Deep Cross-Modal Retrieval for Remote Sensing Image and Audio

GUO Mao[a,b], YUAN Yuan[a], LU Xiaoqiang[a]
[a] Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences,
Xi'an, 710119, Shaanxi, P. R. China.
[b] University of Chinese Academy of Sciences, Beijing, 100049, P. R. China.
guomao2016@opt.cn, {yuany,luxiaoqiang}@opt.ac.cn

***Abstract*—Remote sensing image retrieval has many important applications in civilian and military fields, such as disaster monitoring and target detecting. However, the existing research on image retrieval, mainly including to two directions, text based and content based, cannot meet the rapid and convenient needs of some special applications and emergency scenes. Based on text, the retrieval is limited by keyboard inputting because of its lower efficiency for some urgent situations and based on content, it needs an example image as reference, which usually does not exist. Yet speech, as a direct, natural and efficient human-machine interactive way, can make up these shortcomings. Hence, a novel cross-modal retrieval method for remote sensing image and spoken audio is proposed in this paper. We first build a large-scale remote sensing image dataset with plenty of manual annotated spoken audio captions for the cross-modal retrieval task. Then a Deep Visual-Audio Network is designed to directly learn the correspondence of image and audio. And this model integrates feature extracting and multi-modal learning into the same network. Experiments on the proposed dataset verify the effectiveness of our approach and prove that it is feasible for speech-to-image retrieval.**

Keywords—cross-modal retrieval; remote sensing image; spoken audio; convolutional neural network

I. INTRODUCTION

Remote sensing has been rapidly developed in recent years and widely used in applications, such as resource investigating, surveying and mapping, disaster relief, military, commerce field, etc. With image data increasing greatly, rapid retrieval and easy achievement of the desired image from the mass images are becoming a tricky problem.

Many researchers have studied the problem of image retrieval and it is mainly divided into two branches, text-to-image [1][2] and image-to-image [3][4][5]. The text-to-image approach highly depends on the availability and the quality of manual tags or labels. However, keyboard inputting is much lower in efficiency for some urgent situations, and sometimes even cannot be available. For example, in military targets detecting or in disaster monitoring, it is emergent and not convenient for keyboard input. The image-to-image method needs an example image as input to the query, which usually does not exist in practical applications. These defects of the current retrieval methods are not conducive to the applications of some emergency and special circumstances.

A retrieval system, from the users' perspective, should be fast and easy. It is known that speech is the primary way for human communication and much more convenient and faster than writing or typing. Thus image retrieval through speech will greatly increase the convenience and efficiency, and the widely used speech devices in mobile terminals provide strong support in physical profile. The present methods of image retrieval by speech are mostly based on speech recognition technology. It firstly transcribes speech to text and then applies text to conventional image retrieval. Yet this course of transformation will lose some useful details for retrieval. Then to generate the correspondence between audio and image directly is necessary.

The retrieval from audio to image is a cross-modal task. The main challenge of cross-modal retrieval is the heterogeneous gap [6], which implies that the similarities of different modalities cannot be measured directly. For instance, it is impracticable to measure the distance of image and audio because of their heterogeneity. Many convolutional neural network (CNN) based models have been applied to solve this problem on image and sentence mapping. But there are few for image and audio.

In this paper we propose a *Deep Visual-Audio Network* (DVAN). The correspondence of image and its spoken audio caption is utilized here as supervised information to train the network. Through the DVAN network, high-level semantic information of both visual and audio modalities can be learned. And the correlation of images and audio is directly generated from the data, without the course of speech recognition and text transcription, or any supplementary information. The main contributions of this paper can be summarized as follows:

1) A cross-modal retrieval method of remote sensing images and audio is proposed in this paper. This method proves the feasibility of speech-to-image retrieval, and provides a new way for image retrieval.

2) In this paper, a new convolutional neural network for audio modality, which is a subnetwork of DVAN, is designed to obtain good semantic representations of audio.

3) A remote sensing image audio caption dataset is constructed for training and evaluating our model. The cross-modal retrieval task is carried out on this proposed dataset and achieves good performances.

The remaining part of this paper is structured as follows. Section II briefly overviews some previous works. The details of the new dataset are described in Section III. In Section IV our DVAN model and its implemental details are introduced. Section V provides the experimental results and the related discussions. Finally, we conclude this paper in Section VI.

II. Related Work

A. Remote sensing image retrieval

In recent years, there has been much work toward developing a high-performance remote sensing image retrieval [4][5][7][8] model. Some initial methods focus on text-based retrieval like what we do in natural image retrieval. However, remote sensing image is different from natural scene image. Remote sensing image contains richer and more obvious spatial information, geometric structure, texture information and other details. So the retrieval method used in natural scene cannot be used directly on remote sensing images. Besides, many works are motivated by content-based methods. These methods usually calculate the low-level features of remote sensing images. Most of the work is concentrated on feature extraction and high dimensional index. For example, Zhang *et al*. [4] proposed a hyperspectral remote sensing image retrieval system using spectral and texture features. But like all the content-based retrieval methods, they need an example image as reference. And the low-level features of the image extracted automatically by the computer differ greatly from the users' query command. When users query and judge images, they use high-level words and concepts. These high-level concepts include the understanding of the objects in the image, that is, the semantic features, not the low-level image features. Xia *et al*. [7] focused on the visual feature aspect and delved how to use powerful deep representations in their task. Zhou *et al*. [5] learned low dimensional convolutional neural networks for high-resolution remote sensing image retrieval and indicated that deep feature representations achieved better performance than traditional hand-crafted features. Other approaches have focused on image coding and decoding. Therefore, our work, focusing on semantic features representation, aims to explore a semantic-based retrieval approach for remote sensing images.

B. Cross-modal retrieval

During the past decades, multi-modal modeling of image and text [9][10][11][12] has been widely studied in machine learning field. And many methods focus on accurately annotating objects and regions in images. For example, Karpathy *et al*. [9] introduced a multi-modal embedding of visual and natural language data to fulfil a bidirectional retrieval of images and sentences. Other similar methods also map images and sentences into a common space. And through the common space, the task of image and sentence retrieval can be accomplished. Recently, multimedia retrieval, including image, text, audio and so on, has attracted growing attention. There are a lot of researches focusing on finding a direct mapping between images and sounds [13][14][15][16][17][18]. Torfi *et al*. [14] devoted to finding the correspondence of audio-visual streams. The work most similar to ours is [13], in which the authors associate spoken audio captions to relevant regions in images. It did not split the audio sentence, but generated the spectrogram [19] of the whole sentence. And then it used a CNN to embed the spectrogram into a high dimensional space.

III. Dataset Establishing

Cross modal retrieval of remote sensing image and audio is an innovative problem in artificial intelligence. In this case, a dataset of remote sensing images with corresponding audio captions is quite necessary. However, currently there is only an audio caption dataset for natural images [13] and a sentence caption dataset for remote sensing images [10]. Therefore, a remote sensing image audio dataset is established in this paper.

We construct a new corpus of spoken audio captions for UCM dataset [20], Sydney dataset [20], and RSICD dataset [10]. RSICD dataset consists of 10921 images categorized into 30 different scene classes. UCM dataset contains 2100 images of 21 different classes and Sydney has 613 images of 7 categories, which provide a rich variety of ground object types.

To collect audio captions, each remote sensing image is given 5 sentences. In order to guarantee the quality of image captions, a uniform rule is drawn up in the process of audio annotation. Standard American pronunciation is used here to generate the spoken sentence. And to better satisfy the need of sound diversity, different speakers participate in collecting this dataset. Each sentence of an image is generated by 5 different speakers. Different lengths of spoken audio can also show the diversity of sound. The length of the audio varies from 1 seconds to 15 seconds. Most of them are concentrated in 4 to 6 seconds, about 69%. Approximately 93% is under 10 seconds.

IV. Proposed Approach

In this paper, we address the problem of cross-modal retrieval for remote sensing image and audio. Fig. 1 shows the architecture of our DVAN model, including three main parts: 1) Remote sensing image representation, 2) Audio representation, and 3) Multi-modal embedding. The remote sensing image part uses image subnetwork to extract high-level semantic feature from each remote sensing image. And the audio part extracts high-level semantic feature of the spoken audio caption by audio convolutional subnetwork. The multi-modal part is represented as a fusion layer to combine the image feature vector with audio feature vector.

A. Remote sensing image representation

To perform retrieval of remote sensing image, image representation is the first and important step. We use CNN to extract features from the original image layer by layer through multi-layer network architecture. CNN features are later proved to be very useful for the entire retrieval process because they can better express the high-level semantic information contained in the image. Concretely, the remote sensing image features are extracted by VGG 16 network [21] pretrained on ImageNet. The classification layer is discarded and the output of the FC7 layer is taken as the representation of the image. Therefore, a single 4096-dimension feature vector is gained here via the image branch. Let $I = \{I_i\}_{i=1}^{n}$ denotes the image.

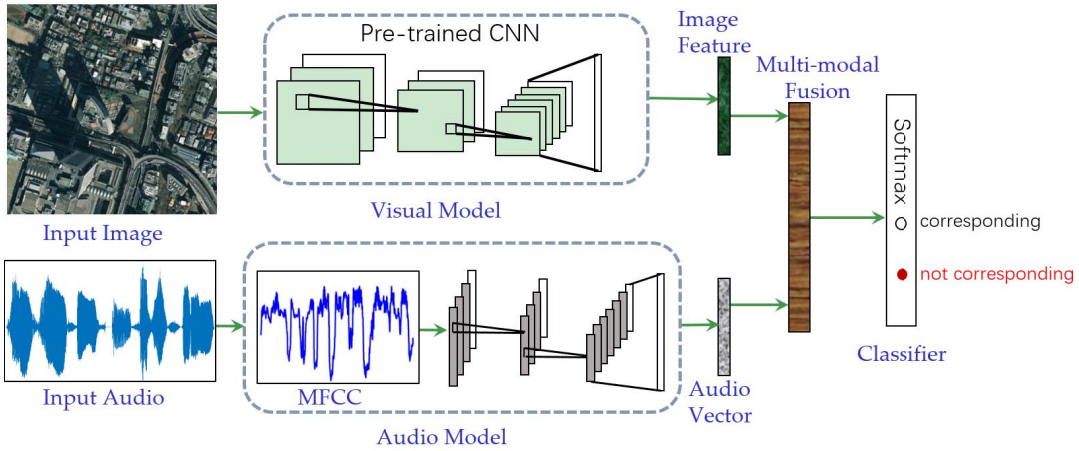

Fig. 1. The architecture of our DVAN network. Extracting features of images and sounds, respectively, these two feature vectors are fused in the fusion layer. Finally, we judge whether or not it is relevant by a softmax layer.

Then the learned representation of remote sensing image I is as follows:

$$V_I = f(I;\theta_I), \tag{1}$$

where θ_I is the parameters of the pretrained VGG 16 network for image modality, $V_I \in \mathbb{R}^{d_1}$ denotes the image feature, and d_1 denotes the dimension of the image feature.

B. Audio representation

Similar to images, audio representation is also a key process for cross-modal retrieval task. Let $S = \{S_j\}_{j=1}^{n}$ denotes the audio caption. When given a raw audio S_j, we first represent each audio caption as a vector with Mel-frequency Cepstral Coefficients (MFCC) [22]. The process can be written as follows:

$$M_j = \mathit{flatten}(\mathrm{MFCC}(S_j)), \tag{2}$$

where $M_j \in \mathbb{R}^{d_2}$ denotes the MFCC feature of j-th audio caption, and d_2 denotes the dimension of the MFCC feature which depends on the length of an audio.

The MFCC feature obtained from equation (2) has too much information redundancy and noise. And the dimension is too high, which would bring large amount of computation. To perform good feature representation for audio caption, we should go further in representing the audio. Therefore, an audio convolutional neural subnetwork is proposed here to get a powerful representation. The specific parameters of the audio subnetwork are shown in TABLE I.

The MFCC feature M is then taken as the input of the audio subnetwork and the learned audio feature is denoted as follows:

$$V_S = g(M;\theta_S), \tag{3}$$

where θ_S is the parameters of the convolutional neural network for audio modality, $V_S \in \mathbb{R}^{d_3}$ denotes the audio feature, and d_3 denotes the dimension of the audio feature.

C. Multi-modal embedding

This section focuses on learning a multi-modal feature to judge whether the input image-audio pair is related. That is, the image feature and the audio feature are both taken into account to produce the final decision. As illustrated in Fig. 1, a fusion layer is set here to represent the multi-modal feature. And the fused feature then passing through a softmax layer, produces a classification result, corresponding or not corresponding. In the process of retrieval, by comparing the probabilities of the corresponding class between the query audio and each image in the database, we determine whether or not they are relevant. The process of multi-modal embedding can be expressed as:

$$h = \tanh\left(W_I \cdot V_I + W_S \cdot V_S + b_I\right), \tag{4}$$

where W_I and W_S are weights, b_I denotes the bias, $h \in \mathbb{R}^{d_4}$ denotes the fusion feature vector, and d_4 is the dimension of the fusion feature.

$$p = \mathrm{softmax}\left(W_h h + b_h\right), \tag{5}$$

where W_h is the weights, b_h denotes the bias, $p \in \mathbb{R}^k$ is a probability vector and k denotes the dimension of p. Each dimension of p is the probability of belonging to the corresponding class.

The cross-entropy loss converges quickly when training the neural networks and is stable when combined with the softmax function. The cross-entropy loss function used in this paper can be defined as follows:

$$J(\theta) = -\frac{1}{B}\sum_{i,j=1}^{B} y^{(i,j)} \cdot \log(p^{(i,j)}) + (1 - y^{(i,j)}) \cdot \log(1 - p^{(i,j)}), \tag{6}$$

where B means each minibatch has B image-audio pairs. $p^{(i,j)}$ is the probability that the i-th image and j-th audio correspond. $y^{(i,j)}$ is the true label, defined as:

$$y^{(i,j)} = \begin{cases} 1 & I_i \text{ and } S_j \text{ correspond,} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

V. EXPERIMENTS

In this section, we trained and tested our proposed model and evaluated it on image retrieval and audio retrieval tasks. The test set of image and audio pairs is sampled from the dataset described in Section III. Given an audio caption, we preform image retrieval by using our model to calculate the probability of the corresponding class between this audio caption and each of the remote sensing images in our test set. The process of audio retrieval is similar, and the only difference is that one fixed remote sensing image and all pieces of audio are computed one by one to get the audio caption which best describes the image.

A. Experiment Setup

For the experiments in this paper, 80% image-audio pairs are taken as training set, 10% as validating set and the rest as test set. Positive samples are those corresponding pairs. And negative ones are randomly sampled from other classes. In the experiment, each training batch has the same number of corresponding pairs and not corresponding pairs. After the DVAN training process is accomplished, the testing set is then divided into query set and retrieval set. For the audio-to-image retrieval task, the query set contains the audio and the retrieval set contains the image, and vice versa.

1) *Remote sensing image representation:* In the proposed model, deep CNN features generated from the FC7 layer of VGG 16 are used to represent the remote sensing images. Before that, we resize the images into the same size 224×224. The dimension of the feature output from the FC7 layer is 4096. So, in Section IV.A, the parameter $d_1 = 4096$.

TABLE I. THE ARCHITECTURE FOR AUDIO SUBNETWORK

layer	output-size	kernel	stride
Conv1	12000x1x6	10x1x6	1x2x1
Conv2	6000x1x12	8x1x12	1x2x1
Pool3	3000x1x12	1x2x1	1x2x1
Conv4	1500x1x24	5x1x24	1x2x1
Conv5	750x1x48	3x1x48	1x2x1
Pool6	375x1x48	1x2x1	1x2x1

2) *Audio representation:* Audio feature extraction includes MFCC representing and audio subnetwork representing. An MFCC is computed with a 16 millisecond window size and a 5 millisecond shift. In order to reduce the amount of calculation and improve efficiency, we force the length of MFCC to be a fixed size *m*, truncating those longer than *m* and padding zero into shorter ones. Thus one dimensional vector of a fixed length takes place of the audio in the wav form as the input of the audio subnetwork. The architecture for audio subnetwork is shown in TABLE I. So, in Section IV.B, the parameters $d_2 = 24000$ and $d_3 = 1024$.

3) *Multi-modal embedding:* Multi-modal embedding utilizes a fusion layer to learn the common space of audio and image. The fusion layer generates a fusion feature of image and audio from the FC layer (see equation (4)). The dimension of the fusion feature is 512. In other words, the parameter $d_4 = 512$. A softmax layer divides the fusion features into two classes, so in Section IV.C the parameter $k = 2$.

B. Evaluation metric

We use four widely used metric indexes to evaluate the performance of our model in this paper.

1) *Mean average precision* (MAP): MAP is the mean value of average precision (AP). AP of top L retrieved instances is defined as:

$$\mathrm{AP} = \frac{1}{NL}\sum_{r=1}^{L} P(r) \times \mathrm{rel}(r), \tag{8}$$

where r is the rank, $P(r)$ is the precision of the top L retrieved instances, N is the number retrieved, and $rel(\cdot)$ is a binary function on the relevance of a given rank, equaling 1 if the item at rank r is relevant, 0 otherwise.

2) *Precision* ($P@K$): Precision is the fraction of relevant instances among the retrieved instances. $P@K$ denotes the precision in top K of the rank. Here, K equals to 1, 5 and 10. The precision is $P@1$, $P@5$ and $P@10$, respectively.

3) *Recall rates* ($R@K$): Recall is the fraction of relevant retrieved instances over the total amount of relevant instances. The same as $P@K$, $R@K$ denotes the recall rate in top K of the rank list. K equals to 1, 5 and 10.

4) *F-score* ($F@K$): F-score combines the results of P and R.

$$F = \frac{2PR}{P+R}. \tag{9}$$

These metric indexes reflect the performance of our model in different views. If the index is higher, the test method is more effective.

C. Results and Analysis

In order to find a better model, we set different parameters on our model architectures. The coefficient of MFCC and its first order derivative and second order derivative are taken as the representation of audio. The standard cepstrum parameter MFCC only reflects the static characteristics of voice parameters, and the dynamic characteristics of audio can be described by the derivative spectrum of these static characteristics.

1) *Matching results*: TABLE II shows the results of matching between image and audio. According to the correspondence of image-audio pairs, they are divided into two classes, matching or not matching. Here, M denotes MFCC coefficient, ΔM and Δ^2M denote the first and second order derivative of MFCC, respectively. CNN denotes the image features represented by VGG 16. SPEC denotes the spectrogram of the audio. SIFT+M means that using SIFT to represent remote sensing image and using MFCC to represent audio. The meanings of other models are the same. As shown in TABLE II, our model achieves better results than those

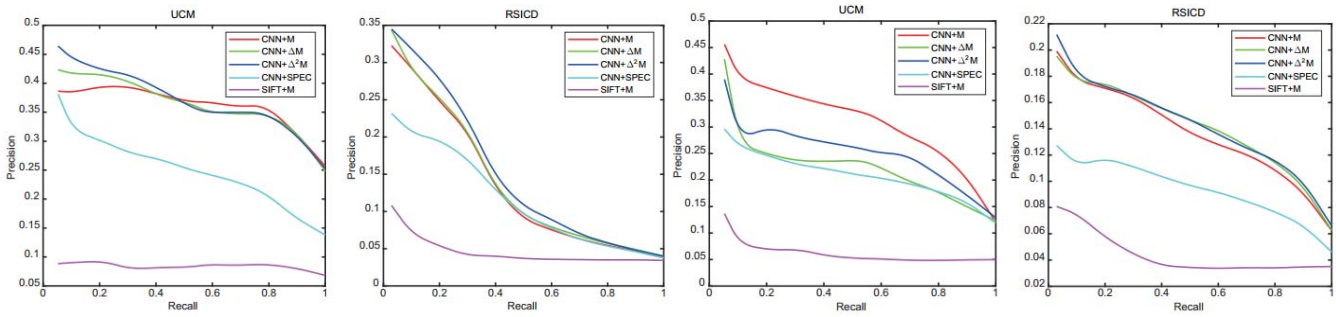

Fig. 2. Precision-Recall curves on UCM and RSICD dataset. (a) (b) represent the performances for the audio-to-image retrieval task and (c) (d) represents the performances for the image-to-audio retrieval task.

using SIFT as image feature and spectrogram as audio feature. SIFT features represent the low-level features of remote sensing images and cannot express the high-level semantic information of images accurately. This has been clearly verified by the result. It is demonstrated that the better performance is largely due to the good representation of CNN features. From the experimental results, it is also obvious that the spectrogram representation of sound is not as good as MFCC features. Spectrogram is a two-dimension picture, and still has much redundant information. MFCC features passing through the audio subnetwork can obtain an effective representation of audio.

TABLE II. MATCHING RESULTS (%)

Model	RSICD	UCM	Sydney
SIFT+M	50.00	50.24	53.02
CNN+SPEC	57.72	72.90	43.10
CNN+M	64.81	**83.21**	70.26
CNN+ΔM	64.08	79.02	**75.00**
CNN+Δ^2M	**65.22**	78.90	74.57

TABLE III. EXPERIMENTAL RESULTS ON SYDNEY DATASET (%)

Task	Model	MAP	P@1	P@5	P@10	R@1	R@5	R@10	F@1	F@5	F@10
	SIFT+M	26.50	34.48	24.48	23.28	0.92	4.52	9.24	1.78	7.62	13.23
	CNN+SPEC	35.72	17.24	27.76	31.21	0.96	7.27	16.21	1.82	11.52	21.33
A→I	CNN+M	63.88	67.24	63.34	67.07	2.95	13.12	29.10	5.65	21.73	40.59
	CNN+ΔM	62.91	**69.83**	63.10	69.05	**3.28**	13.09	31.06	**6.27**	21.68	**42.84**
	CNN+Δ^2M	**64.41**	68.10	**64.14**	**70.26**	2.96	**13.67**	**32.22**	5.67	**22.53**	44.18
	SIFT+M	31.67	11.21	35.00	37.59	0.76	4.74	9.50	1.43	8.35	15.17
	CNN+SPEC	46.67	58.62	55.00	51.64	2.02	9.66	19.03	3.90	16.43	27.81
I→A	CNN+M	**71.77**	75.86	73.62	72.93	3.43	15.85	31.04	6.57	26.09	43.55
	CNN+ΔM	70.77	**77.59**	74.14	**75.00**	**3.56**	15.98	**32.83**	**6.81**	26.30	**45.66**
	CNN+Δ^2M	71.05	75.00	**74.48**	74.24	3.25	**16.33**	31.65	6.24	**26.79**	44.38

TABLE IV. EXPERIMENTAL RESULTS ON UCM DATASET (%)

Task	Model	MAP	P@1	P@5	P@10	R@1	R@5	R@10	F@1	F@5	F@10
	SIFT+M	6.66	3.59	4.41	4.68	0.19	1.16	2.46	0.35	1.83	3.22
	CNN+SPEC	21.79	19.42	19.86	19.23	1.03	5.26	10.19	1.95	8.32	13.32
A→I	CNN+M	**32.38**	**32.37**	**33.91**	**34.34**	**1.73**	**9.00**	**18.23**	**3.28**	**14.23**	**23.82**
	CNN+ΔM	23.62	30.22	23.31	21.25	1.61	6.19	11.27	3.05	9.79	14.73
	CNN+Δ^2M	26.34	29.26	22.59	24.65	1.55	6.01	13.12	2.95	9.50	17.12
	SIFT+M	8.55	4.56	4.65	4.56	0.24	1.24	2.45	0.46	1.96	3.18
	CNN+SPEC	26.25	29.50	25.52	23.65	1.55	6.73	12.48	2.95	10.66	16.34
I→A	CNN+M	36.79	32.37	33.29	33.74	1.72	8.85	17.96	3.27	13.99	23.45
	CNN+ΔM	37.42	36.93	36.98	35.11	1.96	9.81	18.66	3.72	15.51	24.36
	CNN+Δ^2M	**38.42**	**42.21**	**37.60**	**35.90**	**2.24**	**9.98**	**19.05**	**4.26**	**15.77**	**24.89**

TABLE V. EXPERIMENTAL RESULTS ON RSICD DATASET (%)

Task	Model	MAP	P@1	P@5	P@10	R@1	R@5	R@10	F@1	F@5	F@10
	SIFT+M	4.85	3.66	3.60	3.54	0.09	0.47	0.93	0.18	0.83	1.47
	CNN+SPEC	9.96	7.13	7.00	7.44	0.19	0.94	2.00	0.37	1.65	3.15
A→I	CNN+M	**15.71**	**16.18**	**15.10**	**14.76**	**0.46**	**2.11**	4.13	**0.89**	**3.71**	**6.46**
	CNN+ΔM	15.01	10.51	13.93	14.17	0.29	1.97	4.05	0.56	3.45	6.30
	CNN+Δ^2M	15.24	12.43	13.78	14.39	0.34	1.97	**4.15**	0.67	3.44	6.45
	SIFT+M	5.04	6.22	5.34	4.50	0.10	0.47	0.90	0.19	0.86	1.51
	CNN+SPEC	13.24	16.82	16.62	15.69	0.42	2.13	4.02	0.82	3.77	6.40
I→A	CNN+M	16.29	22.49	22.56	21.77	0.66	3.33	6.28	1.28	5.80	9.75
	CNN+ΔM	16.55	23.77	24.30	21.94	0.67	3.45	6.13	1.30	6.05	9.58
	CNN+Δ^2M	**17.84**	**24.95**	**26.34**	**24.27**	**0.72**	**3.83**	**6.92**	**1.40**	**6.69**	**10.77**

2) *Results of retrieval*: As illustrated in TABLE III-V, image retrieval using CNN and MFCC features is better than that using SIFT and spectrogram features. Here, **A→I** means image retrieval using audio, and **I→A** means audio retrieval using image. We can find that in all cases our model can achieve the best precision, recall and F-score. For MAP, our model outperforms others in all cases. Specifically, in image retrieval task, CNN+M achieves the best results and in audio retrieval task, CNN+ΔM is the best. For precision, recall and F-score, MFCC coefficients including first and second order derivative outperform spectrogram, and CNN feature outperforms SIFT feature. In summary, the experiment results indicate the effectiveness of our proposed model.

VI. CONCLUSION

In this paper, we establish a model of cross-modal retrieval for remote sensing images and audio, and create a new dataset of remote sensing images and audio captions. Our proposed DVAN architecture consists of two subnetworks, visual branch and audio branch, and then they are combined at the fusion layer. It directly learns the correspondence of these two modes through looking at remote sensing images and listening to audio captions. In this way, we have accomplished an innovative task to greatly improve the efficiency of human-computer interaction in remote sensing retrieval. Our model has been validated on image retrieval and audio retrieval, which shows its application prospect.

REFERENCES

[1] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao, "Discriminative dictionary learning with common label alignment for cross-modal retrieval," IEEE Transactions on Multimedia, vol. 18, no. 2, pp. 208-218, 2016.

[2] Y. Verma and C. V. Jawahar, "Im2text and text2im: Associating images and texts for cross-modal retrieval," British Machine Vision Conference, 2014.

[3] T. Costachioiu, I. Nita, V. Lazarescu, and M. Datcu, "A semantic framework for data retrieval in large remote sensing databases," IEEE International Geoscience and Remote Sensing Symposium, 2012, pp. 5285-5288.

[4] J. Zhang, W. Geng, X. Liang, J. Li, L. Zhuo, and Q. Zhou, "Hyperspectral remote sensing image retrieval system using spectral and texture features." Applied Optics, vol. 56, no. 16, pp. 4785-4796, 2017.

[5] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," Remote Sensing, vol. 9, no. 5, p. 489, 2017.

[6] J. Zhang, Y. Peng, and M. Yuan, "Unsupervised generative adversarial cross-modal hashing," CoRR, vol. abs/1712.00358, 2017.

[7] G. Xia, X. Tong, F. Hu, Y. Zhong, M. Datcu, and L. Zhang, "Exploiting deep features for remote sensing image retrieval: A systematic investigation," arXiv preprint arXiv:1707.07321, 2017.

[8] P. Li and P. Ren, "Partial randomness hashing for large-scale remote sensing image retrieval," IEEE Geoscience Remote Sensing Letter, vol. 14, no. 3, pp. 464-468, 2017.

[9] A. Karpathy, A. Joulin, and F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," Advances in Neural Information Processing Systems, 2014, pp. 1889-1897.

[10] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 4, pp. 2183-2195, 2017.

[11] A. Yuan, X. Li, and X. Lu, "FFGS: feature fusion with gating structure for image caption generation," Computer Vision - Second CCF Chinese Conference, 2017, pp. 638-649.

[12] J. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," Advances in Neural Information Processing Systems, 2017, pp. 465-476.

[13] D. F. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," Advances in Neural Information Processing Systems, 2016, pp. 1858-1866.

[14] A. Torfi, S. M. Iranmanesh, N. M. Nasrabadi, and J. M. Dawson, "3d convolutional neural networks for cross audio-visual matching recognition," IEEE Access, vol. 5, pp. 22 081-22 091, 2017.

[15] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[16] R. Arandjelovic and A. Zisserman, "Look, listen and learn," IEEE International Conference on Computer Vision, 2017, pp. 609-617.

[17] B. George and B. Yegnanarayana, "Unsupervised query-by-example spoken term detection using segment-based bag of acoustic words," IEEE International Conference on Acoustics, Speech and Signal Processing, 2014, pp. 7133-7137.

[18] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, and H. T. Shen, "Unsupervised cross-modal retrieval through adversarial learning," in IEEE International Conference on Multimedia and Expo, 2017, pp. 1153-1158.

[19] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," International Conference on Deep Learning and Music, 2017.

[20] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," International Conference on Computer, Information and Telecommunication Systems, 2016, pp. 1-5.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Computer Science, 2014.

[22] W. Han, C. Chan, O. C. Choy, and K. Pun, "An efficient MFCC extraction method in speech recognition," International Symposium on Circuits and Systems, 2006.