinese characters. Code and data are available at https://goo.gl/X1J2qR.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China under grant agreements Nos. 61502387, 61522206, 61373118, 61202198, 61672409, 61702415, and 61672427, the Major Basic Research Project of Shaanxi Province under Grant No. 2017ZDJC-31, and the Science and Technology Plan Program in Shaanxi Province of China under grant agreements 2017KJXX-80 and 2016JQ6029; the UK Engineering and Physical Sciences Research Council under grants EP/M01567X/1 (SANDeRs) and EP/M015793/1 (DIVIDEND); and the Royal Society International Collaboration Grant (IE161012).

References

[Brink *et al.*, 2012] A. A. Brink, J. Smit, and M. L. Bulacu. Writer identification using directional ink-trace width measurements. *Pattern Recognition*, 45(1):162–171, 2012.

[Chang *et al.*, 2017] Xiaojun Chang, Zhigang Ma, and Yang Yi. Bi-level semantic representation analysis for multimedia event detection. *IEEE Trans Cybern*, 47(5):1180–1197, 2017.

[Donahue *et al.*, 2015] Jeff Donahue, Lisa Anne Hendricks, and Guadarrama. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 677–691, 2015.

[Gao *et al.*, 2017] Junyu Gao, Qi Wang, and Yuan Yuan. Embedding structured contour and location prior in siamesed fully convolutional networks for road detection. In *IEEE International Conference on Robotics and Automation*, pages 219–224, 2017.

[Greff *et al.*, 2017] Klaus Greff, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE TNNLS*, 28(10):2222–2232, 2017.

[Kalal *et al.*, 2012] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE TPAMI*, 34(7):1409–22, 2012.

[Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, and Shetty. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference*

on Neural Information Processing Systems, pages 1097–1105, 2012.

[Li *et al.*, 2014] Wei Li, Yuping Song, and Changle Zhou. *Computationally evaluating and synthesizing Chinese calligraphy*. Elsevier Science Publishers B. V., 2014.

[Lin *et al.*, 2013] Yuan Lin, Jiangqin Wu, and Pengcheng Gao. Lsh-based large scale chinese calligraphic character recognition. In *Acm/ieee-Cs Joint Conference on Digital Libraries*, pages 323–330, 2013.

[Lu *et al.*, 2011] Weiming Lu, Jiangqin Wu, and Baogang Wei. Efficient shape matching for chinese calligraphic character retrieval. *Journal of Zhejiang University-Science*, 12(11):873–884, 2011.

[Ng *et al.*, 2015] Yue Hei Ng, Matthew Hausknecht, and Sudheendra Vijayanarasimhan. Beyond short snippets: Deep networks for video classification. *CVPR*, 16(4):4694–4702, 2015.

[Qin *et al.*, 2015] Lei Qin, Qiong Hu, and Qingming Huang. Action recognition using trajectories of spatio-tempral feature points. *Chinese Journal of Computers*, 37(6):1281–1288, 2015.

[Rasouli and Attaran, 2012] Atousa Rasouli and Mohammad Attaran. Improve the quality of traditional education of calligraphy in iran by using of collaborative e-learning. *Procedia - Social and Behavioral Sciences*, 51:433–443, 2012.

[Ren *et al.*, 2017a] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2017.

[Ren *et al.*, 2017b] Zhou Ren, Xiaoyu Wang, and Ning Zhang. Deep reinforcement learning-based image captioning with embedding reward. *arXiv*, pages 1151–1159, 2017.

[Shi *et al.*, 2016] Zhenghao Shi, Binxin Xu, and Xia Zheng. An integrated method for ancient chinese tablet images denoising based on assemble of multiple image smoothing filters. *Multimedia Tools & Applications*, 75(19):12245–12261, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 1(4):568–576, 2014.

[Wang *et al.*, 2017a] Qi Wang, Junyu Gao, and Yuan Yuan. A joint convolutional neural networks and context transfer for street scenes labeling. *IEEE TITS*, PP(99):1–14, 2017.

[Wang *et al.*, 2017b] Qi Wang, Jia Wan, and Yuan Yuan. Deep metric learning for crowdedness regression. *IEEE TCSVT*, PP(99):1–1, 2017.

[Wang *et al.*, 2017c] Qi Wang, Jia Wan, and Yuan Yuan. Locality constraint distance metric learning for traffic congestion detection. *Pattern Recognition*, 75, 2017.

[Xia *et al.*, 2013] Yang Xia, Jiangqin Wu, and Pengcheng Gao. Ontology based model for chinese calligraphy synthesis. *Computer Graphics Forum*, 32(7):11–20, 2013.

[Xu *et al.*, 2012] Songhua Xu, Hao Jiang, and Francis C. M. Lau. Computationally evaluating and reproducing the beauty of chinese calligraphy. *IEEE Intelligent Systems*, 27(3):63–72, 2012.

[Xu *et al.*, 2016] Pengfei Xu, Xia Zheng, and Chang Xiaojun. Artistic information extraction from chinese calligraphy works via shear-guided filter. *Journal of Visual Communication & Image Representation*, 40(PB):791–807, 2016.

[Yao *et al.*, 2004] Fenghui Yao, Guifeng Shao, and Jianqiang Yi. Extracting the trajectory of writing brush in chinese character calligraphy. *Engineering Applications of Artificial Intelligence*, 17(6):631–644, 2004.

[Yoo, 2017] Jae Hyeon Yoo. Large-scale video classification guided by batch normalized lstm translator. *arXiv*, pages 1–7, 2017.

[Zhang and Nagy, 2012] Xiafen Zhang and George Nagy. Style comparisons in calligraphy. *Document Recognition & Retrieval XIX*, 8297(2):263–271, 2012.

[Zhang and Zhuang, 2012] Xiafen Zhang and Yueting Zhuang. Dynamic time warping for chinese calligraphic character matching and recognizing. *Pattern Recognition Letters*, 33(16):2262–2269, 2012.

[Zheng *et al.*, 2016] Xia Zheng, Qiguang Miao, and Zhenghao Shi. A new artistic information extraction method with multi channels and guided filters for calligraphy works. *Multimedia Tools & Applications*, 75(14):8719–8744, 2016.