Journal of Ambient Intelligence and Humanized Computing (2020) 11:3101–3108
https://doi.org/10.1007/s12652-017-0656-4

ORIGINAL RESEARCH

# Learning multi-path CNN for mural deterioration detection

Rui Huang[1,2,3] · Wei Feng[1,2] · Mingyuan Fan[1,2] · Qing Guo[1,2] · Jizhou Sun[1,2]

Received: 27 March 2017 / Accepted: 9 December 2017/ Published online: 19 December 2017

**Abstract**

Mural deterioration easily destroys valuable paintings and must be monitored frequently for preventive protection. Deterioration detection in mural images is often manually labeled and is a preprocessing step for mural protection and restoration. Many deterioration forms are commonly invisible with only one lighting condition because mural deterioration is caused by changes in the material and plaster layer. This study addresses mural deterioration detection through a multi-path convolutional neural network (CNN), which takes images of a scene with multiple lightings as inputs and generates a binary map that indicates deterioration regions. We design an eight-path CNN in which seven paths are utilized for basic feature extraction from lighted images, and the remaining path is responsible for cross feature fusion. This mechanism enables our method to not only identify suitable features for different lightings but also utilize these features collaboratively through cross feature fusion. Furthermore, we build two realistic mural deterioration datasets of real-world mural deterioration and briquettes that simulate the cave deterioration. Extensive experiments verify the effectiveness and efficiency of our method.

**Keywords** Mural deterioration detection · Multi-input, max-fusion, multi-path CNN · Dense CRF

## 1 Introduction

Murals, which date back to 3000 B.C., have been widely utilized to record histories and myths in different cultures. However, murals easily suffer from different types of deterioration, which can make these valuable works unrecoverable. Thus, recent studies propose preventive protection methods (Pinchin 2013; Wirilander 2012) to monitor murals by manually marking deterioration regions in these images. Such methods are generally laborious, especially for large-scale and frequency monitoring tasks. Thus, developing an automatic method to detect mural deterioration is necessary.

Given that deterioration is commonly caused by material and plaster layer changes, which manifests as subtle changes in the 3D structure of the wall surface, a single-view scene under single lighting cannot show all deterioration forms. The deterioration forms (marked by yellow and red ellipses) shown in Fig. 1 are invisible in *Input* $\mathrm{x}_{\mathrm{DL}}$ and *Input* $\mathrm{x}_{\mathrm{DSL}_4}$. However, they can be easily observed in images taken under other lighting conditions. Thus, traditional single image parsing and semantic labeling methods (Tighe and Lazebnik 2010, 2013; Zhang et al. 2013; Farabet et al. 2013) fail to detect deterioration due to the lack of 3D information. As shown in Fig. 1, traditional image semantic segmentation method (e.g., BA_CRF) fails to segment deterioration with a single-view image. Depth, hyper-spectral and multi-view cameras can be employed to capture 3D structure of deterioration. However, subtle structure changes that correspond to the deterioration impose high demands on the precision of depth-captured devices or multi-view geometry algorithms, thereby resulting in an expensive deterioration detection solution based on device and algorithm.

We propose a solution apart from calculating the 3D structure from the observation that deterioration can be more visible under different illumination conditions than under single illumination condition by fixing the position of a digital camera. We show a set of images captured with different lighting directions in Fig. 1. When an image is taken under environment lighting conditions (EL denoted by a solid circle), we can  barely recognize the deterioration in the image.

✉ Wei Feng
wfeng@tju.edu.cn

[1] School of Computer Science and Technology, Tianjin University, Tianjin 300350, China

[2] Key Research Center for Surface Monitoring and Analysis of Cultural Relics, State Administration of Cultural Heritage, Beijing, China

[3] School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China

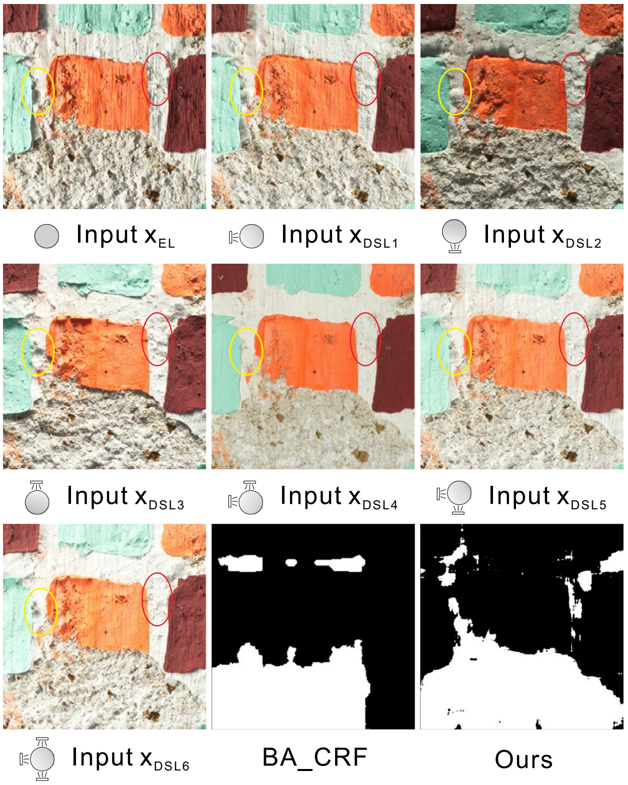


**Fig. 1** Motivation of our MPCN for mural deterioration detection. We show a deterioration detection case of $D_{lab}$. The images are taken for a mural briquette under lightings. Note that subscript EL denotes environment lighting. Subscripts DSL1–DSL6 denote directional side lightings. The yellow and red ellipses mark the deterioration regions that have different visibility in the images that taken with different lightings. We also show the deterioration detection results of our MPCN and a baseline method (e.g., BA_CRF)

However, the deterioration becomes highly visible under directional side lightings (DSLs). Thus, the deterioration within the images under different DSLs appears differently. We label each pixel by considering the images with EL and DSLs simultaneously. Notably multiple lighted images are also utilized for fine-grained change detection in Feng et al. (2015) and Huang et al. (2017b) because of the sufficient details of the EL and DSL images.

Accordingly, we propose to learn a multi-path convolutional neural network (MPCN) which takes seven images of a mural scene as inputs that contain one EL and six DSLs. It then outputs a binary map that indicates the deterioration regions. Apart from utilizing a fixed CNN to extract the deep features of all images, we also design a seven-path CNN in which each path extracts deep features of specific lighted images as shown in Fig. 2. Unlike the Siamese architecture in which two-path CNN shares parameters, each path in the proposed network possesses its own parameters. This design can fully utilize different lighted images. We add a side-path that conducts max-pooling operation across multiple deep features to combine the seven deep features from different lighted images of different layers. We then build two datasets to verify the deterioration detection ability of our model. The first dataset, which is denoted as $\mathbf{D}_{\mathbf{lab}}$, contains briquette images under seven directional lightings. The deterioration in the briquettes is generated in the laboratory through manual control. The second dataset, which is denoted as $\mathbf{D}_{\mathbf{cav}}$, contains real mural images of the Dunhuang grotto under seven directional lightings. Our major contributions are threefold:

- We address a novel problem of automatically detecting mural deterioration from a single scene with multiple lightings.
- We propose to learn a multi-path CNN to solve the aforementioned problem by learning deep features for specific lighted images and fusing them with a max pooling operation. Our method is simple and easy to implement.
- We capture two types of multiple lighted images from the laboratory and real grottoes. We then build two corresponding datasets for evaluation of mural deterioration detection and for public study.

## 2 Related work

Most works related to deterioration detection involve image parsing, which categorizes pixels into several different classes. Image parsing is content-dependent method. Previous works primarily adopt hand-crafted features and simple classifiers to classify each pixel (or superpixel) according to their corresponding labels (Feng and Liu 2008; Feng et al. 2010; Tighe and Lazebnik 2010; Zhang et al. 2013). For example, Tighe and Lazebnik (2010) proposes to use shape, location, texture/sift, color and appearance, totally 1708D superpixel features for local superpixel labeling. Zhang et al. (2013) train a Random Forest classifier (Breiman 2001) with Texton and SIFT features for unary potential in their CRF model. He et al. (2004) use multiscale CRF with local classifier, regional features, and global features for image labeling. Liu and Lu (2007) train a Bayesian classifier with texture features to detect Lacunas for ancient paintings. Apart from using classifier, Liu et al. (2009) proposes to propagate the label from the best matched images by a coarse-to-fine SIFT flow algorithm. Whatever, designing representative features and training a suitable classifier in these works are essential tasks in correctly labeling the images. However, designing representative features to distinguish deteriorated and normal regions is difficult in certain tasks, such as deterioration detection.

The recent breakthrough work (Krizhevsky et al. 2012) and success of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Deng et al. (2009) have renewed

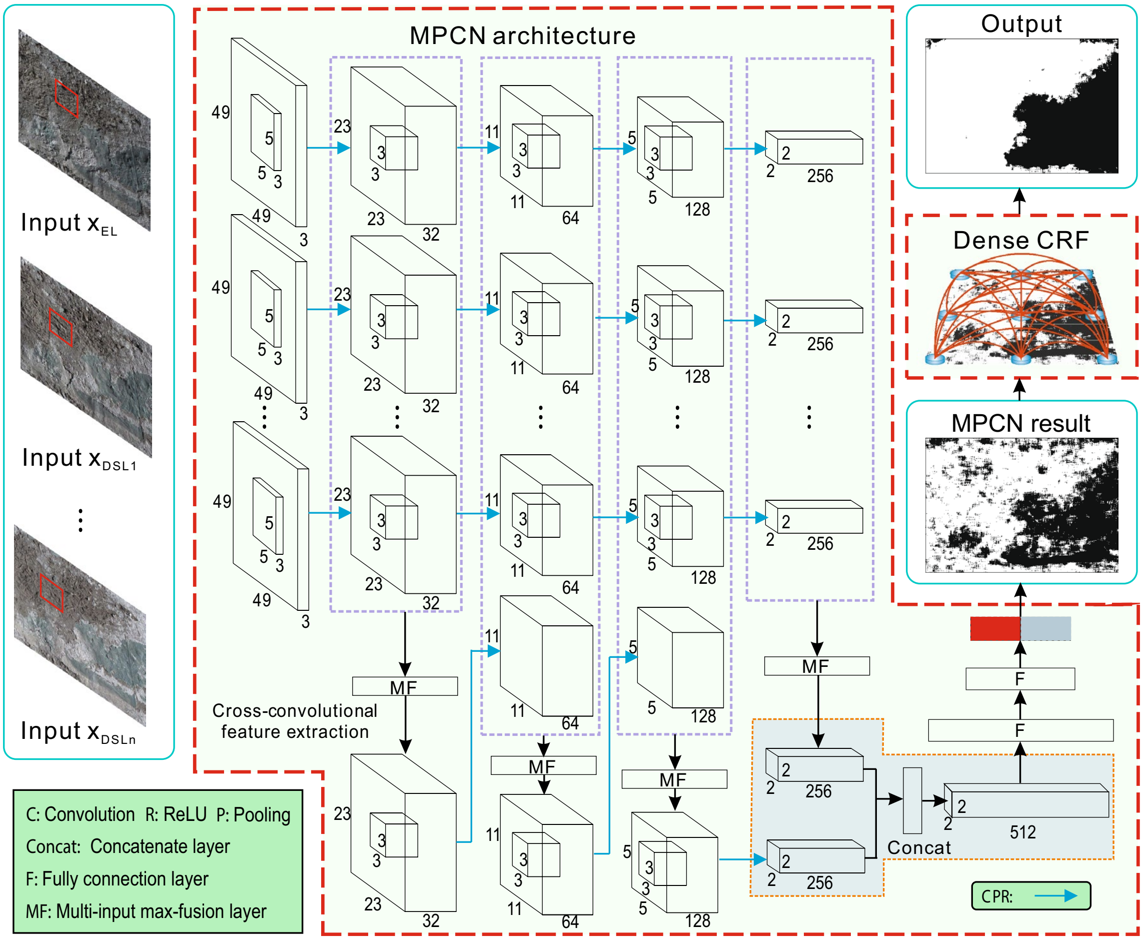


**Fig. 2** The architecture of the proposed MPCN. The inputs are multiple lighted images that taken at fixed camera position. Mainly, we have two types of path, i.e., the basic feature extraction paths and cross-convolutional feature extraction path. Basic feature extraction paths extract deep features for each lighted image, whereas the cross-convolution feature extraction path fuses the features of lighted images from different convolutional layers

the interest of researchers on CNN, especially in terms of computer visions. Most contemporary image parsing or semantic segmentation methods are based mainly on CNN to adaptively extract deep features and classify each pixel into semantic labels (Long et al. 2015; Chen et al. 2015; Hariharan et al. 2015, 2014; Mostajabi et al. 2015). Considering the pairwise relation of the pixel, Zheng et al. (2015) proposed to incorporate CRF in CNN for the end-to-end training of the semantic network by employing the mean-field approximation inference. However, another branch of semantic segmentation methods utilize CRF as postprocessing for fine-grained segmentation (Chen et al. 2015; Papandreou et al. 2015) because of the efficient inference of the dense CRF (Koltun 2011). Several recent works have designed two-tower CNN architectures that match two inputs, such as human parsing (Liu et al. 2015), optical flow estimation (Bai et al. 2016; Dosovitskiy et al. 2015) and video tracking (Guo et al. 2017a, b). However, no readily available method exists for mural deterioration detection with multiple lighted images. The proposed method employs multi-path CNN architectures in which each path of the CNN is responsible for feature extraction from one lighted image. We also design a cross-convolutional feature extraction path to fuse multi-level convolutional features. The final decision is made by the softmax classifier with collaborative features. Our network architectural may facilitate multi-input applications, such as image co-saliency detection (Huang et al. 2017a) and tone-mapping (Feng et al. 2016).

## 3 The method

Given a group of multiple lighted mural images $X = [x_{EL}, x_{DSL_1}, \ldots, x_{DSL_n}]$ as shown in Fig. 2, we propose a multi-path CNN (MPCN) framework to solve the mural deterioration detection problem. An image with subscript EL is taken under environment lighting, and another image with subscript $DSL_i$ is taken with the $i$th directional side lighting.

The proposed MPCN takes square image patches that are cropped from all lighted images with same center position and outputs a probability of the center pixel that belongs to deterioration. We employ the sliding window approach to compute the probability map of the entire image. After the probability map is obtained, we employ dense CRF (Koltun 2011) to obtain the final mural deterioration detection result.

### 3.1 Deterioration detection with MPCN

*Architectures of MPCN* MPCN consists of eight convolutional feature extraction paths. Seven of these paths are responsible for basic feature extraction from lighted images and one path is responsible for convolutional feature fusion. The first path extracts features of the EL image, $\mathrm{x}_{\mathrm{EL}}$. The second to seventh paths are designed to extract features of the six DSL images. We design a cross-convolutional feature extraction path to fuse and fully utilize the features of lighted images of the different convolutional layers. ReLU, max-pooling and dropout are always conducted after convolution or fully connected layers. Thus, we do not partition them as single layers. We adopt four convolutional layers by using max-pooling and ReLU for basic feature extraction paths. The filters are set to $5 \times 5 \times 3 \times 32$, $3 \times 3 \times 32 \times 64$, $3 \times 3 \times 64 \times 128$ and $3 \times 3 \times 128 \times 256$.[1] We use $2 \times 2$ max-pooling with a stride of 2. The cross-convolutional feature extraction path consists of three convolutional layers with filter's sizes of $3 \times 3 \times 32 \times 64$, $3 \times 3 \times 64 \times 128$ and $3 \times 3 \times 128 \times 256$. As shown in Fig. 2, each convolutional layer of the cross convolutional feature extraction path takes the output features of the multi-input max-fusion layer (*MF*) as input and then outputs convolutional feature for the subsequent *MF*. *MF* conducts multi-input max-fusion operations to fuse the effective convolutional features of the same layer but with different paths. The third convolutional layer of the cross-convolutional feature extraction path outputs $2 \times 2 \times 256$ convolutional features, which are concatenated with the output features of *MF* on the basic feature extraction paths, and forms $2 \times 2 \times 512$ features. We utilize three fully connected layers with lengths of 1024, 512 and 2. The first two fully connected layers are followed with the ReLU and dropout, whereas the third layer is followed with softmax operation to generate the corresponding deterioration probability. We adopt a dropout ratio of 0.5 in our training.

*Multi-input max-fusion* The basic convolutional feature extraction path and the cross-convolutional feature extraction path differ mainly in that we utilize max-pooling operation to fuse the input features in the cross-convolutional feature extraction path. Formally, we let $\mathrm{F}_{\mathrm{ij}}^{\mathrm{x}_{\mathrm{L}_\mathrm{n}}}$ represents the *j*th convolutional feature map at the *i*th layer for the *n*th lighted input patch. $n = 0$ denotes the input image with EL, and $n = 1, \ldots, 6$ denotes the inputs are images with DSLs. We calculate the output of the multi-input max-fusion layer, i.e., $\mathrm{F}_{\mathrm{ij}}^{\mathrm{MF}}$ by

$$\mathrm{F}_{\mathrm{ij}}^{\mathrm{MF}}(\mathrm{p}) = \begin{cases} \max(\mathrm{F}_{\mathrm{ij}}^{\mathrm{x}_{\mathrm{L}_0}}(\mathrm{p}), \ldots, \mathrm{F}_{\mathrm{ij}}^{\mathrm{x}_{\mathrm{L}_6}}(\mathrm{p})), & i = 1, 4 \\ \max(\mathrm{F}_{\mathrm{ij}}^{\mathrm{x}_{\mathrm{L}_0}}(\mathrm{p}), \ldots, \mathrm{F}_{\mathrm{ij}}^{\mathrm{x}_{\mathrm{L}_6}}(\mathrm{p}), \Psi(\mathrm{F}_{\mathrm{i-1j}}^{\mathrm{c}})(\mathrm{p})), & i = 2, 3, \end{cases} \tag{1}$$

where $p$ is pixel position and $\Psi(\mathrm{F}_{\mathrm{i\,-\,1j}}^{\mathrm{MF}})$ denotes the convolution, max-pooling and ReLU operations on $i - 1$th multi-input max-fusion layer. The cross-feature extraction starts after the first convolutional layers of the basic feature paths finish the computations.

*Training procedure* We sample the patches for multiple lighted images at the same pixel position to train the MPCN. We crop patches from the pure clean images and pure deteriorated images for $\mathbf{D_{lab}}$. The patches with the positive (or negative) labeled pixels exceed 80% of the total pixel number in the patch are selected for $\mathbf{D_{cav}}$. Notably, the center pixel of the patch must have the same label as the patch label. The patch size of the two datasets is set to $49 \times 49$. We utilize a $1 : 1$ ratio of the positive and negative samples to balance the two data classes. Finally, we collect 5.4*M* and 4.96*M* for simulated and grotto data, respectively. We utilize the 90% of the samples as training sample set while the remaining 10% are employed to form the validation sample set. We build our mural deterioration framework on basis of the deep learning framework CAFFE (Jia et al. 2014). We adopt multinomial logistic loss for training, utilize a learning rate of 0.0001, and decrease it by a factor of 0.1 when the loss is stabilized. Our batch size, weight decay and momentum are set to 50, 0.0005 and 0.99, respectively. The parameters are updated by stochastic gradient descent. One week is needed for MPCN to converge on each dataset. Thus, training two MPCNs for $\mathbf{D_{lab}}$ and $\mathbf{D_{cav}}$ takes about two weeks.

### 3.2 Post processing with dense CRF

MPCN outputs the deterioration probability of a pixel sitting at the center of the mural image patch, which only considers the local image context to endow the deterioration probability. In this paper, we employ dense CRF (Koltun 2011) to infer the final label of each pixel by considering the entire image context. CRF generally optimizes the following energy function:

$$\mathrm{E}(l) = \sum_{p} \phi_p(l_p) + \sum_{p,q} \psi_{p,q}(l_p, l_q), \tag{2}$$

[1] We denote the filter's size by $H \times W \times \#Input \times \#Output$, where $H$ denotes filter's height, $W$ denotes filter's width, $\#Input$ denotes input dimension and $\#Output$ denotes output dimension.

where $l$ is the label assignment for the pixel, $\phi_p(l_p)$ is the unary potential, $\psi_{p,q}(l_p, l_q)$, as a pairwise potential, and $p$ and $q$ are the pixel positions.

*Unary potential* We set the unary potential $\phi_p(l_p) = -\log \mathrm{P(p)}$, where P(p) is the output deterioration probability of pixel $p$ by MPCN.

*Pairwise potential* The pairwise potential $\psi_{p,q}(l_p, l_q)$ smoothens the noise labels by jointly considering the relation of the location and the appearance of neighbor pixels and is defined as follows:

$$\psi_{pq}(l_p, l_q) = \mu(l_p, l_q)k(p, q), \tag{3}$$

where $\mu(l_p, l_q) = [l_p \neq l_q]$ is the indication function and kernel $k\ (p, q)$ is defined as follows:

$$k(p, q) = \omega_1 \exp\left(-\frac{\|p-q\|^2}{2\sigma_\alpha^2} - \frac{\|I_p - I_q\|^2}{2\sigma_\beta^2}\right) + \omega_2 \exp\left(-\frac{\|p-q\|^2}{2\sigma_\gamma^2}\right), \tag{4}$$

where $\sigma_\alpha$ is 20, $\sigma_\beta$ is 3, $\sigma_\gamma$ is 3, $\omega_1$ is 5 and $\omega_2$ is 3. The first term encourages the nearby pixels with similar features to possess same label. The second term removes the isolated pixels.

## 4 Data collection

We build two real-world datasets of multiple lighted images for mural deterioration detection, namely, $\mathbf{D_{lab}}$ and $\mathbf{D_{cav}}$. $\mathbf{D_{lab}}$ is captured in the laboratory and attempt to observe the process of deterioration generation, whereas $\mathbf{D_{cav}}$ is captured in real caves. The images in both datasets are captured with a Canon 5D Mark III single-lens reflex camera. The standard prime of 50 mm is utilized. We open the camera and turn off the flashlight when capturing the images. Additional area light with low color temperature is employed to provide uniform light for shooting. We adopt drive-by-wire technology and send the shooting order by computer to avoid shaking caused by pushing the shutter button. All image groups are labeled by our staff members who are experienced in labeling disruption deterioration. Each image group is labeled by considering all directional images and is double checked.

*Laboratory data* $\mathbf{D_{lab}}$ We create several briquettes to simulate deterioration and design a shooting platform for the DSLs. The briquettes are placed in the middle of the platform while taking an image. Four incandescent lamps are then turned on or off to provide seven different illumination conditions. Given that the briquettes are made circular, we crop the images to obtain square images and thus produce different image sizes in this dataset. We collect 89 groups of disruption images. We utilize 50 groups as the training image set and 39 groups as the testing image set. In the training image set, we collect 9 groups of pure clean images (i.e., deterioration free) and 41 groups of pure deteriorated images (i.e., pixels belonging to disruption).

*Cave data* $\mathbf{D_{cav}}$ We collect the second deterioration data from real caves with murals. Given that all captured murals are vertical to the ground, we use a tripod to set up our camera. The shooting area is parallel to the camera lens. Further more, we can achieve a dark environment without natural light because all the grottoes are concealed by doors. We provide area light in seven directions. One setting is vertical to the murals, whereas the other lights are near the murals by approximately 30 degrees at clockwise intervals of 1, 3, 5, 7, 9, 11. We capture nine groups of disruption images with a resolution of 2736 × 2192 under seven lighting directions. We divide each image into small images to generate training and testing set. Finally, we obtain 85 groups of small images for training and 24 groups of small images for testing.

## 5 Experiments

### 5.1 Setup

*Baseline* We employ two conventional methods as our baselines. The first baseline is a simple classifier [i.e., SVM (Chang and Lin 2014)] with 76 dimensional handcrafted features, including RGB, LAB, HSV, HOG (Felzenszwalb et al. 2010) and Gabor (Feichtinger and Strohmer 1998) features. We randomly sample 49 × 49 patches from the training images and extract the mean features of the patches. We randomly sample 0.7$M$ training examples from 350 mixed lighted images for simulate dataset. We then randomly sample 0.7$M$ training examples from 595 mixed lighted images for the grotto dataset. In our experiment, we keep the ratio of the positive and negative at 1. We call this baseline BA_SVM.

We learn a one-versus-all boosted decision tree classifier for positive and negative labels for the second baseline. A total of 669 dimensional hand-crafted features are utilized in this experiment, which includes a bank of 17 filtered features, RGB color, dense HOG features (Felzenszwalb et al. 2010), and LBP-like features. We then calibrate the output of the boosted decision trees through a multiclass regression classifier. The output of the regression classifier forms a unary potential in the CRF model. The pairwise term encodes a contrast-dependent smoothness prior to the image labeling. The weight term is learned by direct searching. In particular, several parameter values are tested, and the one that generates the best results on a subset of training images is retained. We employ the public implementation of Darwin framework (He et al. 2004;

**Table 1** Quantitative comparisons of deterioration detection of different methods on test images of $\mathbf{D_{lab}}$

| Methods | Precision | Recall | F-measure | MAE |
|---|---|---|---|---|
| BA_SVM | 0.6725 | **0.9263** | 0.6881 | 0.2464 |
| BA_CRF | 0.7702 | 0.5468 | 0.7451 | 0.1979 |
| Ours | **0.8825** | 0.8516 | **0.8823** | **0.0956** |

**Table 2** Quantitative comparisons of deterioration detection of different methods on test images of $\mathbf{D_{cav}}$

| Methods | Precision | Recall | F-measure | MAE |
|---|---|---|---|---|
| BA_SVM | 0.6675 | **0.9335** | 0.6835 | 0.3318 |
| BA_CRF | 0.7782 | 0.7567 | 0.7764 | 0.2249 |
| Ours | **0.8055** | 0.8832 | **0.8114** | **0.1244** |

Gould et al. 2009; Shotton et al. 2006), train it with our training sets, and report the results on our test set. We call the second baseline BA_CRF.

We average all lighted results and utilize a threshold of 0.5 to obtain the final result for BA_SVM and BA_CRF, as they generate results for every lighted image.

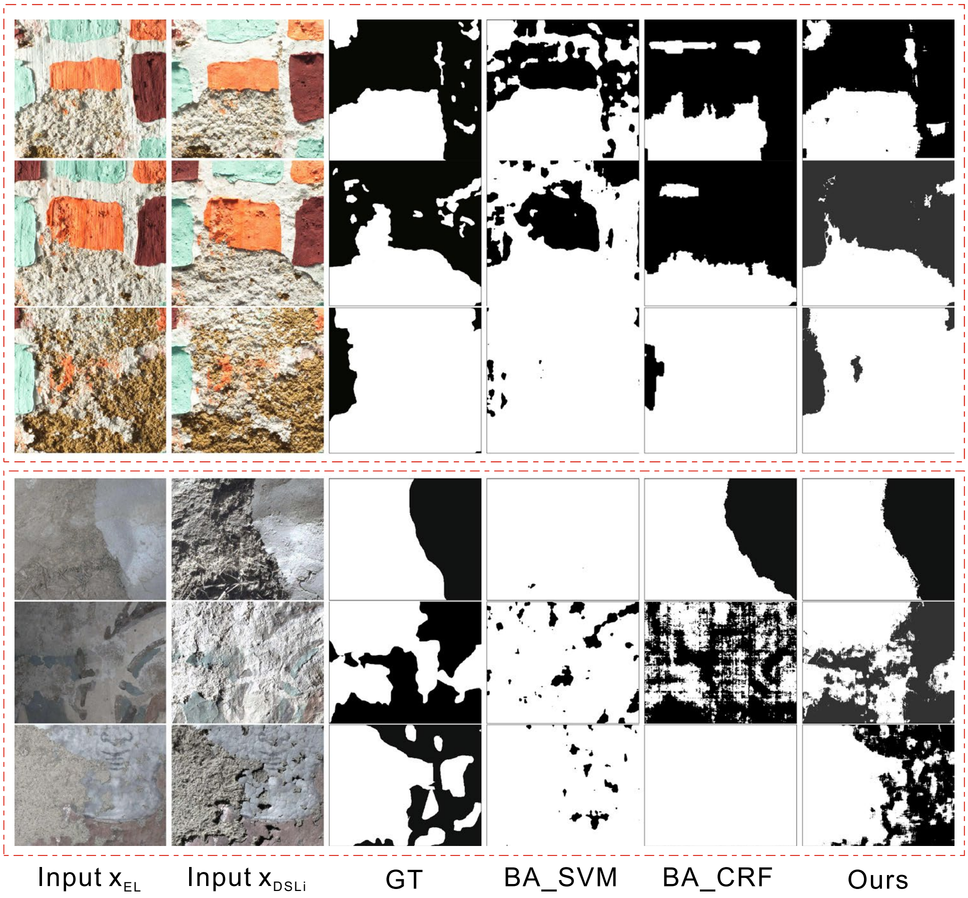
*Criteria* We employ mean precision, recall, F-measure and mean absolute error (MAE) to evaluate our deterioration detection results. $\beta = 0.3$ in F-measure emphasizes the precision. All criteria are averaged on all test images.

## 5.2 Result analysis

This section details the quantitative comparisons of the different methods on the test images of $\mathbf{D_{lab}}$ and $\mathbf{D_{cav}}$ in Tables 1 and 2, respectively. Several typical mural deterioration detection cases are shown in Fig. 3.

As shown in Fig. 3, the briquettes are simply painted with four colors (i.e., white, orange yellow, dull-red and light blue) that are basically utilized in real grottoes. Nearly no complicated painting and texture exist on the briquettes. Thus, the deterioration detection on $\mathbf{D_{lab}}$ easier than that on $\mathbf{D_{cav}}$. Unlike $\mathbf{D_{lab}}$ images, the grotto images of $\mathbf{D_{cav}}$ possess complex painting contents and textures. Certain parts of the painting walls are also repaired by concrete, and several parts have lost the entire patches of painting. We show three mural deterioration detection cases with increasing degradation degrees on $\mathbf{D_{lab}}$ in the top pink box in Fig. 3. BA_SVM fails to detect the deterioration forms in these

**Fig. 3** Comparisons of different deterioration detections methods on $\mathbf{D_{lab}}$ and $\mathbf{D_{cav}}$ datasets, respectively. We show three typical deterioration detection cases of each dataset. From top to bottom the degradation level of deterioration is increasing



relatively simple cases. BA_CRF achieves better detection results compared with BA_SVM, but BA_CRF still fails in several complex regions. By contrast, our method attains the best deterioration detection results, which are near the ground truth. The bottom pink box of Fig. 3 shows three difficult cases on $\mathbf{D}_{\mathbf{cav}}$. The conventional methods BA_SVM and BA_CRF, fail to implement true deterioration detection in these cases. However, our method can still obtain good mural deterioration results.

In Table 1, we show the quantitative comparisons of different methods on $\mathbf{D}_{\mathbf{lab}}$. From Table 1, we find that a simple classifier with hand-crafted features performs poorly in mural deterioration detection. The F-measure of BA_SVM can reach only 0.6881, which is the lowest score. The MAE of BA_SVM is 0.2464, which is the highest score. The F-measure of mural deterioration detection can improve by 0.057 whereas MAE can decrease by 0.0485 by utilizing a complicated classifier, such as boosted decision tree classifier with elegant features (i.e., BA_CRF). However, the proposed method obtains greater improvement than the baselines. We obtain 28.2% increase in F-measure (our F-measure is **0.8823**), and 61.2% decrease in MAE (our MAE is **0.0956**) compared with BA_SVM.

Table 2 shows quantitative comparisons between the baselines and our method on the $\mathbf{D}_{\mathbf{cav}}$ test images captured in real grottoes. From Table 2, we find that the performance of all methods is lightly low, which demonstrates that mural deterioration detection is difficult in real-world scenes. BA_SVM still performs worse than the other methods on $\mathbf{D}_{\mathbf{cav}}$ test images with F-measure of 0.6835 and MAE of 0.3318. BA_CRF achieves improvement over BA_SVM on F-measure by 0.929, and decreases MAE by 0.1069. Compared with BA_CRF, our method achieves 4.5% improvement relative to F-measure (our F-measure is **0.8114**) and decreases MAE (our MAE is **0.1224**) by 44.7%.

As shown in Tables 1 and 2, BA_SVM always presents the highest Recalls and the lowest Precisions on both $\mathbf{D}_{\mathbf{lab}}$ and $\mathbf{D}_{\mathbf{cav}}$. From Fig. 3, we find that BA_SVM nearly fails on all deterioration detection cases, even for the simplest case (the first case) on $\mathbf{D}_{\mathbf{lab}}$. BA_SVM is prone to classify the mural into deterioration more than to normal, which caused the highest recall on two datasets. This finding demonstrates that simple classifier and general hand-crafted features cannot distinguish the normal murals from deteriorated ones.

The running times for a 912 × 730 image of our method, BA_SVM and BA_CRF are 61, 39 and 196 m, respectively. BA_CRF exhibits the longest running time, most of which is spent on unary potential. The main reason for the long running time of our method and BA_SVM is per-pixel classification. Although our method takes about 1h to generate final output, it presents the highest detection precision. In real applications, mural deterioration cares more on detection precision rather than on running time.

The above-mentioned analysis shows that our method is highly qualified for mural deterioration detection in images captured under multiple lighting conditions.

## 6 Conclusion

This study presents an MPCN for mural deterioration detection, which can be utilized to automatically detect deterioration in murals for preventive protection. MPCN takes multiple lighted image patches as inputs and outputs the corresponding label of the center pixel. We design a cross-convolutional feature extraction path to fuse the convolutional features of different lighted images from different layers. We then employ multi-input max-fusion operation in this path to integrate the convolutional features from the different paths. We build two real-world datasets to validate our method. The proposed method shows better performance than the two baselines on both real-world datasets. We plan to add image groups with a wide range of disease types in our future work. Thus, we can train a large CNN with a higher capacity for mural deterioration detection. The running time of our method can also be decreased for fast mural deterioration detection.

**Acknowledgements** This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61671325 and 61572354, and in part by the National Science and Technology Support Project 2013BAK01B01.

## References

Bai M, Luo W, Kundu K, Urtasun R (2016) Exploiting semantic information and deep matching for optical flow. In: European conference on computer vision. https://doi.org/10.1007/978-3-319-46466-4_10

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Chang C-C, Lin C-J (2014) Libsvm : a library for support vector machines. ACM Trans Intell Syst and Technol 2(3):1–27. https://doi.org/10.1145/1961189.1961199

Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2015) Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: International conference on learning representations. arXiv:1412.7062

Deng J, Dong W, Socher R, Li LJ, Li K, Li FF (2009) Imagenet: A large-scale hierarchical image database. In: Computer vision and pattern recognition, pp 248–255. https://doi.org/10.1109/CVPRW.2009.5206848

Dosovitskiy A, Fischery P, Ilg E, Hausser P, Hazirbas C, Golkov V, Smagt PVD, Cremers D, Brox T (2015) Flownet: Learning optical flow with convolutional networks. In: International conference on computer vision. https://doi.org/10.1109/ICCV.2015.316

Farabet C, Couprie C, Najman L, LeCun Y (2013) Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell 35(8):1915–1929. https://doi.org/10.1109/TPAMI.2012.231

Feichtinger HG, Strohmer T (1998) Gabor analysis and algorithms: theory and applications. Birkhauser, Boston

Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645. https://doi.org/10.1109/TPAMI.2009.167

Feng W, Liu ZQ (2008) Region-level image authentication using bayesian structural content abstraction. IEEE Trans Image Process 17(12):2413–2424. https://doi.org/10.1109/TIP.2008.2006435

Feng W, Jia J, Liu ZQ (2010) Self-validated labeling of markov random fields for image segmentation. IEEE Trans Pattern Anal Mach Intell 32(10):1871–1887. https://doi.org/10.1109/TPAMI.2010.24

Feng W, Tian FP, Zhang Q, Zhang N, Wan L, Sun J (2015) Fine-grained change detection of misaligned scenes with varied illuminations. In: International conference on computer vision. https://doi.org/10.1109/ICCV.2015.149

Feng W, Yang Y, Liang W, Yu C (2016) Tone-mapped mean-shift based environment map sampling. IEEE Trans Vis Comput Graph 22(9):1077–2626. https://doi.org/10.1109/TVCG.2015.2500236

Gould S, Fulton R, Koller D (2009) Decomposing a scene into geometric and semantically consistent regions. In: International conference on computer vision. https://doi.org/10.1109/ICCV.2009.5459211

Guo Q, Feng W, Zhou C, Huang R, Wan L, Wang S (2017a) Learning dynamic siamese network for visual object tracking. In: International conference on computer vision

Guo Q, Feng W, Zhou C, Pun CM, Wu B (2017) Structure-regularized compressive tracking with online data-driven sampling. IEEE Trans Image Process 26(12):5692–5705. https://doi.org/10.1109/TIP.2017.2745205

Hariharan B, Arbeláez P, Girshick R, Malik J (2014) Simultaneous detection and segmentation. In: European conference on computer vision. https://doi.org/10.1007/978-3-319-10584-0_20

Hariharan B, Arbeláez P, Girshick R, Malik J (2015) Hypercolumns for object segmentation and fine-grained localization. In: Computer vision and pattern recognition. https://doi.org/10.1109/CVPR.2015.7298642

He X, Zemel RS, Carreira-Perpiñán MÁ (2004) Multiscale conditional random fields for image labeling. In: Computer vision and pattern recognition

Huang R, Feng W, Sun J (2017) Color feature reinforcement for co-saliency detection without single saliency residuals. IEEE Signal Process Lett 24(5):569–573. https://doi.org/10.1109/LSP.2017.2681687

Huang R, Feng W, Wang Z, Fan M, Wan L, Sun J (2017b) Learning to detect fine-grained change under variant imaging conditions. In: International conference on computer vision workshop on E-Heritage

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: ACM MM. https://doi.org/10.1145/2647868.2654889

Koltun V (2011) Efficient inference in fully connected crfs with gaussian edge potentials. In: International conference on neural information processing systems, pp 109–117

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: International conference on neural information processing systems. https://doi.org/10.1145/3065386

Liu C, Yuen J, Torralba A (2009) Nonparametric scene parsing: label transfer via dense scene alignment. In: Computer vision and pattern recognition. https://doi.org/10.1145/3065386

Liu J, Lu D (2007) Knowledge based lacunas detection and segmentation for ancient paintings. In: International conference on virtual systems and multimedia, pp 121–131

Liu S, Liang X, Liu L, Shen X, Yang J, Xu C, Lin L, Cao X, Yan S (2015) Matching-cnn meets knn: Quasi-parametric human parsing. In: Computer vision and pattern recognition. https://doi.org/10.1109/CVPR.2015.7298748

Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Computer vision and pattern recognition. https://doi.org/10.1109/CVPR.2015.7298965

Mostajabi M, Yadollahpour P, Shakhnarovich G (2015) Feedforward semantic segmentation with zoom-out features. In: Computer vision and pattern recognition. https://doi.org/10.1109/CVPR.2015.7298959

Papandreou G, Chen LC, Murphy K, Yuille AL (2015) Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: International conference on computer vision. https://doi.org/10.1109/ICCV.2015.203

Pinchin S (2013) Historical perspectives on preventive conservation. J Archit Conserv 4119(1):1–3

Shotton J, Winn J, Rother C, Criminisi A (2006) Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: European conference on computer vision. https://doi.org/10.1007/11744023_1

Tighe J, Lazebnik S (2010) Superparsing: scalable nonparametric image parsing with superpixels. In: European conference on computer vision. https://doi.org/10.1007/978-3-642-15555-0_26

Tighe J, Lazebnik S (2013) Finding things: image parsing with regions and per-exemplar detectors. In: Computer vision and pattern recognition. https://doi.org/10.1109/CVPR.2013.386

Wirilander H (2012) Preventive conservation: a key method to ensure cultural heritages authenticity and integrity in preservation process. E-Conserv Mag 6(24):164–176

Zhang H, Wang J, Tan P, Wang J, Quan L (2013) Learning crfs for image parsing with adaptive subgradient descent. In: International conference on computer vision. https://doi.org/10.1109/ICCV.2013.382

Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, Huang C, Torr PH (2015) Conditional random fields as recurrent neural networks. In: International conference on computer vision. https://doi.org/10.1109/ICCV.2015.179