Intelligent Automation And Soft Computing, 2020
Vol. 26, no. 4, 841–846
DOI: 10.32604/iasc.2020.010118

The Instance-Aware Automatic Image Colorization Based on Deep Convolutional Neural Network

Hui Li[1], Wei Zeng[1], Guorong Xiao[2] and Huabin Wang[1]

[1]School of Information Science and Technology, Huizhou University, Huizhou 516000, Guangdong, China
[2]Key Laboratory of Science & Technology and Finance, Guangdong University of Finance, Guangzhou 510521, Guangdong, China

ABSTRACT
Recent progress on image colorization is substantial and benefiting mostly from the great development of the deep convolutional neural networks. However, one type of object can be colored by different kinds of colors. Due to the uncertain relationship between the object and color, the deep neural network is unstable and difficult to converge during the training process. In order to solve this problem, this paper proposes an instance-aware automatic image colorization algorithm, which uses the semantic features of the object instance as prior knowledge to guide the deep neural network to do the colorization task. Meanwhile, we design a discrete loss function to train the deep network and this network can be trained from end to end. Experiments show that this algorithm can obtain satisfactory colorful results on the images containing object instance and achieves state-of-the-art results.

KEY WORDS: Image colorization, instance aware, deep convolutional neural network, semantic capture.

1 INTRODUCTION

THE image colorization technique assigns the vector RGB to each pixel based on its gray levels. This technique has a wide application prospect in the areas of historical photo processing, video processing and artwork recovery. Currently, there are three different types of image colorization algorithms; graffiti-based, transformed-based and automated-based. The graffiti-based and transformed-based methods require human involvement.

The early image colorization methods such as the graffiti-based and transformed-based obtain more propagation by solving the gradient differences between the pixel and its neighboring pixels together with border information or by utilizing the texture similarity. These kinds of methods usually need the user to decide the colors on some areas. The use of prior knowledge of the chosen colors largely depends on the input from the user, and a relatively good result can be achieved by repeat testing. For the transformed-based method, image colorization is done through the mapping between the image features such as the lightness, texture, etc. and the gray level of the image and it needs the user to provide reference images similar to the input image, and maps the color of the reference image to the gray images been inputted. For the automated-based colorization method, Deshpande et al. (2015) used many manually designed features and solved the linear system for the images. However, these manually designed features lack of the ability of the deep semantic information representation of the images.

With the great success of the deep convolutional neural network on the object's recognition and detection, its ability of representing the deep sematic information of the images has attracted wide attention, which would be helpful for the area of image colorization. Since the image deep semantic features provides additional information related to color, such as the grass is green, the sky is blue, etc. Recent research has focused on feature learning via data estimation and prediction. The recent image colorization algorithms all used the deep convolutional neural network to extract the semantic features of the gray images and did the colorization with the semantic information, the reference images. These algorithms show good results for the image colorization. For example, Cheng et al. (2015) presented an automated-based image colorization algorithm with high-level features extracted from the convolutional neural network, while improving the

CONTACT Wei Zeng ✉weizeng@hzu.edu.cn, Guorong Xiao ✉newducky@126.com

colorization results together with the both-side threshold. Larsson et al. (2016) proposed the deep network to extract the detailed features of the low layer and the semantic features of the high layer to achieve the image automatic colorization. Iizuka et al. (2016) utilized the convolutional neural network to capture the global and local features of the images, and the image colorization results were satisfactorily, especially outdoor photos. The generated color images make most people feel that the color images were real. Cristina caridade, et al. (2015) presented an automatic analysis method for microarray image to correct it for grid rotation, and to identify and evaluate the visible markers.

However, it is very easy to get confused when the current image colorization algorithms based on the deep convolutional neural network are applied to photos with object instances (such as human, cat, dog, car, etc.), as shown in Figure 1. It is because there are many instances of the same object on the photo but the color for the different instances of the same object may have difference colors, (for example, the color of the bag may be red or purple), which results in non-unique results provided by the deep convolutional neural network. Therefore, it is difficult for the network to get converged when it is trained with certain color objects.

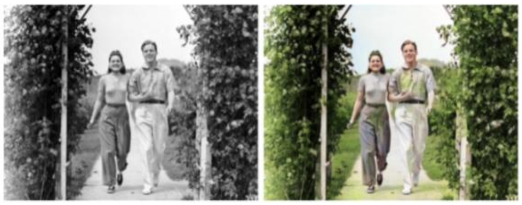

Figure 1. Problems Exist at the Object Instance Colorization.

In order to solve the problem of the object instance colorization, this paper proposes an instance-aware automatic image colorization algorithm based on the deep convolutional neuron network. This algorithm uses the instance feature of images as prior knowledge and designs two pipelines of the convolutional neuron network to extract the object instance semantic features and colorize the whole image automatically. One of the pipelines uses an auxiliary classifier to detect the category of each image to help the network learn the instance semantic information of the image. On the other pipeline, the instance level features and the image itself are combined, and the combined semantic features are used to colorize the gray image, which achieves the best effect of image colorization including the instances on the image.

This paper evaluates and tests the instance-aware automatic coloring algorithm based on the deep convolution neural network on the Pascal VOC data set. The experimental results show that the algorithm achieves a good coloring effect on most gray images with different instances and a state-of-the-art performance. At the same time, the instance-aware automatic coloring algorithm based on the deep convolution neural network achieves a real-time processing speed.

2 THE NETWORK FOR INSTANCE-AWARE AUTOMATIC IMAGE COLORIZATION

IN order to combine individual instances and the advanced semantic information of the full image to achieve the automated-colorization, this paper designs the network for the instance-aware automatic image colorization as shown in Figure 2.

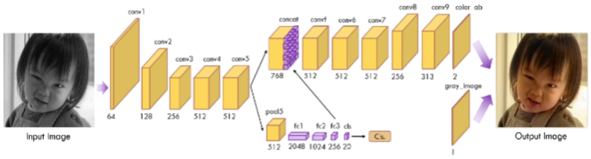

Figure 2. The Network for Instance-aware Automatic Image Colorization.

The network uses gray images as the input, and the output will be the colored images, shown in Figure 2. This network first extracts the image features by applying a serial of convolutional operations, and the corresponding image features are generated after each set of operations. Some of the convolutional operations take the stride of 2, which decrease the resolution of the feature image by 50%, while increasing the number of feature images, and the results are almost the same size of the computational space for each set of feature images.

Then, the network is divided into two paths after the five sets of convolutional operations (conv5) to arrive at the feature image. On one path, how to extract the instance semantic information is considered as a classification problem. The semantic tags of the images (human, dog, airplane, etc.) are used as the supervision information for the image to guide the training of the network, and the deep instance semantic information of the images can be obtained at the deep network. On the other path, the color images corresponding to the input images are used as the supervision information, and the instance semantic information obtained above is added to the fifth set of the convolutional operation to gain each position of the feature images to allow each of the pixels of the feature images to have the instance feature perception of the images. Then, five sets of the full convolutional operations are performed to obtain the image color information. Among the five sets of the convolutional operations, the feature channel of the conv9 is designed to be 313 and this number is decided by the number of bins of the color channel. Finally, the original gray image is combined with the color channels generated from the deep convolutional network to get the final color image. This network achieves end-to-end training.

2.1 Obtaining the Instance Semantic Features based on Category Labels

In order to extract the instance level features of the image, this paper uses category labels to be the supervision information, which enables the network to have the instance perception ability and to gain the semantic features of the image. Since the category labels are the semantic labels of the instance in the image, the network learns the instance semantic information under the same category with the detection of different images through the training of the deep network.

As shown in Figure 2, the network has been divided into two paths after conv5. For the bottom path of the network, we adopt the down sample operation to decrease the length and the width of the feature image shared by both paths by 50%, and then conduct through three sets of the full connected layer for further processing, and finally we use the classifier to do the category detection for the image. Here, this branch uses the following cross-entropy loss function to train:

$$L\ _{ls}(t_i, z_i) = -\sum_{j=1} t_{i,j} log z_{i,j} \qquad (1)$$

where the trained label of the image is represented as $\mathrm{t_i}$， and $\mathrm{t_i} \in \{0,1\}$ indicates the $\mathrm{t_i}$is the one-hot vector, and the category of ith image is represented as y_i, then $\mathrm{t_{i,y_i}} = 1$， $\mathrm{t^T_{i,j} t_{i,j}} = 1$. Furthermore, $\mathrm{z_{i,j}}$ represents the probability of ith image that belongs to class j. The total value of the loss for such functions is the sum of the probabilities of such image that belongs to all different classes. However, since $\mathrm{t_{i,j}}$is a one-hot vector, the loss function is:

$$L\ _{ls}(t_i, z_i) = -logP(t = t_i | x_i) \qquad (2)$$

where P represents a conditional probability and the loss function is equal to the negative logarithm of the training set.

Feature vector fc3 obtained from this path contains the instance semantic information on the image, i.e. it gets the instance perception on the image. In Iizuka et al. (2016), the image semantic information extracted from adding the category labels as the extra supervision information is called the global feature, and that is the instance semantic feature in this paper. Since there are enough big number of objects in the data set Pascal VOC, hence all the global features can be represented as instance level features. This paper duplicates the feature vector fc3 many times to adapt the feature image size on the conv5 and concatenate the feature vector with each position of the conv5 feature image, as shown in Figure 2.

After the instance features of the image has been added to the network, the network is able to get the perception of the instance semantic information, and such that the generated color for the same instance will be consistent.

2.2 The Loss Function on Training

As mentioned above, since the object can have different colors (such as a bag can be blue or can be red), hence it will cause the network training to be unstable if we would use the color as the supervision information to train the neural network. For the same object, if at times we let the network consider it to be red, and other times we let the network consider it to be blue, this will cause the network unable to converge or generate a gray photo.

Therefore, we discretize the color channels into a few intervals to allow the network to learn the relationship among the intervals instead of the fixed color values. In order to better utilize the existing gray information of the image, we transform the RGB color space into the Lab color space. The gray information L is already known, and what the network must learn are the a and b channels. We divide the color channels a and b into two intervals with a range of 10, and the total number of the relationships of all the intervals of channels a and b is 313, which is the feature image number of conv9 on Figure 2.

For each pixel, if its color belongs to bin_i , then $\mathrm{P}(bin_i) = 1$ ， where $\mathrm{P}(bin_i) \in [0,1]$. This formula represents the probability that the pixel belongs to bin_i. Here, we treat the problem of which pixel belongs to which bin as a problem of the classification and can be solved by using the cross-entropy loss function. Hence the final loss function $L\ _{olor}$is defined as below:

$$L\ _{olor}(T, Z) = \\ -\sum_{h,w} w(Z_{h,w}) logP(Z_{h,w,i} = T_{h,w,i} | x_i) \qquad (3)$$

where T represents the image the true color and Z represents the network predicting color. The function has been changed by using the sum of the weighted cross-entropy of each pixel as a whole of the loss function value, where the weight is represented by w, and the meanings of other parts are the same as function (2). Weight w is used to balance the number of pixels that belong to different color bin intervals. When there are a large number of pixels that belong to bin_a and while there are only a few pixels that belong to bin_b, we will assign a smaller weight to bin_awhile assigning a bigger weight to bin_b , and that will guarantee a balanced sample of the network.

3 EXPERIMENT AND RESULTS ANALYSIS

3.1 Datasets and Technical Details

THE PASCAL VOC provides a good and full standard dataset for the image recognition and classification. This dataset contains a large amount of object instances with semantic labels, which can be used as the supervision information. This paper has conducted the assessment test based on this dataset, where the training set includes 10821 pieces of photos and the training set includes 500 pieces of photos.

During the training, we use the Adam method to train the network. If we would train the whole network from scratch, then the training time on a single TITAN X GPU would be about one month. In order to speed up the training, we used the weights trained by Zhang et al. (2016) and Russakovsky et al. (2015) to initialize the classification network and colorize the shallow layers shared by the network. The last convolutional layers (conv6-conv9) on Figure 2 also used the initial weights trained by Zhang et al. (2016), and the other layers as the Gaussian noise as initial weights. After the meaningful initial weights had been used, we decreased the training time from the original one month to one week.

3.2 The Visualization Results

The image colorization effects from the algorithm based on the deep learning is much better than the algorithm based on the manually designed features. This paper mainly compares the automatic colorization algorithms based on the deep learning in recent years, and the colorization effect comparison of the methods can be found in reference papers of Larsson et al. (2016) and Iizuka et al. (2016). Figure 3 shows the image colorization effect. We can see that the colors on the instances of the images are reasonable, and the confusion of the instances against the background is minimum.

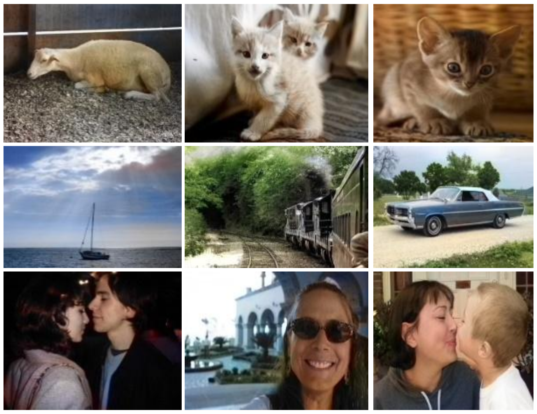

Figure 3. The Resulting Photos Colorized by the Instance-aware Automatic Colorization Algorithm.

Figure 4 shows the comparison results of our method against the results processed from the methods of Larsson et al. (2016) and Iizuka et al. (2016). From the figure, we can see that our algorithm has obvious advantages on the degree of naturalness and harmonization, and degree of richness of the different colors than other algorithms; especially the colorization effect is much better than the other two algorithms.

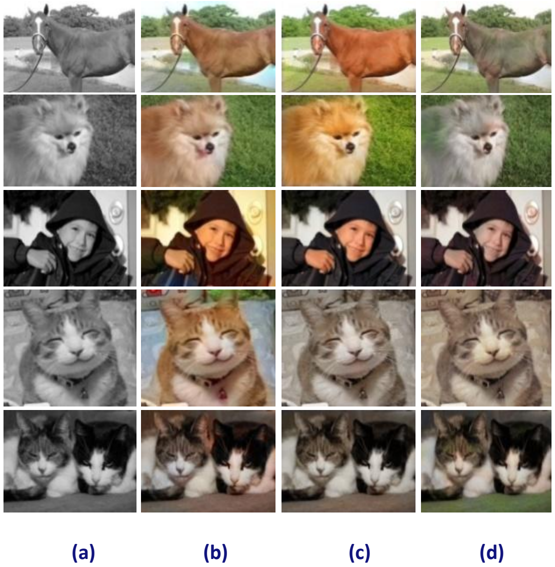

Figure 4. Experimental Results Comparison: (a) Input Images; (b) Our Output Images; (c) Larsson's Method; (d) Iizuka's Method.

3.3 User Experience Studies

In order to further assess our algorithm with the two other algorithms based on deep learning, we have conducted the user experience studies and there were 60 people that participated in the assessment test. We used 500 photos from the VOC dataset as test samples and compared three automated-based methods including our algorithm and two other deep-neural-network-based automatic colorization algorithms, i.e. the Larsson's method and the Iizuka's method. We asked users to give scores ranging from 1 to 10 based on the degree of the color naturalness (DCN), the photo color and harmonization (PCH), and the richness of the different colors (RDC). The higher the scores, the better effects of the results. For the degree of the color naturalness, users used only their intuition and would not spend too much time looking at the image details to check if the image were natural. The image color and harmonization means that the users will check for the details to discriminate if it were reasonable or not for the image foreground and background. The richness of the different colors refers to the users' comprehensive assessments on the image colorization effects, i.e. the users needed to determine if it is true, or if it is reasonable.

Table 1. The Average Scores by the Users.

	DCN	PCH	RDC
Our algorithm	8.73	8.62	8.45
Larsson's method	7.98	7.52	7.85
Iizuka's method	6.95	6.82	6.72

Table 2. The Standard Deviation of the Scores.

	DCN	PCH	RDC
Our algorithm	1.46	1.42	1.25
Larsson's method	1.40	0.72	1.32
Iizuka's method	1.25	1.05	1.36

Table 3. The Statistical Analysis Results by using a Significance Test.

	DCN	PCH	RDC
Our algorithm	Yes	Yes	Yes
Larsson's method	Yes	Yes	Yes
Iizuka's method	Yes	Yes	Yes

Tables 1 and 2 show the comparison of the average scores of the users and the corresponding standard deviations among our algorithm and the two methods. Table 3 shows the comparison of the significance test statistical analysis among our algorithm and the two methods. We can see from Table 1 that all results from the three automatic colorization methods are acceptable, (all score more than 6.0). Our algorithm score is the highest, and better than the other two methods. From Table 1 we can observe that the score of our algorithm is significantly different from the scores of the other two methods in most cases, (the t test result is $p < 0.05$，where $p < 0.05$ means there is a significant difference between the scores). All the evidence indicates that the algorithm presented in this paper has better results on the image colorization effects including the degree of the color naturalness, the photo color and harmonization, and the richness of the different colors.

4 CONCLUSION

THIS paper has proposed an instance-aware automatic colorization algorithm based on the deep neural network. It has added the image level of the classification network to help the network learn the image semantics, provide the colorization network with the instance level feature to achieve color consistencies. Furthermore, this paper has divided the color values into several intervals to prevent an unstable training problem because the same object on the training images might be represented as different colors. The experiments indicate that the algorithm presented in this paper achieved good results of the automatic colorization on gray images including object instances.

5 ACKNOWLEDGMENT

THIS work was supported by the Indigenous Innovation's Capability Development Program of the Huizhou University (Grant No. hzu201810), Huizhou Municipal Science and Technology Project Fund (Grant No. 2017C0405021), Educational Science Planning Project of the Guangdong Province and "Research on the innovative talent fostering pattern with deep converging of school and industries under new engineering disciplines " (Grant No. 2018GXJK183), Guangdong Provincial Key Laboratory of Technology and Finance & Big Data Analysis (Grant No. 2017B030301010), Construction on New technology Credit Service Platform based on the O2O Model (Grant No. 2017B080802004), Guangdong Key Research based of Technology and Finance (Grant No. 2014B030303005), Platform of Credit Financing and Trade for Guangdong Technological Enterprises (Grant No. 2014B080807035).

6 REFERENCES

Agrawal, P. and J. Carreira, J, (2015) Malik, Learning to see by moving. *In: Proceedings of the IEEE International Conference on Computer Vision*,37–45.

Caridade, C.M.R., André R.S. Marcal, P. Albuquerque, M.V. Mendes, F. Tavares (2015). Automatic analysis of dot blot images, Intelligent Automation and Soft Computing, Vol. 21, No. 4, 607-622.

Charpiat, G., M. Hofmann, and B. Schölkopf, (2008). Automatic Image Colorization via Multimodal Predictions. *Lecture Notes in Computer Science*, 126-139.

Cheng, Z., Q. Yang, and B. Sheng, (2015). Deep colorization. *Proceedings of the IEEE International Conference on Computer Vision*,415–423.

Deshpande, A., J. Rock, and D. Forsyth, (2015). Learning Large-Scale Automatic Image Colorization. *IEEE International Conference on Computer Vision. IEEE Computer Society*, 567-575.

Doersch, C., A. Gupta, A.A. Efros, (2015). Unsupervised visual representation learning by context prediction. *In: Proceedings of the IEEE International Conference on Computer Vision*, 1422–1430.

Donahue, I., P. Kra¨henbu¨hl, T. Darrell, (2016). Adversarial feature learning. *Ar Xiv preprint arXiv:*1605.09782.

Everingham, M., L. V. Gool, C. K.I. Williams, J. Winn, A. Zisserman (2010).The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2), 303-338.

Huang, Y. C., Y. S. Tung, J. C. Chen, S. W. Wang, and J. L. Wu, (2005). An adaptive edge detection-based colorization algorithm and its applications. *ACM International Conference on Multimedia, Singapore, November. DBLP*, 351-354.

Iizuka, S., E. Simoserra, and H. Ishikawa, (2016). Let there be color: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *Acm Transactions on Graphics*, 35(4), 1-11.

Irony, R., D Cohen-Or, and D. Lischinski, (2005). Colorization by example. *Eurographics Symposium on Rendering Techniques, Konstanz, Germany, June 29 - July. DBLP*, 201-210.

Jayaraman, D. and K. Grauman, (2015). Learning image representations tied to ego-motion. *In: Proceedings of the IEEE International Conference on Computer Vision*, 1413–1421.

Kingma, D. P. and J. Ba, (2014). Adam: A Method for Stochastic Optimization. *Computer Science.*

Larsson, G., M. Maire, and G. Shakhnarovich, (2016). Learning Representations for Automatic Colorization, 577-593.

Levin, A., D. Lischinski, and Y. Weiss, (2004). Colorization using optimization. *Acm Transactions on Graphics*, 23(3), 689-694.

Lotter, W., G. Kreiman, and D. Cox, (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104.*

Luan, Q., W. Fang, D. Cohen-Or, L Lin, Y. Q. Xu, and H. Y. Shum, (2007). Natural image colorization. *Eurographics Conference on Rendering Techniques,* 309-320.

Owens, A., J. Wu, J.H McDermott, W.T. Freeman, and A. Torralba, (2016). Ambient sound provides supervision for visual learning. *In: ECCV.*

Owens, A., P. Isola, J. McDermott, A. Torralba, E.H. Adelson, and W.T. Freeman, (2016). Visually indicated sounds. *CVPR.*

Pathak, D., P. Kra¨henbu¨hl, J. Donahue, T. Darrell, and A. Efros, (2016). Context encoders: Feature learning by inpainting. *In: CVPR.*

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, H. Zhiheng, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision,* 115(3), 211-252.

Wang, X, and A. Gupta, (2015). Unsupervised learning of visual representations using videos. *In: Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

Welsh, T., M. Ashikhmin, and K. Mueller, (2002). Transferring color to greyscale images. *ACM Transactions on Graphics,* 21(3), 277-280.

Zhang, R., P. Isola, and A. A. Efros, (2016). Colorful Image Colorization. *Ar Xiv preprintarXiv:* 1603.08511.

7 DISCLOSURE STATEMENT

NO potential conflict of interest was reported by the authors.

8 AUTHOR BIOGRAPHIES

Hui Li received M.S. degree from the National Aviation University in Ukraine in 2003. She is currently a lecturer at the School of Information Science and Technology, Huizhou University. Her research interests include image processing, computer vision and image recognition.

Wei Zeng received M.S. degree in Computer Science from the University of Ottawa, Ontario province, Canada in 1995. He has taught at several universities in China and has been a lecturer at the Huizhou University, P. R. China since 2017. His current research interests include patter recognition, sparse representation, image recognition and neural network.

Guorong Xiao received B.S. degree in 2001, M.S. degree in 2004 from the South China University of Technology in Computer Sciences and Technology. He is an associate professor of Computer Application Technology. Currently he serves as the deputy director of the Guangdong Provincial Key Laboratory of Technology and Finance & Big Data Analysis, and the chief technical officer of the Guangdong Province Science & Technology Financial Comprehensive Information Service Platform, and the director of the Guangdong Provincial Science and Technology Finance Big Data Engineering Technological Research Centre, and a member of the Internet Professional Committee of the China Computer Federation. His main research is big data analysis, intelligence information processing and computer application. He is co-corresponding author.

Huabin Wang received M.S degree in Software Engineering from the Central South University, Changsha, PR China, in 2006. Since 2013, he has been an associate professor at the Huizhou University, PR China. His current research interests include image recognition, intelligent algorithms and wireless sensor network.