IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

RECOGNITION OF GEOTAGGED AUDIOVISUAL AERIAL SCENE

Kiran Onapakala
Integration software Developer
Pacific Source health plans
Portland, Oregon, USA
kiran.onapakala@pacificsource.com

ABSTRACT

Aerial scene recognition is an essential task in remote sensing and has recently received enlarged attention. This paper studies the improving performance on the aerial scene recognition. This explores a novel audiovisual aerial scene recognition task using both images and sounds as input. Based on an observation that some specific sound events are more likely to be heard at a given geographic location, we propose to exploit the knowledge from the sound events to improve the performance on the aerial scene recognition. For this purpose, we have used dataset named AuDio Visual Aerial sceNe reCognition dataset. With the help of this dataset, we evaluate three proposed approaches for transferring the sound event knowledge to the aerial scene recognition task in a multimodal learning framework, and show the benefit of develop the audio information for the aerial scene recognition.

1. Introduction

Scene recognition is a longstanding, hallmark problem in the field of computer vision, and it refers to assigning a scene-level label to an image based on it overall contents. Most scene recognition approaches in the community make use of ground images and have achieved remarkable performance. The success of current state-of-the-art aerial scene understanding models can be attributed to the development of novel convolutional neural networks (CNNs) that aim at learning good visual representations from images. Albeit successful, these models may not work well in some cases, particularly whenthey are directly used in worldwide applications, sufering the pervasive factors, such as diff erent remote imaging sensors, lighting conditions, orientations, and seasonal variations. A study in neurobiology reveals that human perception usually benefit ts from the integration of both visual and auditory knowledge. In spired by this investigation, we argue that aerial scenes soundscapes are partially free of the aforementioned factors and can be a helpful cue for identifying scene categories.

2. Related work

Some related works in aerial scene recognition, multimodal learning, and cross-task transfer. Aerial Scene Recognition. Earlier studies on aerial scene recognition [1] mainly focused on extracting low-level visual attributes and/or modeling mid level spatial features [2]. Recently, deep networks, especially CNNs, have achieved a large development in aerial scene recognition [3]. Moreover, some methods were proposed to solve the problem of the limited collection of aerial images by employing more efficient networks [4]. Although these methods have achieved great empirical success, they usually learn scene knowledge from the same modality, i.e., image. Different from previous works, this paper mainly focuses on exploiting multiple modalities (i.e. image and sound) to achieve robust aerial scene recognition performance. Multimodal Learning. Information in the real world usually comes as different modalities, with each modality being characterized by very distinct statistical properties, e.g., sound and image [5]. An expected way to improve relevant task performance is by integrating the information from different modalities. In past decades, amounts of works have developed promising methods on the related topics, such as reducing the audio noise by introducing visual lip information for speech recognition [6], improving the performance of facial sentiment recognition by resorting to the voice signal [7]. Recently, more attention is paid to the task of learning to analyze real-world multimodal scenarios [8] and events [9]. These works have con rmed the advantages of multimodal learning.

3. Dataset

To our knowledge, the audiovisual aerial scene recognition task has not been explored before. Salem et al. [10] established a dataset to explore the correlation between geotagged sound clips and overhead images. For further facilitating the research in this field, we construct a new dataset, with high-quality images and scene labels, named as ADVANCE6, which in summary contains 5075 pairs of aerial images and sounds, classifed into 13 classes.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

4. Methodology

In this paper, we focus on the audiovisual aerial scene recognition task, based on two modalities, i.e., image and audio. We propose to exploit the audio knowledge to better solve the aerial scene recognition task. In this section, we detail our proposed approaches for creating the bridge of knowledge transfer from sound event knowledge to the scene recognition in a multi-modality framework. We take the notations from Table 1, note that the data x follows the empirical distribution of our built dataset ADVANCE. For the multimodal learning task with deep networks, we adopt the model architecture that concatenates rep resentations from two deep convolutional networks on images and sound clips. So our main task, which is a supervised learning problem for aerial scene recognition, can be written

$$L_s = -\log\left[f_s(\boldsymbol{x}, N_{v+a})\right]_t \quad , \tag{1}$$

which is a cross-entropy loss with t-th class being the ground truth.

Table 1. Main notations.

$\boldsymbol{a}$,$\boldsymbol{v}$	audio input, visual input
$\boldsymbol{x}$,t	paired image and sound clip, $\boldsymbol{x} = \{\boldsymbol{v}, \boldsymbol{a}\}$, and the labeled ground truth t for aerial scene classification
N_*	network, which can be one of the network for extracting visual representation, the network for extracting audio representation, the pretrained (fixed) one for extracting audio representation, *i.e.*, $\{N_v, N_a, N_a^{(0)}\}$; also the one that concatenates N_v and N_a, *i.e.* N_{v+a}
f_*	classifier, which can be one of $\{f_s, f_e\}$, for aerial scene classification or sound event recognition; f_* takes the output of the network as input, and predicts the probability of the corresponding recognition task
$\boldsymbol{s}$,$\boldsymbol{e}$	probability distribution over aerial scene classes and sound event classes
s_k,s_t	k-th scene class' probability, and the t-th class being the ground truth
e_k	k-th sound event class' probability
$C(p, q)$	binary KL divergence: $\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$

Our proposed model architecture for addressing the multimodal scene recognition task, and present our idea of exploiting the audio knowledge following three directions: (1) avoid forgetting the audio knowledge during training by preserving the capacity of recognizing sound events; (2) construct a mutual representation that solves the main task and the sound event recognition task simultaneously, allowing the model to learn the underlying relation between sound events and scenes; (3) directly learn the relation between sound events and scenes. Our total objective function L is

$$L = L_s + \alpha L_\Omega \quad , \tag{2}$$

where controls the force of L , and L is one of the three approaches that are respectively.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

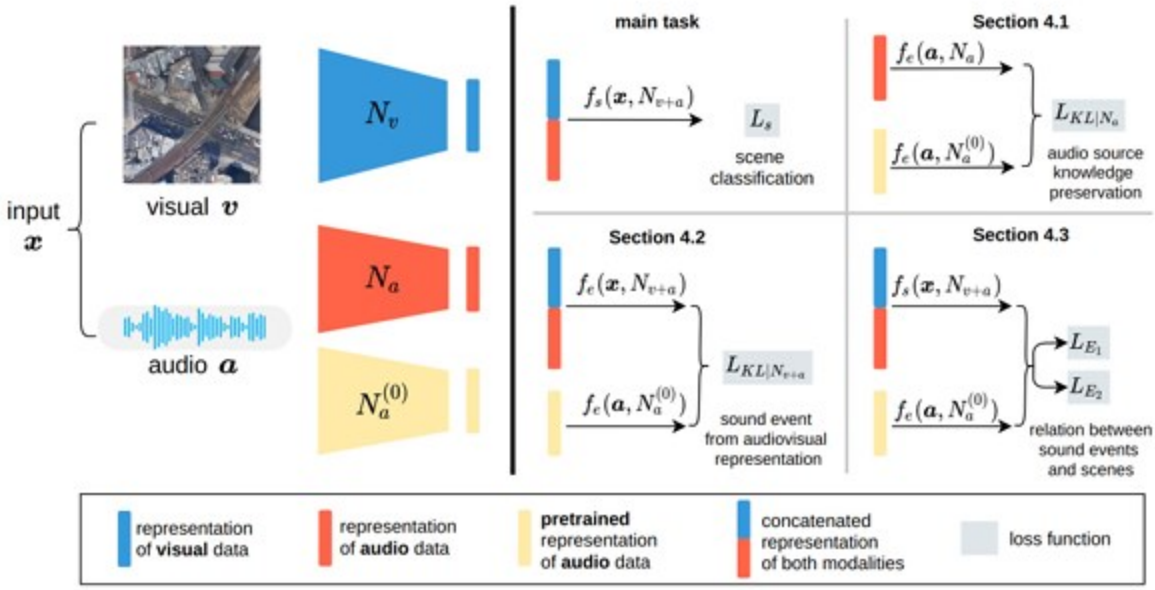

Fig 1. cross-task transfer approaches

In the above fig 1, We recall the notations: Nv, with trainable parameters, extracts visual representations, pretrained on the AID dataset; Na, also with trainable parameters, extracts audio representations, pretrained on the AudioSet dataset; N(0) a ,is the same as Na except parameters being xed; Nv+a simply applies both Nv and Na. The classi er at the last layer of the network is presented by ftask(input datanetwork), where the choice of task is s : scene classi catione : sound event recognition , input data is one of vax ,andthe set for network is NvNaN(0) a Nv+a .Onthe left of this gure, our model takes a paired data sample x of an image v and a sound clip a as input, and extracts representations from di erent combinations of modalities and models (shown in di erent colors); On the right, the top-left block introduces our main task of aerial scene recognition, and the rest three blocks present the three cross-transfer approaches.

5. Experiments

Our built ADVANCE dataset is employed for evaluation, where 70% image sound pairs are for training, 10% for validation, and 20% for testing. Note that, these three sub-sets do not share audiovisual pairs that are collected from the same coordinate. Before feeding the recognition model, we sub-sample the sound clips at 16 kHz. Then, following the short-term Fourier transform is computed using a window size of 1024 and a hop length of 400. The generated spectrogram is then projected into the log-mel scale to obtain an audio matrix in RT F, where the time T = 400 and the frequency F = 64. Finally, we nor malize each feature dimension to have zero mean and unit variance. The image data are all resized into 256 256, and horizontal fipping, color, and brightness jittering are used as data augmentation means.

Table 2: Aerial scene recognition results on the ADVANCE dataset

Approaches	L_{E_1}	$L_{E_1}+\beta L_{E_2}$	$L_{KL\|N_{v+a}}$	$L_{SQ\|N_{v+a}}$
Precision	43.37±0.59	54.23±1.14	3.08±0.14	2.95±0.07
Recall	49.26±0.36	52.57±0.72	9.69±0.43	9.28±0.17
F-score	42.50±0.42	48.65±0.85	4.46±0.20	4.24±0.07

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

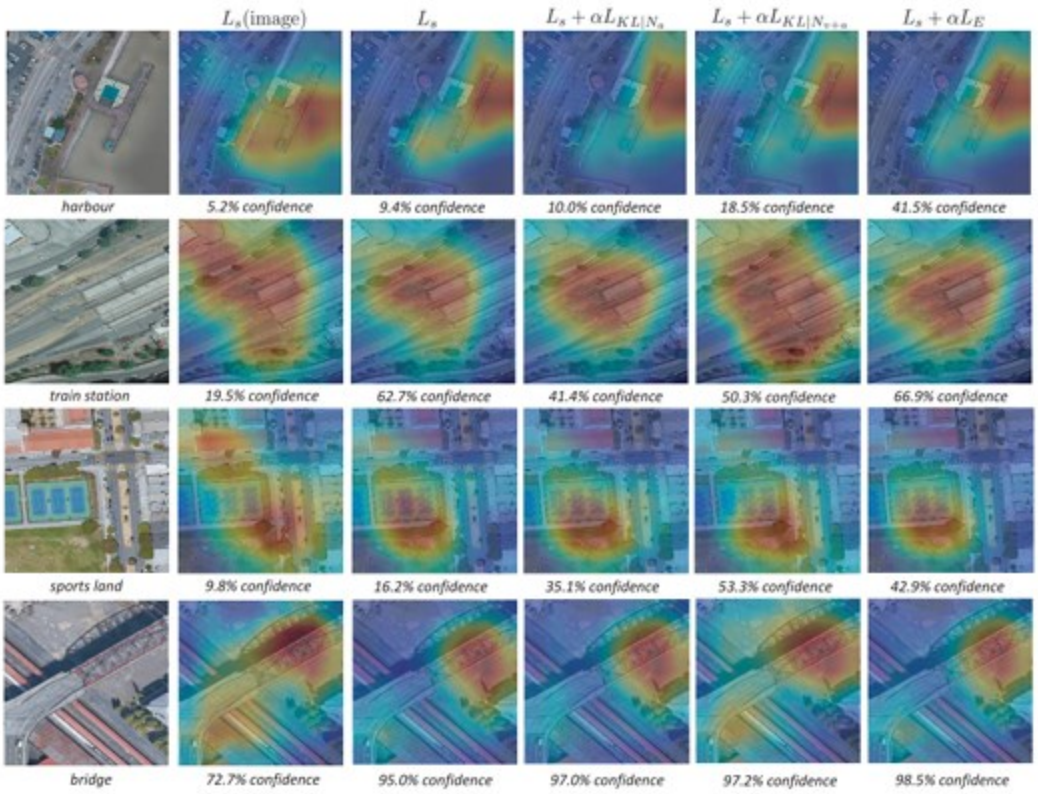

Fig 2 The class activation map

In the Fig 2, The class activation map generated by different approaches for different categories, as well as the corresponding predict probabilities of ground-truth category. Ls(image) means the learning objective of Ls is just performed with image data.

6. Conclusions

In this paper, we explore a novel multimodal aerial scene recognition task that considers both visual and audio data. We have constructed a dataset consists of labeled paired audiovisual worldwide samples for facilitating the research on this topic. We propose to transfer the sound event knowledge to the scene recognition task for the reasons that the sound events are related to the scenes and that this underlying relation is not well exploited. Amounts of experimental results show the effectiveness of three proposed transfer approaches, conforming the benefits of exploiting the audio knowledge for the aerial scene recognition

References

1. Assael, Y.M., Shillingford, B., Whiteson, S., De Freitas, N.: Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599 (2016)

2. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in neural information processing systems. pp. 892900 (2016)

3. Baltrusaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence 41(2), 423443 (2018)

4. Castelluccio, M., Poggi, G., Sansone, C., Verdoliva, L.: Land use classi ca tion in remote sensing images by convolutional neural networks. arXiv preprint arXiv:1508.00092 (2015)

ISSN 2347–3657

Volume 13, Issue 1, 2025

5. Cheng, G., Yang, C., Yao, X., Guo, L., Han, J.: When deep learning meets met ric learning: Remote sensing image scene classi cation via learning discriminative cnns. IEEE transactions on geoscience and remote sensing 56(5), 28112821 (2018)

6. Ehrlich, M., Shields, T.J., Almaev, T., Amer, M.R.: Facial attributes classi cation using multi-task representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 4755 (2016)

7. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE in ternational conference on computer vision. pp. 26502658 (2015)

8. Gan, C., Zhao, H., Chen, P., Cox, D., Torralba, A.: Self-supervised moving vehicle tracking with stereo sound. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 70537062 (2019)

9. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776780. IEEE (2017)

10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015), http: //arxiv.org/abs/1503.02531