Future Generation Computer Systems 113 (2020) 468–473

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A novel deep quality-aware CNN for image edge smoothening

Hongpeng Zhu, TongCheng Huang *

Shaoyang University, Shaoyang 422000, Hunan, China

ARTICLE INFO

Article history:
Received 21 March 2020
Received in revised form 6 July 2020
Accepted 7 July 2020
Available online 17 July 2020

Keywords:
Edge smoothing
Deep neural network
Multi-view learning
Active learning
Gaussian mixture model

ABSTRACT

Image-edge smoothing is a practical technique in many image-editing implementations such as Visual Aesthetic Enhancement, Action Movie Production, and Privacy Protection. In the literature, there have been many well-adapted smoothing techniques used for image edges such as Gaussian and Bilateral Blurred utilizing visual semantics perceived by the human visual system. However, it is observed that the previously employed image-edge smoothing techniques are not both content-aware and semantically aware. In other words, the image edges are not smoothsimple and efficient, its performance might be less satisfactory. We tend to smooth semantically non-important regions while smoothing semantically important regions slightly in practice. The motivation behind this is that we propose a deep architecture for semantically aware image-edge smoothing. Hence, provided massive image patches extracted from highly aesthetic images crawled from Flickr, we extract multimodal visual features to characterize each. Then, these multimodal features are optimally combined via a multi-view learning framework. Hence, we employ an active learning framework to extract the visually attractive image patches from each image. Thus, we train a deep CNN to find out the edge attributes of actively selected image patches. Utilizing the trained CNN feature of each image patch, we construct a Gaussian mixture model to conduct image-edge smoothing. To verify the proposed method, extensive experimental runs on images with different styles are employed to present the attractiveness of image edges processed by the proposed method. Besides, the efficiency of the proposed method is compared to the well-known edge smoothing algorithms.

1. Introduction

Smoothing Image edge is a typical operation in image processing. The captured images might be noisy or too sharp due to the various imaging environments in practice. When those kinds of cases are attained, we can employ the smoothing edge algorithms to reduce image noises and to try to make the image look milder. In the literature, many image edge-smoothing algorithms have been proposed and commercialized such as the well-known Gaussian blurring. These methods are efficient and easy to implement and have been used for many successful image-processing platforms such as Photoshop. A successful image-edge smoothing tool can facilitate many implementations in practice. When a weak lighting condition in visual aesthetic enhancement is a concern, captured photographs are usually noisy that require an image-edge smoothing technique to remove the noisy points within the photograph. On the other hand, edge smoothing is particularly important to blur some regions such as wrinkles and stains on faces, especially for portrait photography. When privacy issues are a concern, broadcasting the individuals' faces might not be possible under some circumstances like news reports that deal with employing an edge smoothing algorithm to blur the face regions.

Many successful image-edge smoothing algorithms have been suggested in the literature. However, they have been still further away from producing satisfying outcomes due to the following shortcomings.

(1) The previously employed image-edge smoothing algorithms handle all the image edges uniformly, which smooth all the edges within an image not taking into account the visual semantics. By doing so, they produced outcomes of image-edge smoothing that might be sub-optimal. As a result, the edges related to semantically important regions are smoothed less, namely, the portraits and salient architectures. On the other hand, the edges from the non-salient background regions such as the sky and oceans are smoothed heavily even though they are less important based on human visual perception. Hence, a semantically aware image-edge smoothing algorithm could overcome this issue.
(2) Designing an image-edge smoothing algorithm that is semantically aware is a highly difficult task. There are massive-scale image patches within each image and their

* Corresponding author.
E-mail address: tongchenghuang@yeah.net (T. Huang).

https://doi.org/10.1016/j.future.2020.07.019
0167-739X/

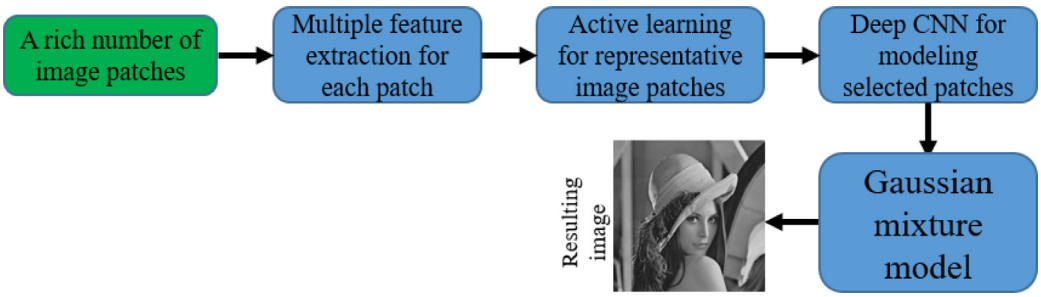

Fig. 1. The overview of the proposed deep-learning utilizing image edge-smoothing framework.

corresponding semantic categories are complex. Designing a flexible semantic model well representing the categories towards these massive-scale image patches is difficult. The previously employed methods have only supported a fixed number of pre-defined semantic categories. On the other hand, how to design a mathematical model effectively finding out the edge attributes from million-scale image patches has been still unsolved. The potential challenges contain (1) how to design an efficient model that fast computes its inherent parameters, and (2) how to leverage the trained mathematical model to guide the image-edge smoothing process.

(3) Evaluating the effect of image-edge smoothing is a subjective task. In other words, users with different aesthetic perceptions, education backgrounds, occupations, and genders might demand different options on the result of the same image-edge smoothing. Hence, how to fuse optimally the experiences of the aesthetic perceptions from multiple users is a tough job. Moreover, we need an effective method to compare the performances of different image-edge smoothing algorithms.

To resolve or to alleviate at least the aforementioned challenges, a novel deep model framework is suggested for semantically aware image-edge smoothing [1–4]. The proposed method can adapt smooth different edges within an image utilizing its semantic saliency. Fig. 1 above depicts the steps of the proposed method. Firstly, we extract the internal image patches from a rich number of training images utilizing features such as color, texture, and visual semantics to capture the appearance and semantics of each image patch. Secondly, we leverage an active learning algorithm to select a few highly representative but less redundant image patches that can contribute to finding out the degree of the image-edge smoothing. Finally, we build a large-scale deep CNN to find out the underlying distribution of such actively discovered image patches. Utilizing the extracted deep features of each image patch, we employ the GMM to find out their distributions, which provide further to guide the process of image-edge smoothing. Comprehensive experimental runs on our compiled data set have presented the visual attractiveness of the outcomes generated by the proposed method that can facilitate many image-edge smoothing applications.

The key contributions of this manuscript are expressed as follows: (1) we proposed a novel method that adaptively conducts image-edge smoothing whose degree can be dynamically tuned utilizing the semantics of each image patch. (2) An active learning algorithm automatically determining a concise set of highly representative image patches from the massive-scale image patches sampled from a collection of images is employed. (3) A deep CNN model converting the actively selected higher quality image patches into the corresponding deep features is conducted utilizing a GMM that can be conducted to guide the degree of edge smoothing.

The rest of the manuscript is organized as follows: Section 2 briefly introduces the previously employed researches closely related to one that we presented in this manuscript. Section 3 provides the proposed method with details containing three key modules, which are called (1) multi-view feature extraction to characterize each image patch, (2) active learning for highly representative and less redundant image patches extraction, (3) a deep learning framework computing the deep representation for each image patch. Section 4 experimentally investigates the proposed algorithm that smooths a deep edge. Section 5 concludes the manuscript and suggests some future work.

2. Related work

2.1. Image modeling based on deep learning

The traditional image processing algorithms employ handcraft features such as SIFT (Scale Invariant Feature Transform) [5] and HOG (Histogram of Oriented Gradients) [6] for both geometry and reconstruction. In 2006, Hinton et al. [7] proposed the concept of deep learning whose layer-by-layer training algorithm could train deep neural networks well. Since then, deep learning and neural networks have received great attention from researchers and widely employed in areas such as image classification, speech recognition, object recognition, and other fields [8]. Ji et al. [8] researched the RGB video-based behavior recognition by constructing a three-dimensional convolutional neural network (CNN) model in which a series of fixed kernel functions were employed firstly to generate multi-channel information for each frame. Then, the 3-dimensional convolution to capture motion information among multiple adjacent frames was utilized. Finally, they combined the information of all channels to attain the latest feature representation. Le et al. [9] combined the advantages of independent subspace analysis (ISA) and CNN that first employed the invariant spatiotemporal features in the ISA learning behavior video. Then, they determined the features as input to the multi-layer CNN network. Thus, utilizing the CNN to find out higher-level and more abstract features of the behavioral video was conducted. Du et al. [10] extracted both spatial and temporal features from optical flow information between video frames and consecutive frames. Then, they employed two deep networks to extract high-level features for human behavior recognition. Tran et al. [11] employed the RGB video whose features were extracted by 3-dimensional convolution directly to achieve better outcomes.

Recurrent neural networks (RNNs) are powerful instruments to extract implicit features in time or spatial domain sequence. The history of the RNN has been gone back to the Elman RNN in the early 1990s [12]. Although the RNN was originally designed to find out the long-term dependencies, a large number of practices suggested that the standard RNNs often had difficulty in achieving the long-term preservation of information. Bengio et al. [13] suggested that the standard RNN had the problems of the gradient

disappearance and explosion, which were caused by the iterative nature of the RNN. Hence, it has not been widely employed since its introduction. To resolve the problem of the long-term dependency, Hochreiter et al. [14] suggested a Long-Short Time Memory (LSTM) network enhance the traditional RNN model. On the other hand, LSTM has also currently become the most efficient sequence model in practical applications. When compared to the hidden unit of the RNN, the internal structure of the hidden unit of the LSTM is more complicated. By adding a linear intervention during the flow of information along with the network, the LSTM can selectively add or reduce information. The RNN has a variety of excellent variant structures such as the Gated Recurrent Unit (GRU), which has been widely employed in applications. Both LSTM and GRU preserve long-term dependencies by appending internal gating mechanisms. Hence, their loop structure only has the dependencies on all past states, and correspondingly, the current state might be dependent on future information. Schuster et al. [15] suggested a Bidirectional Neural Network (BRNN), which can find out the context in two directions on the time domain. The BRNN includes two different hidden layers, and the input is processed in both directions. Graves et al. [16] employed the Bidirectional LSTM (BLSTM) to achieve excellent outcomes in the phenomenon recognition.

2.2. Previous algorithms based on multi-view feature learning

Blum and Mitchell [17] suggested a co-training multi-view learning algorithm under a semi-supervised framework exploiting the complementary information between different views. More specifically, the algorithm that attempts to maximize the consistency of each classifier was conducted by leveraging different views for each co-training iteration. Nevertheless, the above co-training framework greatly depended on the observations whose assumptions are summarized as follows: (1) the redundancy among different views is huge, and (2) each view corresponds to a carefully trained classifier. However, such an assumption cannot be satisfied resulting in affecting the performance of the algorithm badly in practice. In [18], Collins and Singer suggested an upgraded co-training scheme assuming that pairwise views are weakly correlated and non-independent. It can jointly optimize the functions computed by multiple views. Therefore, the inconsistency among multiple views could be minimized. To represent multiple views without redundancy, Nigam suggested a probabilistic co-EM framework attempting to maximize the expectation of classifiers attained from different views. Brefeld suggested a Support Vector Machine (SVM), instead of the Bayes, which is capable of representing small-scale data sets. In [19], Goldman suggested a noise-tolerant multi-view learning framework utilizing the assumption-free approach where multiple views are highly correlated. They updated the paradigm framework to characterize different classifiers. In [20], Zhou et al. suggested a tri-training paradigm free from the restriction when multiple views are redundant and the classifiers must be pre-specified. The algorithm attains a different subset of training features to compute classifiers. Therefore, it derives the unknown features utilizing resemble learning. In [21], Zhou et al. further suggested a co-forest scheme to employ the random forest to replace the classifiers in [20]. More recently, collaborative learning has been deployed in multi-view feature learning. Zhou et al. [20] suggested a semi-supervised multi-view regression model termed GoReg by computing the highly confident unlabeled samples during training. Brefeld et al. [22] suggested a collaborative learning framework to minimize the regression function from different views. In [23], Wang et al. suggested to update the multi-view clustering by computing the correlations among multiple views. Xu et al. [24] suggested the multi-view intact space learning integrating the encoded complementary features from multiple views to generate a latent intact feature representation of the data. Besides, Zhang et al. [25,26] suggested a multi-view subspace clustering that reconstructed the subspace by leveraging the original features utilizing the underlying distribution from multiple views.

3. The proposed method

3.1. Multi-view feature description of each image patch

In the proposed method, we sample a rich number of image samples. Afterward, the visual features of the appearance to represent each sample are extracted. Generally, the features called color, texture, and silhouette are jointly utilized to describe each sample, since they are called three channels complementary to each other in characterizing the appearance of each sampled image patch. In the implementation, while the color channel is represented by the 9-dimensional color moment, the texture channel is featured by the 224-dimensional histogram of gradient (HOG) [22,23]. Finally, the silhouette channel is represented by the 64-dimensional Wavelet transform (WT). The appearance character of each photo is represented by a 297-dimensional feature vector composing of the total dimension.

Besides, we need to represent the semantic features of each image utilizing the aforementioned appearance features. Recently, weak semantic tags have been cheaply available and accessible due to the advancement of massive-scale data retrieval/ classification algorithms. Provided a set of image patches sampled from each image, an L-dimensional augmented feature vector called the *fea* to characterize the distribution of its tags is employed. To alleviate the contaminated weak semantic tags randomly being occurred using computational aspects, the frequently occurring top K most weak semantic tags are attained by ranking. Then, the tag set of each sample is handled as a document called Y. In the proposed method, the weakly supervised semantic encoding to transfer the weak labels into different image patches is employed [1]. Then, the H-dimensional semantic feature for each image patch is attained In detail.

3.2. Active learning to select highly representative image patches

Utilizing the above low/high-level visual features, we combine them into a vector to represent each image patch. In the proposed method, effective active learning is suggested to select multiple highly representative image patches. Utilizing the locality of image patches extracted from one image, each image patch can be linearly reconstructed by its spatial neighbors. Then, the coefficients of the optimal reconstruction are computed by

$$\text{argmin}_H \sum_{i=1}^{N} \left\| z_i - \sum_{j=1}^{N} \mathbf{Q_{ij}} z_j \right\|, s.t., \sum_{j=1}^{N} \mathbf{Q}_j = 1,$$
$$\mathbf{Q}_{ij} = 0, if\ y_j \notin \mathcal{N}(z_i), \quad (1)$$

where $\{z_1, \ldots, z_N\}$ is the post-embedding image patches, $\mathbf{Q_{ij}}$ is a matrix denoting the contribution of the jth image patch to reconstruct the ith image patch, respectively. N denotes the total number of image patches within each image, and $\mathcal{N}(z_i)$ contains the spatial neighboring image patches towards the ith image patch.

To assess the representativeness of the actively selected image patches, an approach is suggested dealing with the reconstruction of the image patch. The reconstruction error of image patches can gauge the quality of the selected image patch. $\{h_1, \ldots, h_N\}$

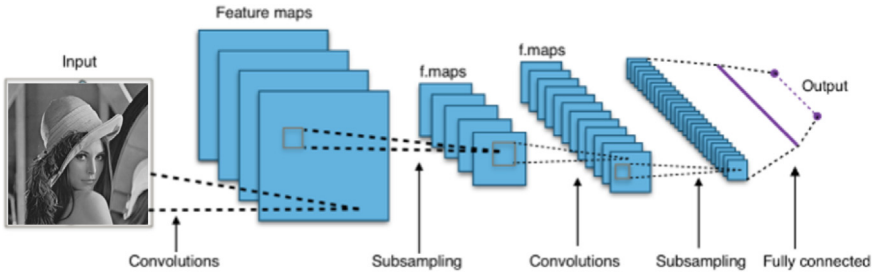

Fig. 2. The overview of the different layers of the proposed hierarchical CNN.

is a set denoting the reconstructed image patches determined by minimizing the objective function defined by

$$\tau\left(\mathrm{h}_1,\ldots,\mathrm{h}_\mathrm{N}\right)=\sum_{i=1}^{K}\left\|h_i-z_{s_i}\right\|^2+\eta\sum_{i=1}^{N}\left\|h_i-\sum_{j=1}^{N}\mathrm{Q}_{ij}h_{ij}\right\|^2, \quad (2)$$

where η is the regularization parameter, and K represents the number of actively selected image patches, $\{\mathrm{h}_1,\ldots,\mathrm{h}_{s_k}\}$ includes the selected image patches, $\{1,\ldots,\mathrm{s_k}\}$ denotes a set of indices of the selected image patches. While the first term is the cost function fixing the entire image patches to the selected image patches, the second term ensures the selected image patches to be locally preserved.

Let $\mathbf{H}=[\mathrm{h}_1,\ldots,\mathrm{h}_\mathrm{N}]$ and $\mathbf{Z}=\{\mathrm{z}_1,\ldots,\mathrm{z}_N\}$ be two sets and $\boldsymbol{\Delta}$, an N × N, is a diagonal matrix indicating whether each image patches is selected. Then, the cost function above can be transformed into the matrix form defined by

$$\tau\left(\mathbf{Z}\right)=\mathrm{tr}\left((\mathbf{H}-\mathbf{Z})^{\mathrm{T}}\,\Delta\,(\mathbf{H}-\mathbf{Z})\right)+\tau\mathrm{tr}(\mathbf{Z}^{\mathrm{T}}\mathbf{V}\mathbf{Z}), \quad (3)$$

where $\mathbf{V}=(\mathbf{I}-\mathbf{H})^{\mathrm{T}}(\mathbf{I}-\mathbf{H})$. To minimize Eq. (3), we set the objective Eq. (3) to zero, and we attain

$$\mathbf{H}=(\tau\mathbf{V}+\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}\mathbf{Y}, \quad (4)$$

Utilizing the derived reconstructed image patches, the reconstructed error is gauged by

$$\tau\left(\mathrm{z}_1,\ldots,\mathrm{z}_{s_k}\right)=\|\mathbf{H}-\mathbf{Z}\|_F^2=\|(\tau\mathbf{V}+\Delta)\,\tau\mathbf{V}\mathbf{Z}\|, \quad (5)$$

By minimizing the objective Eq. (5), the K selected image patches are attained.

3.3. Deep CNN model to determine the degree of edge smoothing

The traditional algorithms used for image modeling always performed global feature extraction utilizing simple clipping. Then, the histogram is constructed employing the extracted features. Although these algorithms enhanced the local performance of the features, they also led to both redundant information and image noise. To resolve this issue, we leveraged the CNN architecture to extract the deep representation of training image patches, since the deep models exhibit its competitive performance in visual representations. Some key patches such as foreground flowers, portraits as well as the architectures are important illustrations for the edge smoothing. Thus, we suggested a hierarchical CNN architecture to extract the deep features of different actively selected image patches.

The hierarchical CNN consists of multiple components depicted in Fig. 2. More specifically, a rich number of actively selected image patches are plugged into the neural network. The architecture of each network consists of four convolution layers and a maximum pooling layer where convolution layers are employed to extract deep features. On the other hand, the maximum pooling layer is employed to downgrade the dimensions of deep features. Then, a fully connected layer associates the four convolution layers and the maximum pooling layer. The second last layer Id is called a binary mask that can indicate the location of different key patches of each image patch. In the implementation, the size of the mask is 36 × 36. The details of the employed CNN architecture are presented in Table 1. Each key patches and face region are resized to 224×224.

Table 1
The details of proposed CNN architecture (While Conv represents convolutional layers, the FC represents fully connected layers).

Layer	Filter Size/Number of kernels
Conv_1	16 × 16/128
Conv_2	7 × 7/256
Conv_3	5 × 5/384
Conv_4	1×1/64
FC_KP1	768
FC_FG	12 048
FC1	2048
FC2	128

Utilizing the proposed CNN architecture, each image patch can be represented by a 128-dimensional feature vector. The proposed architecture can achieve an end-to-end learning model. To downgrade the dimensions of the attained feature vector, we leverage Principle Component Analysis (PCA) to map the feature vectors to a 16-dimensional subspace further, which can optimally represent the original feature vector. This process can be defined by

$$\phi_s=\sum_{i=1}^{N}\|\left(\mathbf{T}+\sum_{i=1}^{N'}a_{ki}E_j\right)-\mathbf{t}_i\|^2, \quad (6)$$

where $\mathbf{T}$ denotes the averaged feature vector extracted from the original image patches $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$. When $\boldsymbol{t}_1$, $\boldsymbol{t}_2$, …, $\boldsymbol{t}_{N'}$ are the eigenvectors corresponding to the N'largest eigenvalues of the scatter matrix, Eq. (6) can attain the minimum value. The adopted PCA reconstructs the image patches on the different principle component dimensions. When the principal components are different, the reconstruction effect of the human face would be different. When the principal component value is larger, the PCA algorithm has a better effect on the reconstruction of the image patch.

Utilizing the PCA, we attain the deep features of each image patch. Then, the GMM is employed to find out their distribution, i.e.,

$$\mathrm{degree}=\sum_{h=1}^{H}\beta_h\mathcal{N}(\mu_h,\,\Sigma_h), \quad (7)$$

where H denotes the number of Gaussian components, both u and Σ are the parameters of each Gaussian component, β_h denotes the weight of the hth Gaussian component. The calculated degree indicates the edge smoothing degree of each image patch (see Fig. 3).

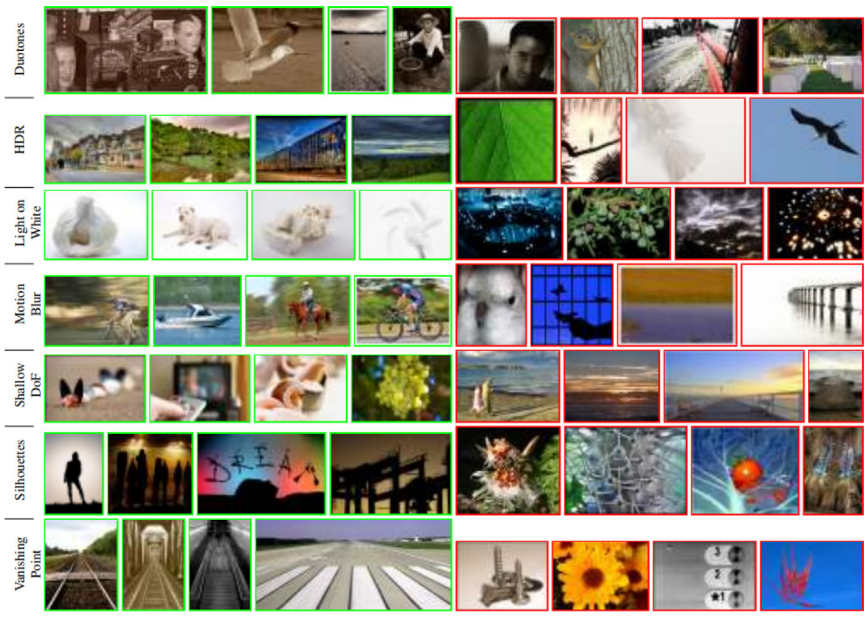

Fig. 3. The sample images collected from the AVA [2].

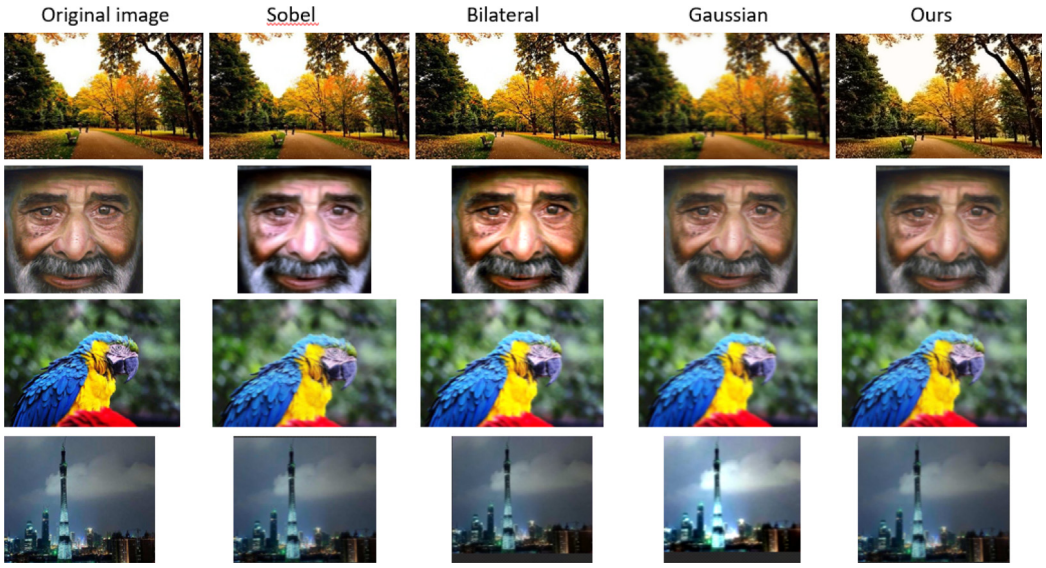

Fig. 4. The comparative edge smoothing by using different algorithms.

4. Experimental results and analysis

This sub-section provides the experimental outcomes of the proposed method utilizing semantically aware image-edge smoothing. Then we provide a brief explanation about the experimental platform and experimental data set. The outcomes of the edge smoothing by employing the proposed method and the traditional methods are presented.

In our research, all experiments were carried out using a Dell workstation equipped with an Intel E5-2860 CPU, dual NVidia K3100 GPUs, 16 GB RAM, and a 1TB SSD. All baseline-clustering algorithms were implemented utilizing the MATLAB 2011a package and C# platforms.

4.1. The data set of the AVA visual aesthetic

To our best knowledge, the AVA [2,27] is known as the largest data set to assess the image aesthetics, which is suitable for the performance assessment of image-edge smoothing. The entire AVA data set composed of over 250,000 images joined with a rich number of multi-level semantic tags containing a rich variety of aesthetic scores for each image, semantic labels for over 60 categories as well as labels pertinent to photographic style.

4.2. The comparative study

In this sub-section, we reported the edge smoothing generated by different algorithms. The comparisons with three well-known algorithms are made, which are called the Sobel, Bilateral, and Gaussian operator, respectively. In Fig. 4, we presented some examples of the smoothed images generated by different algorithms. To compare the generated outcomes, a user study is conducted.

An online user study based on the https://www.wjx.cn/ *is conducted*. We invited over 100 students majoring in Computer Science, who were graduate students experienced in photography

Table 2
The comparative outcomes of the smoothed edges produced by different methods.

	Sobel	Bilateral	Gaussian	Ours
First row	8.763%	11.435%	10.324%	69.478%
Second row	12.211%	13.325%	9.547%	64.917%
Third row	10.321%	11.440%	13.226%	65.013%
Last row	11.324%	14.546%	13.335%	60.795%

and could easily identify the visual attractiveness of each image. As presented in Table 2, the vote percentages of the users providing the most aesthetically pleasing picture were reported. As seen, the proposed method outperformed significantly than did the competitors. The outcomes indicated the advantage of the proposed method over the others.

5. Conclusions

This research suggested a novel adaptive image edge-smoothing algorithm that utilized a deep semantic model encoding the distribution of image patches into a GMM framework to predict the degree of edge smoothing. We employed an active learning algorithm to select a few highly representative image patches within each image on which the deep CNN was employed to deeply represent the visual content of each image patch. The comparative study between the proposed method and a set of traditional non-adaptive image-edge smoothing algorithms have shown the better performance of the proposed method. In the future, we plan to design a more generic and semantically aware image patch-encoding model to represent the massive-scale image patches.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the Shaoyang Science and Technology Bureau Plan Project on Data Desensitization System of Information System (No. 2019GZ55).

References

[1] Junwei Han, Xiang Ji, Xintao Hu, Dajiang Zhu, Kaiming Li, Xi Jiang, Guangbin Cui, Lei Guo, Tianming Liu, Representing and retrieving video shots in human-centric brain imaging space, IEEE Trans. Image Process. 22 (7) (2013) 2723–2736.
[2] Tuo Zhang, Lei Guo, Kaiming Li, Changfeng Jing, Yan Yin, Dajiang Zhu, Guangbin Cui, Lingjiang Li, Tianming Liu, Predicting functional cortical ROIs via DTI-derived fiber shape models, Cerebral Cortex 22 (4) (2012) 854–864.
[3] Dingwen Zhang, Deyu Meng, Junwei Han, Co-saliency detection via a self-paced multiple-instance learning framework, IEEE Trans. Pattern Anal. Mach. Intell. 39 (5) (2017) 865–878.
[4] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, Xuelong Li, Detection of co-salient objects by looking deep and wide, Int. J. Comput. Vis. 120 (2) (2016) 215–232.
[5] Navneet Dalal, Bill Triggs, Histograms of oriented gradients for human detection, in: International Conference on Computer Vision & Pattern Recognition (CVPR'05). Vol. 1, IEEE Computer Society, 2005.
[6] David G. Lowe, Object recognition from local scale-invariant features, ICCV 99 (2) (1999).
[7] Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[8] Shuiwang Ji, Wei Xu, Ming Yang, Kai Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2012) 221–231.
[9] Quoc V. Le, Will Zou, Serena Yeung, Andrew Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, 2011.
[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[11] Karen Simonyan, Andrew Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014.
[12] Jeffrey L. Elman, Finding structure in time, Cogn. Sci. 14 (2) (1990) 179–211.
[13] Yoshua Bengio, Patrice Simard, Paolo Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5 (2) (1994) 157–166.
[14] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[15] Mike Schuster, Kuldip K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.
[16] Alex Graves, Jürgen Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5–6) (2005) 602–610.
[17] Michael Collins, Yoram Singer, Unsupervised models for named entity classification, in: EMNLP, 1999.
[18] Sally A. Goldman, Yan Zhou, Enhancing supervised learning with unlabeled data, in: ICML, 2000, pp. 327–334.
[19] Zhi-Hua Zhou, Ming Li, Semi-supervised regression with co-training, in: IJCAI, 2005, pp. 908–916.
[20] Zhi-Hua Zhou, Ming Li, Tri training: Exploiting unlabeled data using three classifiers, IEEE Trans. Knowl. Data Eng. 17 (11) (2005) 1529–1541.
[21] Ulf Brefeld, Tobias Scheffer, Co-EM support vector learning, in: ICML, 2004.
[22] Zhe Wang, Songcan Chen, Multi-view kernel machine on single-view data, Neurocomputing 72 (10–12) (2009) 2444–2449.
[23] Zenglin Xu, Rong Jin, Haiqin Yang, Irwin King, Michael R. Lyu, Simple and efficient multiple kernel learning by group lasso, in: ICML, 2010, pp. 1175–1182.
[24] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, D. Xu, Generalized latent multi-view subspace clustering, IEEE Trans. Pattern Anal. Mach. (2018).
[25] Tian Xia, Dacheng Tao, Tao Mei, Yongdong Zhang, Multiview spectral embedding, IEEE Trans. Syst. Man Cybern. B 40 (6) (2010) 1438–1446.
[26] Naila Murray, Luca Marchesotti, Florent Perronnin, AVA: A large-scale database for aesthetic visual analysis, in: CVPR, 2012, pp. 2408–2415.
[27] Luming Zhang, Yi Yang, Yue Gao, Yi Yu, Changbo Wang, Xuelong Li, A probabilistic associative model for segmenting weakly supervised images, IEEE Trans. Image Process. 23 (9) (2014) 4150–4159.