IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 37, No. 5, October, 2020, pp. 847-854
Journal homepage: http://iieta.org/journals/ts

Retrieval of Multiple Spatiotemporally Correlated Images on Tourist Attractions Based on Image Processing

Shuang Lu[1], Qian Zhang[1], Yi Liu[1*], Lei Liu[1], Qing Zhu[2], Ke Jing[1]

[1] School of Art & Design, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[2] BIM Center, Henan Wujian Construction Co. LTD, Zhengzhou 450000, China

Corresponding Author Email: ylz_ly@yeah.net

https://doi.org/10.18280/ts.370518

Received: 1 June 2020
Accepted: 8 September 2020

Keywords:
image processing, tourist attractions, multiple spatiotemporally correlated images (MSCIs), image retrieval

ABSTRACT

The thriving of information technology (IT) has elevated the demand for intelligent query and retrieval of information about the tourist attractions of interest, which are the bases for preparing convenient and personalized itineraries. To realize accurate and rapid query of tourist attraction information (not limited to text information), this paper proposes a spatiotemporal feature extraction method and a ranking and retrieval method for multiple spatiotemporally correlated images (MSCIs) on tourist attractions based on deeply recursive convolutional network (DRCN). Firstly, the authors introduced the acquisition process of candidate spatiotemporally correlated images on tourist attractions, including both coarse screening and fine screening. Next, the workflow of spatiotemporal feature extraction from tourist attraction images was explained, as well as he proposed convolutional long short-term memory (ConvLSTM) algorithm. After that, the ranking model of MSCIs was constructed and derived. Experimental results demonstrate that our strategy is effective in the retrieval of tourist attraction images. The research results shed light on the fast and accurate retrieval of other types of images.

1. INTRODUCTION

The thriving of information technology (IT) has elevated the demand and expectation of tourists for high-quality tourism services. To make the itinerary more convenient and personalized, tourists need to query and retrieve information about the tourist attractions of interest [1, 2].

In the era of big data, it is an irresistible trend to retrieve various information about tourist attractions rapidly and accurately, in addition to text information [3-5]. Fortunately, the research results on content-based image retrieval of tourist attractions have greatly promoted the intelligence and informatization in the tourism industry [6-9].

With the help of advanced modern technologies (e.g. computer vision, intelligent video analysis, and digital image retrieval), it is very innovative to develop image retrieval methods for tourist attractions based on image processing. Quite many tourist attraction images are large in data volume and difficult to express. For these images, the retrieval algorithms generally cover two steps: image description by feature descriptor, and matching of salient image features.

The existing studies mainly aim to improve the description ability of feature descriptor and the matching efficiency of salient features [10-12]. Beaudoin, [13] performed template matching between the target images and a landmark image database, and obtained a descriptor for long-distance visual features of images similar to the templates. To improve the image retrieval effect, Houaria and Zaoui, [14] described the images in a landmark image database with discriminative visual phrases, selected the relatively important descriptors, and converted the target images into intuitive histograms through soft coding. Ordonez et al. [15] described the images in a landmark image database with three-dimensional (3D) visual phrases, and sorted a series of spatiotemporally correlated images, such that the spatiotemporal feature points of the images could be extracted more accurately.

The above methods face such disadvantages as complex models and low retrieval efficiency. Therefore, many scholars have turned their attention to the dimensionality reduction of the massive data of image processing, the simplification of the retrieval operations, and the storage of the data processed in related operations [16-19]. Kim et al. [20] described the images in a landmark image database with high-dimensional visual words, which greatly promotes the feature matching between the target images and the templates. Considering visual content and background environment, Fukada et al. [21] completed the position perception and description of images with landmark discrimination descriptors. Latorre-Martiánez, et al. [22] generated image descriptor symbols that can iteratively optimize the segmentation of image descriptors and their geographic locations. After compressing landmark images, Yoshihara et al. [23] realized the image retrieval of a landmark image database through binary hash code mapping. Based on three-stage learning process, Kawase et al. [24] presented a multi-modal binary hash code mapping method for discrete images.

The above methods cannot iteratively optimize the hash map, unless the salient image features have been fully extracted. Compared with the hash mapping of end-to-end learning, these methods perform poorly in the generalization and migration on the acquired image features.

On the application of image retrieval for tourist attractions, An et al. [25] provided a relatively complete design for image retrieval system of tourist attractions, and developed an

Android-based system software, which supports various functions, namely, matching of similar tourist attractions, acquisition of the name and location of tourist attractions, and personalized tour guide. Their system enables tourists to acquire valuable information from the images on tourist attractions.

The existing content-based image retrieval methods are grounded on the low-level visual features of the images. These features are loosely correlated with the high-level semantics of the images. What is worse, the massive spatiotemporal information contained in the images on tourist attractions has not been fully utilized or fused. For accurate retrieval of tourist attraction images, this paper proposes a spatiotemporal feature extraction method and a ranking and retrieval method for multiple spatiotemporally correlated images (MSCIs) on tourist attractions based on deeply recursive convolutional network (DRCN).

The remainder of this paper is organized as follows: Section 2 introduces the acquisition process of candidate spatiotemporally correlated images on tourist attractions, which includes two steps: coarse screening and fine screening; Section 3 details the spatiotemporal feature extraction from tourist attraction images, and explains the proposed convolutional long short-term memory (ConvLSTM) algorithm; Section 4 constructs and derives the MSCIs ranking model; Section 5 verifies the effectiveness of our strategy in the retrieval of tourist attraction images through experiments; Section 6 puts forward the conclusions.

2. ACQUISITION OF SPATIOTEMPORALLY CORRELATED IMAGES

The spatiotemporally correlated images on tourist attractions were acquired through two steps: coarse screening, and fine screening.

2.1 Coarse screening

The pixel difference PD against the tourist attraction image library was adopted to determine whether a tourist attraction image is eligible for coarse screening:

$$PD = I_{i+1} - I_i \tag{1}$$

The PD value is negatively correlated with the similarity between images. The coarse screening was implemented in the following steps:

(1) The tourist attraction image, which had gone through wavelet analysis, was reconstructed by similarity matrix. The reconstructed image was binarized, using the mean pixel value in the non-zero value area as the threshold. The threshold can be calculated by:

$$T_{PM} = \sum_{a=0}^{W-1}\sum_{b=0}^{H-1} g(a,b) / [W \times H - N_B] \tag{2}$$

where, $g(a, b)$, W, and H are the gray value, width, and height of the input image, respectively; N_B is the number of elements in the set $UB=\{(a, b)|g(a, b)=0\}$ of pixels with mean gray value of zero and all background points. Through binarization, the mask image of the region of interest (ROI) can be obtained as:

$$MI(a,b) = \begin{cases} 1 & g(a,b) > T_{PM} \\ 0 & g(a,b) \le T_{PM} \end{cases} \tag{3}$$

(2) The mask image (3) and the input image were convoluted to obtain the ROI that characterizes the salient features of the input image. Then, the regional extraction results were sorted in top-down sequence of space, forming the transition set $\{T_1, T_2, T_3, \ldots, T_N\}$ for coarse screening. This step clearly lowers the computing load of the coarse screening of candidate spatiotemporally correlated images.

(3) T_1 was selected from $\{T_1, T_2, T_3, \ldots, T_N\}$ as the current target of coarse screening. After i had been increased by 1, the selected target was added to the set of coarse screening targets. Then, the PD was calculated by formula (2). If PD is smaller than T_{PM}, the image similarity is high, and the i+1-th image should be added to the set of coarse screening targets; if PD is greater than the threshold, the image similarity is low, and the i+1-th image should not be added.

(4) If all images in the library have been matched, the coarse screening should be terminated; otherwise, Steps 3-4 should be executed repeatedly until all images in the library have been matched.

Figure 1 explains the workflow of the coarse screening of spatiotemporally correlated images.

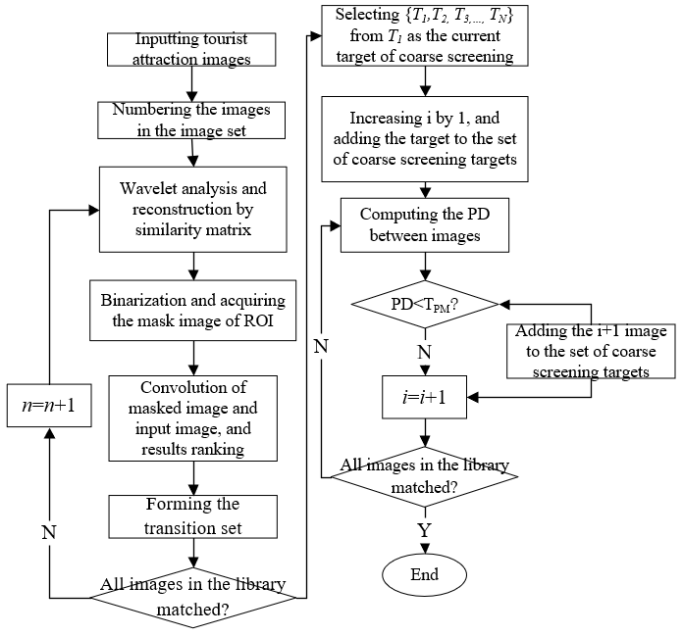

Figure 1. The workflow of the coarse screening of spatiotemporally correlated images

2.2 Fine screening

The above coarse screening produces the set of coarse screening targets. To improve the accuracy of the target images for fine screening, the set of coarse screening targets was measured by normalized correlation coefficient and mutual information between images. The normalized correlation coefficient can be calculated by:

$$NC(p_i, p_j) = \frac{\sum_{a=0}^{W-1}\sum_{b=0}^{H-1}[p_i(a,b) - p_i'][p_j(a,b) - p_j']}{\sqrt{\sum_{a=0}^{W-1}\sum_{b=0}^{H-1}[p_i(a,b) - p_i']^2 \sum_{a=0}^{W-1}\sum_{b=0}^{H-1}[p_j(a,b) - p_j']^2}} \tag{4}$$

where, $p_i(a, b)$ is the value of pixel p_i in the i-th image; $p_j(a, b)$ is the value of pixel p_j in the j-th image; p'_i and p'_j are the mean pixel values of the i-th image and the j-th image, respectively. If the normalized correlation coefficient is small, the image has a low spatiotemporal similarity with the images in the library, and should be removed from the set. The mutual information can be calculated by:

$$MIV(p_i, p_j) = \sum_{A\in p_i, B\in p_j} PD_{p_i p_j}(A,B)\, log\left[PD_{p_i p_j}(A,B) / PD_{p_i}(A) PD_{p_j}(B)\right] \quad (5)$$

where, $PDp_i(A)$ is the probability density of pixel p_i in the i-th image; $PDp_j(B)$ is the probability density of pixel p_j in the j-th image; $PDp_ip_j(A, B)$ is the joint probability density of pixels p_i and p_j. The mutual information characterizes the gray correlation between a target image and the images in the library. The smaller the mutual information, the greater the gray value difference between the image and the library.

3. SPATIOTEMPORAL FEATURE EXTRACTION

3.1 Extraction principle

As a typical tool for deep feature learning, ConvLSTM combines the merits of convolution and LSTM. In this paper, the spatiotemporal features of each target image are acquired by DRCN, a ConvLSTM-based high-resolution image reconstruction method. Figure 2 presents the structure of spatiotemporal feature extraction algorithm based on ConvLSTM and spatial attention.

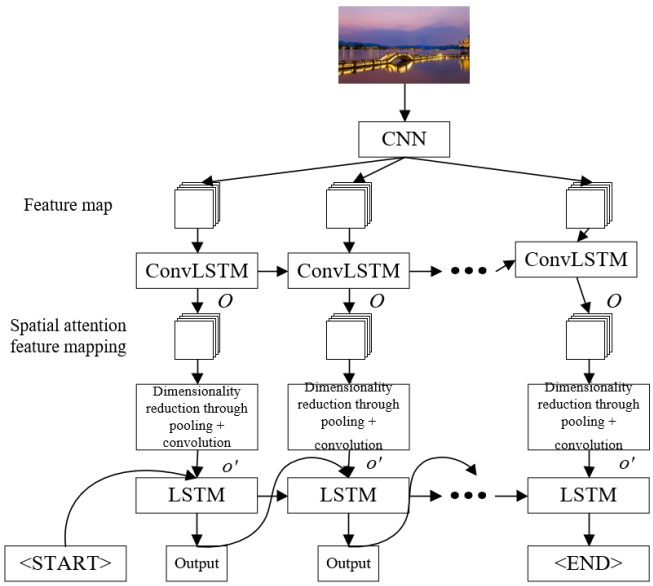

Figure 2. The extraction of spatiotemporal features based on ConvLSTM and spatial attention

Let TI be the given input image on tourist attraction, $DS=\{d_0, d_1, \ldots, d_{M-1}\}$ be the description sentence generated by the DRCN for the image, and $DS'=\{d'_0, d'_1, \ldots, d'_{M-1}\}$ be the reference sentence, where M is the length of the description sentence, and V is the volume of the image description vocabulary. If d_i, and d'_i both belong to $\mathbb{R}^V$, then the descriptive words of the image have been one-hot encoded.

Some spatiotemporal information might get lost, when the features are extracted based on the output of the fully-connected layer. To prevent the information loss, the spatiotemporal features were extracted from the target image on a relatively low convolution layer in the convolutional neural network (CNN):

$$IF = C(TI) \quad (6)$$

where, $C(*)$ is the equivalent function of the relatively low convolution layer in the CNN; $IF=\{IF_0, IF_1, \ldots, IF_{P-1}\}$ is the set of features of the target image, which consists of P $c\times c$ two-dimensional (2D) feature maps. Any feature IF_i belongs to $\mathbb{R}^{c\times c}$. To make the spatial information of the target image more stereo, the spatiotemporal features of the image were extracted once more by the ConvLSTM:

$$IF^* = CL(IF) \quad (7)$$

where, $CL(*)$ is the equivalent function of the ConvLSTM; $IF^*=\{IF^*_0, IF^*_1, \ldots, IF^*_{Q-1}\}$ is the set of secondarily extracted features of the target image, with Q being the number of filters. Let $k\times k$ be the size of convolution kernels. Then, any feature IF_i belongs to $\mathbb{R}^{k\times k}$, where $k=c$.

Capable of describing images containing time-varying information, the LSTM only supports inputs in the form of one-dimensional (1D) vectors. Let u be the number of hidden layer nodes in the LSTM. Since the hidden layer outputs IF^* and IF of ConvLSTM are both 3D tensors, a fully-connected layer was adopted to reduce the dimensionality of the outputs, turning Q into u.

Being a decoder, the LSTM requires its inputs to be the spatiotemporal image features at time t and the descriptive words of the image at time t-1. Note that the word is not a one-hot code d_i, but an embedded word mapped to E-dimensional embedded space:

$$e_{t+1} = U_p d_t, t \in \{0,\ldots,M-1\} \quad (8)$$

where, $U_p \in \mathbb{R}^{E\times V}$ is a parameter matrix updated constantly through training. The index of non-zero elements in the one-hot code corresponds to the column vector of e_t. The parameters of the expanded cell in the LSTM must be in a consistent state at any time. Thus, the output of the hidden layer can be treated as network output:

$$O_t = L(O_{t-1}, E_t, O'_t), t \in \{0,\ldots,M-1\} \quad (9)$$

where, $L(*)$ the equivalent function of the LSTM network; $U_t \in \mathbb{R}^E$ is the output of the function; O'_t is the spatiotemporal features of the target images obtained through a CNN extraction and a ConvLSTM extraction. Based on the features of the current time and the embedded words generated in the previous time, the output probability of a descriptive word can be calculated by:

$$EP_t \propto \exp(\delta(e_{t-1} + \delta_O O_t + \delta_{O'} O'_t)) \quad (10)$$

where, $\delta\in\mathbb{R}^{V\times E}$; $\delta_O\in\mathbb{R}^{E\times M}$; $\delta_{O'}\in\mathbb{R}^{E\times P}$. All these three parameters need to be trained. During the training of model parameters, the descriptive words generated at time t can be obtained, if

the target image TI and the reference sentence $y'_{1(t-1)}$ at time t-1 are given. The loss function of our model can be defined as:

$$loss(TI) = -\sum_{t=1}^{M} \log EP_t\left(O_t \middle| O^*_{1(t-1)}, TI\right) \tag{11}$$

3.2 ConvLSTM algorithm

Drawing on the abovementioned principle of feature extraction, the convolution structure of the ConvLSTM can fully memorize and learn the spatial information of the target image, and bridge it up with the description sentence in time, through the information transforms related to state. The hidden and storage units of the ConvLSTM can be updated by:

$$\begin{aligned} i_t &= \sigma\left(W_{iI} * I_t + W_{iO} * O_{t-1} + W_{iF} \circ F_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_{fI} * I_t + W_{fO} * O_{t-1} + W_{fF} \circ F_{t-1} + b_f\right) \\ F_t &= f_t \circ F_{t-1} + i_t \circ \tanh\left(W_{FI} * I_t + W_{FO} * O_{t-1} + b_F\right) \\ h_t &= \sigma\left(W_{hI} * I_t + W_{hH} * H_{t-1} + W_{hF} \odot F_t + b_h\right) \\ F_t &= h_t \circ \tanh\left(F_t\right) \end{aligned} \tag{12}$$

where, I_t, F_t, and O_t are the input gate, forget gate, and output gate of the ConvLSTM, respectively; * and º are convolution and the inner product operation, respectively. According to the previous analysis, the network outputs $O=\{O_0, O_1, \ldots, O_{Q-1}\}$, where $O_i \in \mathbb{R}^{k \times k}$.

To preserve as much spatiotemporal information of the target image as possible, the spatial attention method was adopted to assign a dynamic weight to each $O_i \in \mathbb{R}^{k \times k}$. If the weight is large, the image area corresponding to O_t could be described by the embedded words generated at that time. The dynamic weights of different image areas can be calculated by:

$$\mu_{ti} = W_\mu \tanh\left(W_O O_i + W_o o_{t-1}\right) \tag{13}$$

$$\omega_{ti} = \frac{\exp\left(\mu_{ti}\right)}{\sum_{l=1}^{c} \exp\left(\mu_{tl}\right)} \tag{14}$$

a) b) c)

Figure 3. The two spatiotemporal feature extractions

The spatial attention mapping tensor can be obtained through the weighted calculation of O_i and dynamic weights:

$$SA_t = \rho\left(\Psi_\varphi\left(o_{t-1}\right)\right) \sum_{i=1}^{E} \omega_i O_i \tag{15}$$

where, Ψ_φ means the hidden layer state of the network at time t-1 is inputted to the multilayer perceptron, reflecting the relative importance between sentence construction or spatiotemporal information preservation in the generation of image description sentences. Figures 3(b) and 3(c) show the two spatiotemporal feature extractions of the target image (a).

4. MSCIS RANKING AND RETRIEVAL

4.1 Ranking model

The tourist attraction image dataset can be independently represented by spatiotemporal features. But the spatial and temporal features differ clearly in form. Therefore, a heterogeneous graph was constructed for the dataset:

$$HG^{(IG)} = \left(V^{(IG)}, S^{(IG)}, \Omega^{(IG)}\right) \tag{16}$$

where, $IG \in \{Time, Space\}$ is the image set independently represented by spatiotemporal features; $V^{(IG)}$ is the set of vertices corresponding to every images; $S^{(IG)}$ is the set of edges between the vertices; $\Omega^{(IG)} \in \mathbb{R}^{W \times H}$ is the weight matrix of the edges, whose elements $\varepsilon^{(IG)}_{ij}$ correspond to every edges $S^{(IG)}_{ij}$.

Then, the similarity between the two feature maps HG^{Time} and HG^{Space} was measured by different methods. The vertices in temporal feature map HG^{Time} form a binary matrix about the presence/absence of temporal correlations between an image and its description sentences $IF^{Time}{}_i$ and $IF^{Time}{}_j$. The binary matrix can be defined by the Jaccard index:

$$\Omega^{Time}_{ij} = \frac{\left|NZ\left(IF_i^{Time}\right) \cap NZ\left(IF_j^{Time}\right)\right|}{\left|NZ\left(IF_i^{Time}\right) \cup NZ\left(IF_j^{Time}\right)\right|} \tag{17}$$

where, $NZ(*)$ is the set of nonzero elements in the real matrix; $|*|$ is the volume of the real matrix. The vertices in spatial feature mapHG^{Space} form a real matrix about the similarity between spatial features $IF^{Space}{}_i$ and $IF^{Space}{}_j$ in terms of saliency. The real matrix can be defined by the Gaussian kernel function:

$$\Omega^{Space}_{ij} = \exp\left(-\frac{DIS^2\left(IF_i^{Space}, IF_i^{Space}\right)}{\gamma^2}\right) \tag{18}$$

where, $DIS(*)$ is the distance between spatial features; γ is the local zoom parameter. Since the spatiotemporal features of tourist attraction images reflect the actual intention of the retriever, this paper treats the spatiotemporal features as two independent feature sets.

The MSCIs ranking model basically encompasses a loss function and a regularization term. Let $RS=[RS_1, RS_2, \ldots, RS_n]^T$ be the initial ranking scores of n candidate spatiotemporally correlated images; $RS^*=[RS^*_1, RS^*_2, \ldots, RS^*_n]^T$ be the optimal ranking scores of these images; $Label=[Label_1, Label_2, \ldots, Label_n]^T$ be the label indicating whether a target image has spatiotemporal correlation (if yes, $Label_i=1$; if no, $Label_i=0$). Then, the optimal ranking score vector RS^* of the candidate spatiotemporally correlated images can be calculated by:

$$RS^* = \arg\min_{RS} \begin{bmatrix} regular\left(RS, TI\right) \\ +\theta loss\left(RS, Label\right) \end{bmatrix} \tag{19}$$

where, $regular(*)$ is the regularization term ensuring that similar images have similar scores; $loss(*)$ is the loss function minimizing the difference between the initial and optimal rankings; θ is an adjustable positive parameter.

4.2 Derivation of the ranking model

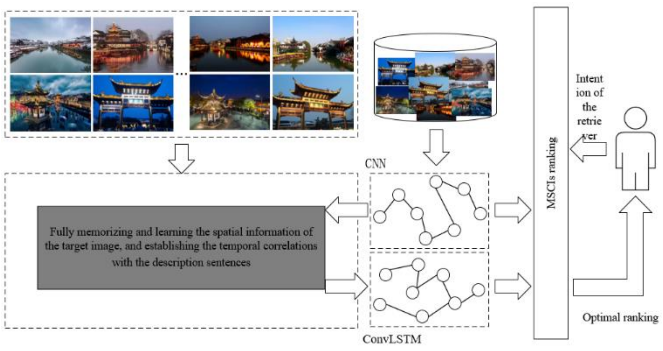

Figure 4. The workflow of the MSCIs ranking algorithm

Figure 4 explains the workflow of the MSCIs ranking algorithm. First, a similarity matrix SM^{Time}_{ij} was constructed based on the temporal features of the target tourist attraction image, and another similarity matrix SM^{Space}_{ij} was constructed based on the spatial features. Then, the objective function of the sorting model for two spatiotemporally correlated images can be expressed as:

$$\begin{aligned} OF(RS) = {} & \frac{\tau}{2}\sum_{i,j=1}^{n} SM_{ij}^{Time}\left(\frac{RS_i}{\sqrt{DM_{ii}^{Time}}}-\frac{RS_j}{\sqrt{DM_{jj}^{Time}}}\right)^2 \\ & +\frac{1-\tau}{2}\sum_{i,j=1}^{n} SM_{j}^{Space}\left(\frac{RS_i}{\sqrt{DM_{ii}^{Space}}}-\frac{RS_j}{\sqrt{DM_{jj}^{Space}}}\right)^2 \\ & +\frac{\theta}{2}\left\|RS-Label\right\|^2 \end{aligned} \tag{20}$$

where, $\tau\in[0, 1]$ can adjust the weights of the two feature maps; $D^{(IG)}$ is the diagonal matrix.

In formula (20), the loss function and regularization term are both real-valued differentiable functions on the convex subset. Hence, the closed solution of the objective function was directly derived.

First, find the first-order partial derivative of RS in formula (20) and set it to zero:

$$\begin{aligned} & \frac{\partial OF}{\partial RS}\Big|_{RS=RS^*} = \tau\left(RS^* - NSM^{Time}RS^*\right)+ \\ & (1-\tau)\left(RS^* - NSM^{Space}RS^*\right)+\theta\left(RS^*-Label\right)=0 \end{aligned} \tag{21}$$

where, NSM^{Time} and NSM^{Space} are the symmetrically normalized similarity matrices:

$$\begin{cases} NSM^{Time} = \sqrt{DM^{Time}}\,SM^{Time}\sqrt{DM^{Time}} \\ NSM^{Space} = \sqrt{DM^{Space}}\,SM^{Space}\sqrt{DM^{Space}} \end{cases} \tag{22}$$

Formula (22) can be transformed into:

$$\begin{aligned} RS^* - \frac{\tau}{1+\theta}NSM^{Time}RS^* - \frac{1-\tau}{1+\theta}NSM^{Space}RS^* \\ -\frac{\theta}{1+\theta}Label = 0 \end{aligned} \tag{23}$$

Let $u=\tau/(1+\theta)$, $v=(1+\tau)/(1+\theta)$, and $w=\theta/(1+\theta)$. The above formula can be converted into:

$$\left(I_1 - uNSM^{Time} - vNSM^{Space}\right)RS^* = wLabel \tag{24}$$

Since $u+v+w=1$, I_1-$uNSM^{Time}$-$vNSM^{Space}$ is irreversible. Thus, the closed solution of the objective function can be obtained by:

$$RS^* = w\left(I_1 - uNSM^{Time} - vNSM^{Space}\right)Label \tag{25}$$

Through the above derivation, the closed solution obtained can be directly used to calculate the optimal ranking score of each image. However, the similarity matrices will be extremely large, if the candidate spatiotemporally correlated images are high-dimensional. In this case, it will be very complex to normalize or invert the matrices. To solve the problem, the objective function was solved by the iterative method:

$$\begin{aligned} RS(t+1) = {} & \left(\frac{\tau}{1+\theta}NSM^{Time}+\frac{1+\tau}{1+\theta}NSM^{Space}\right)RS(t) \\ & +\frac{\theta}{1+\theta}Label \end{aligned} \tag{26}$$

During the sorting of MSCIs, the image set was expanded from two feature maps to multiple feature maps. Then, the objective function of MISCs sorting model can be expressed as:

$$\begin{aligned} OF'(RS) = {} & \sum_{z=1}^{Z}\sum_{i,j=1}^{n}\eta_z SM_{z,ij} \\ & \left(\frac{RS_i}{\sqrt{DM_{z,ii}}}-\frac{RS_j}{\sqrt{DM_{z,jj}}}\right)^2+\upsilon\left\|\eta\right\|^2 \end{aligned} \tag{27}$$

5. EXPERIMENTS AND RESULTS ANALYSIS

To verify the effects of coarse and fine screenings, this paper carries out image retrieval experiments on a standard library of 156,352 images on tourist attractions. Figure 5 presents the PD distribution between each target image and each image in the library.

During the screening of candidate spatiotemporally correlated images, the number N_B of points in UB should be configured reasonably to prevent the loss of highly similar images. A reasonable N_B naturally leads to the rationality of the threshold. The proposed MSCIs sorting algorithm was introduced to process 10 target images on tourist attractions. As shown in Table 1, the number of fine screened images was about 18% of that of coarse screened images.

To describe the distribution of fine screened images in the set of candidates obtained through coarse screening, 35 images were chosen at equal intervals from the candidate set for comparison. Figure 6 clearly displays the distribution of these fine screened images. It can be seen that the fine screened images were not uniformly distributed in the set of candidates obtained through coarse screening. Further analysis shows that the fine screened images concentrated in areas with large spatiotemporal changes, that is, these images can effectively reflect the change law of spatiotemporal feature.

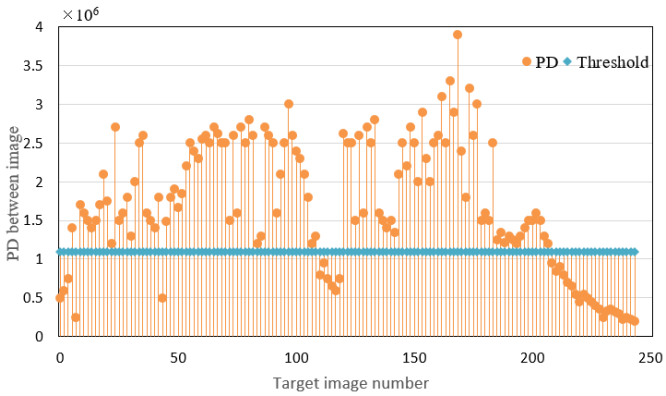

Figure 5. The PD distribution between each target image and each image in the library

Table 1. The set of candidate spatiotemporally correlated images

Target image number	Number of coarse screened images	Number of fine screened images
1	132	28
2	121	57
3	122	31
4	173	29
5	177	22
6	180	16
7	141	38
8	119	41
9	104	20
10	192	32

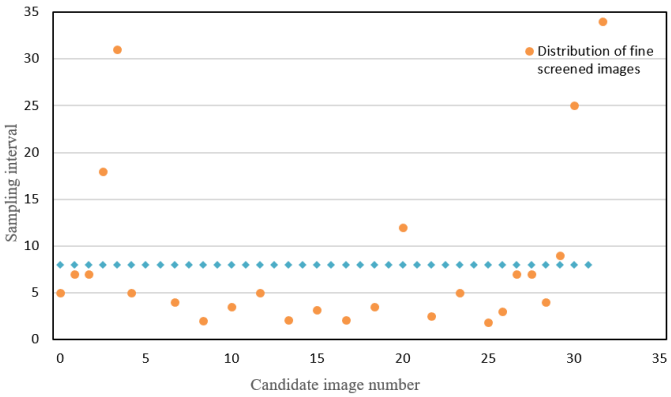

Figure 6. The distribution of fine screened images in the set of candidates obtained through coarse screening

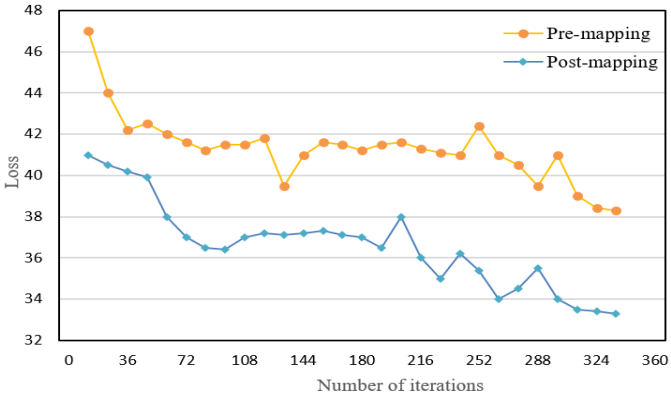

Figure 7. The training losses of our neural network before and after feature mapping by spatial attention

Further, a contrastive experiment was designed to verify the promoting effect of spatial attention method on neural network performance. Figure 7 compares the training losses of our neural network before and after feature mapping by spatial attention. Before the mapping, the training loss of the neural network before feature mapping descended slower and less significantly than that after feature mapping. This means the spatial attention method can improve the learning ability of the neural network by dynamically allocating the weights to the network output O_t.

Through the previous analysis, it can be known that there are 3 parameters in the solution of the objective function of the MSCIs ranking model, namely, u, v, and w. For convenience, the value of w was set to 0.01, and only one of v and w was adjusted, because the sum of them is roughly 1. Figure 8 shows the influence of u adjustment on the mean average precision (MAP) of the retrieval of tourist attraction images with different noise levels. As the noise level rose from 0 to 0.6, the u value continued to growth. In other words, the u value should be smaller for a less noisy image. The inverse is also true.

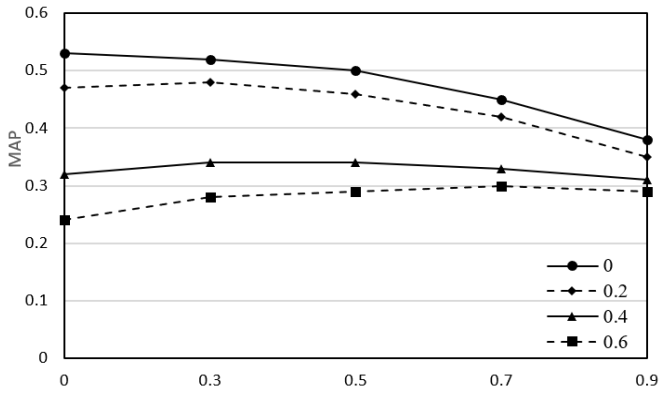

Figure 8. The influence of u adjustment on MAP

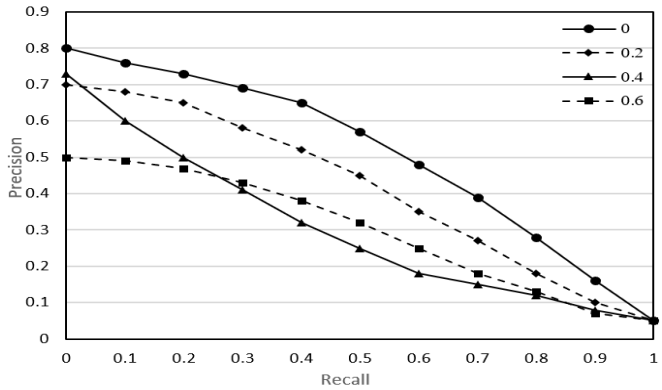

Figure 9. The P-R curves of the proposed algorithm

Figure 9 presents the precision-recall (P-R) curves of our algorithm in the retrieval of tourist attraction images with different noise levels. The x axis and y axis represent the recall and precision, respectively. It can be seen that our algorithm is not sensitive to the noise level of the target image. As the noise level rose from 0 to 0.6, the proposed algorithm became increasingly adaptive and robust.

Figure 10 compares the accuracies of our algorithm with scale invariant feature transform (SIFT), binary robust independent elementary features (BRIEF), and features from accelerated segment test (FAST) on candidate image sets of different volumes. Obviously, the proposed spatiotemporal

feature extraction algorithm had clear advantages over the other algorithms in retrieval accuracy.

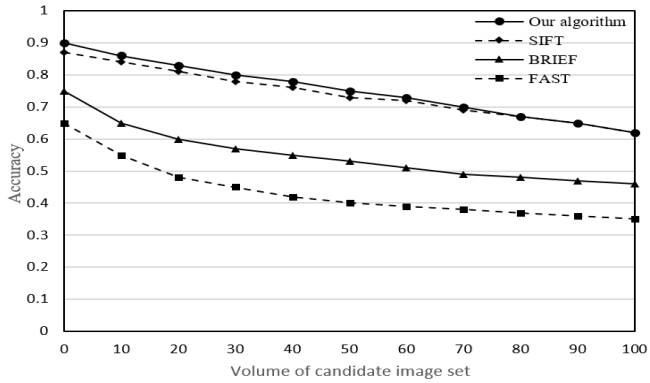

Figure 10. The retrieval accuracies of different methods on candidate image sets of different volumes

6. CONCLUSIONS

Based on DRCN, this paper proposes a spatiotemporal feature extraction method, and an MSCIs ranking and retrieval method for tourist attraction images. Firstly, the candidate spatiotemporally correlated images were acquired through coarse and fine screenings. According to the experimental results on the PD distribution between each target image and each image in the library, the fine screened images can effectively reflect the change law of spatiotemporal features. Next, the authors detailed the principle of spatiotemporal feature extraction, and the algorithm of the proposed ConvLSTM. Experimental results show that the spatial attention method can promote the learning ability of our neural network. Finally, the MSICs sorting model was constructed and derived, and proved superior in image retrieval accuracy through experiments.

ACKNOWLEDGMENT

Research on Strategies for Memory Protection and Inheritance of Industrial and Trade Traditional Villages in Henan from the Perspective of Village Culture, Project No. 2021-ZZJH-453; Research on Spatial Satisfaction Evaluation and Regeneration Protection Strategy of Traditional Village Context Inheritance in Southern Henan Province.2021-ZDJH-0422; Research on the characteristic landscape color recognition system and planning approach of traditional villages in western Henan province, Humanities and Social Sciences research project of Education Department of Henan province in 2020, Project No. 2020-ZZJH-513; Research on Spatial Feature Improvement design of Traditional Village Landscape in Southern Henan Under Protection Early Warning Strategy, Project No. 2020-ZZJH-519; Research on the Memory Inheritance Strategy of industrial and trade Traditional Villages in the Central Plains under the background of rural revitalization, Research project of Henan Federation of Social Sciences, Project No.SKL-2020-1126; Research on the public elective courses of art in Colleges and Universities under the background of quality education construction. Subject of Henan province's 13th five-year plan for education, Project No. (2019) -JKGHYB-0082; A study on the construction of curriculum system of design in Henan province under the mode of Chinese-foreign cooperation in Running schools, Subject of Henan province's 13th five-year plan for education, Project No. (2019) -JKGHYB-0080; Research on promoting the characteristic development of Henan cultural industry with social innovation, Subject of Henan social science planning, Project No. 2018BYS022.

REFERENCES

[1] Valem, L.P., Pedronette, D.C.G. (2020). Unsupervised selective rank fusion for image retrieval tasks. Neurocomputing, 377: 182-199. https://doi.org/10.1016/j.neucom.2019.09.065

[2] Pinjarkar, L., Sharma, M., Selot, S. (2018). Deep CNN combined with relevance feedback for trademark image retrieval. Journal of Intelligent Systems, 29(1): 894-909. https://doi.org/10.1515/jisys-2018-0083

[3] Hoxha, G., Melgani, F., Demir, B. (2020). Toward remote sensing image retrieval under a deep image captioning perspective. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 4462-4475. https://doi.org/10.1109/JSTARS.2020.3013818

[4] Tu, N.A., Dinh, D.L., Rasel, M.K., Lee, Y.K. (2016). Topic modeling and improvement of image representation for large-scale image retrieval. Information Sciences, 366: 99-120. https://doi.org/10.1016/j.ins.2016.05.029

[5] Suditu, N., Fleuret, F. (2016). Adaptive relevance feedback for large-scale image retrieval. Multimedia Tools and Applications, 75(12): 6777-6807. https://doi.org/10.1007/s11042-015-2610-9

[6] Mohamadzadeh, S., Farsi, H. (2016). Content-based image retrieval system via sparse representation. IET Computer Vision, 10(1): 95-102. https://doi.org/10.1049/iet-cvi.2015.0165

[7] Mathan Kumar, B., PushpaLakshmi, R. (2020). An approach for image search and retrieval by cluster-based indexing of binary MKSIFT codes. The Computer Journal, 63(6): 857-879. https://doi.org/10.1093/comjnl/bxz145

[8] Zhuang, Y., Chiu, D.K., Jiang, G., Hu, H., Jiang, N. (2013). Effective location-based geo-tagged image retrieval for mobile culture and tourism education. In International Conference on Web-Based Learning, pp. 152-161. https://doi.org/10.1007/978-3-642-41175-5_16

[9] Sokic, E., Konjicija, S. (2016). Phase preserving Fourier descriptor for shape-based image retrieval. Signal Processing: Image Communication, 40: 82-96. https://doi.org/10.1016/j.image.2015.11.002

[10] Boato, G., Dang-Nguyen, D.T., Muratov, O., Alajlan, N., De Natale, F.G. (2016). Exploiting visual saliency for increasing diversity of image retrieval results. Multimedia Tools and Applications, 75(10): 5581-5602. https://doi.org/10.1007/s11042-015-2526-4

[11] Raisi, Z., Mohanna, F., Rezaei, M. (2011). Content-based image retrieval for tourism application. In 2011 7th Iranian Conference on Machine Vision and Image Processing, pp. 1-5. https://doi.org/10.1109/IranianMVIP.2011.6121857

[12] Guo, K., Zhang, R., Zhou, Z., Tang, Y., Kuang, L. (2016). Combined retrieval: A convenient and precise approach for internet image retrieval. Information Sciences, 358:

151-163. https://doi.org/10.1016/j.ins.2016.04.001
[13] Beaudoin, J.E. (2016). Content-based image retrieval methods and professional image users. Journal of the Association for Information Science and Technology, 67(2): 350-365. https://doi.org/10.1002/asi.23387
[14] Houaria, A.B.E.D., Zaoui, L. (2016). Improving image retrieval using a data mining approach. Inteligencia Artificial, 19(57): 97-113. https://doi.org/10.4114/ia.v18i56.1147
[15] Ordonez, V., Han, X., Kuznetsova, P., Kulkarni, G., Mitchell, M., Yamaguchi, K., Stratos, K., Goyal, A., Dodge, J., Mensch, A., Daumé, H., Berg, A.C., Choi, Y., Berg, T.L. (2016). Large scale retrieval and generation of image descriptions. International Journal of Computer Vision, 119(1): 46-59. https://doi.org/10.1007/s11263-015-0840-y
[16] Passalis, N., Iosifidis, A., Gabbouj, M., Tefas, A. (2020). Variance-preserving deep metric learning for content-based image retrieval. Pattern Recognition Letters, 131: 8-14. https://doi.org/10.1016/j.patrec.2019.11.041
[17] Selamat, A., Ismail, M.K. (2008). Gustafson-Kessel algorithm in content based image retrieval for Malaysia tourism website. In 2008 International Symposium on Information Technology, 3: 1-6. https://doi.org/10.1109/ITSIM.2008.4632018
[18] Zhao, S., Potdar, V., Chang, E. (2007). A practical image retrieval framework for tourism industry. In 2007 IEEE International Symposium on Industrial Electronics, pp. 2928-2932. https://doi.org/10.1109/ISIE.2007.4375079
[19] Chantrapornchai, C., Bunlaw, N., Choksuchat, C. (2018). Semantic image search: Case study for western region tourism in Thailand. JIPS, 14(5): 1195-1214. https://doi.org/10.3745/JIPS.04.0088
[20] Kim, S.E., Lee, K.Y., Shin, S.I., Yang, S.B. (2017). Effects of tourism information quality in social media on destination image formation: The case of Sina Weibo. Information & Management, 54(6): 687-702. https://doi.org/10.1016/j.im.2017.02.009
[21] Fukada, H., Kasai, K., Ohtsu, S. (2015). A field experiment of system to provide tourism information using image recognition type AR technology. In New Trends in Networking, Computing, E-learning, Systems Sciences, and Engineering, pp. 381-387. https://doi.org/10.1007/978-3-319-06764-3_47
[22] Latorre-Martínez, M.P., Iñíguez-Berrozpe, T., Plumed-Lasarte, M. (2014). Image-focused social media for a market analysis of tourism consumption. International Journal of Technology Management, 64(1): 17-30. https://doi.org/10.1504/IJTM.2014.059234
[23] Yoshihara, T., Nishina, D., Tanaka, T., Kawase, K., Takagishi, H. (2017). A study on the psychological evaluation of tourism landscape images in Hiroshima a psychological evaluation by Korean subjects. Journal of Asian Architecture and Building Engineering, 16(1): 223-229. https://doi.org/10.3130/jaabe.16.223
[24] Kawase, H., Nishina, D., Lu, W., Jin, H., Tanaka, T., Yoshihara, T. (2015). A study on the psychological evaluation of tourism landscape images in Hiroshima-Psychological Evaluation by Chinese students, Chinese foreign students and Japanese students. Archit. Plann. Environ. Eng., AIJ, 708: 99-108.
[25] An, L., Zou, C., Zhang, L., Denney, B. (2016). Scalable attribute-driven face image retrieval. Neurocomputing, 172: 215-224. https://doi.org/10.1016/j.neucom.2014.09.098