Estimating Visual Information From Audio Through Manifold Learning

Fabrizio Pedersoli[1], Dryden Wiebe[1], Amin Banitalebi-Dehkordi[2], Yong Zhang[2], George Tzanetakis[3], and Kwang M. Yi[1]

[1] University of British Columbia {fpeder,drydenw,kmyi}@cs.ubc.ca
[2] Huawei Technologies Canada Co., Ltd
{amin.banitalebi,yong.zhang3}@huawei.com
[3] University of Victoria gtzan@uvic.ca

Abstract. We propose a new framework for extracting visual information about a scene only using audio signals. Opposed to the commonly-used end-to-end learning paradigm, we show that learning a per-modal manifold is critical when linking vision with audio. In more detail, we first train a Vector-Quantized Variational Auto-Encoder to learn the data manifold of the particular visual modality we are interested in. Second, we train an Audio Transformation network to map multi-channel audio signals to the latent representation of the corresponding visual sample. We show empirically that this two-stage setup is critical for our method to produce meaningful images from audio using publicly available datasets. In particular, we consider the prediction of the following visual modalities from audio: depth and semantic segmentation. Code is available at: https://github.com/ubc-vision/audio_manifold.

Keywords: Manifold Learning; Audio Transformation Network

1 Introduction

Line-of-Sight is a fundamental requirement of many computer vision algorithms and applications [16, 17, 5, 13]. Computer vision tasks such as object localization [23, 2, 6, 27] and autonomous navigation [18, 19, 21] are particularly affected by the line-of-sight problem. Moreover, when it comes to indoor vision applications [9, 22, 1, 15, 20, 3], meeting the line-of-sight requirement becomes even more challenging. On the other hand, potential *audio-based* methods do not require a direct line-of-sight. More recent research works on audio have shown that even more involved audio-based methods can be designed, such as methods that estimate rough visual characteristics of a scene [7, 26, 10] in terms of depth and semantic segmentation.

In this paper we propose a method for extracting visual information from audio which is not limited to sounding objects. Specifically, we propose a method capable of predicting: depth, and semantic segmentation of a scene. The proposed method is based on VQ-VAE [25] for learning the manifold of the visual and data, and an audio transformation network for mapping the input sound to

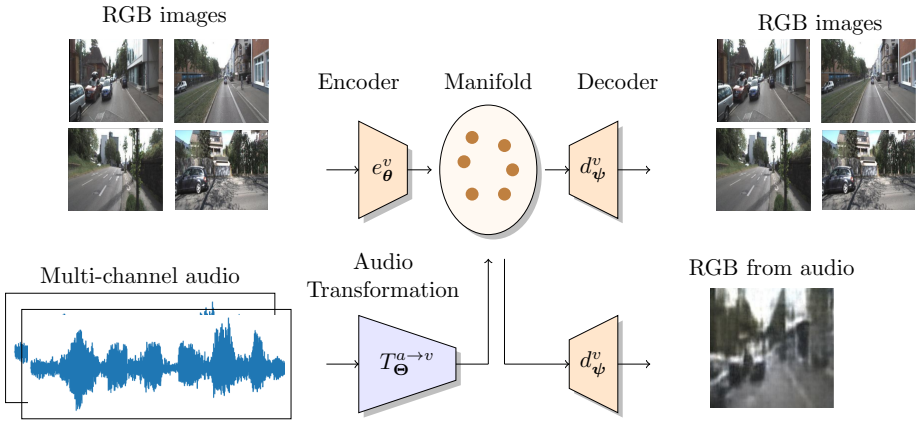


Fig. 1. Overview of our method. At first stage the data manifold is learnt for each visual modality by using a VQ-VAE. After that, the audio transformation network maps an audio sample to the closet visual sample in manifold space.

most similar manifold sample. Critically, our method operates in two stages, instead of an end-to-end setup.

2 Method

In this paper we propose a two-stage method for extracting visual information from audio, as shown in Fig. 1. The first stage consists of training a VQ-VAE on a particular visual modality, which can be but not limited to: depth maps, and semantic segmentation maps. The second stage consists of training a domain transformation network, which we refer to as Audio-Transformation network (AT-net), to map audio signals to the visual modality.

We propose this two-stage approach because it provides advantages compared to single-stage (end-to-end) approaches. End-to-end models have limitations when used to extract visual information from audio signals. We have empirically observed that such models converge to an average representation of the data-set that has low quality and lacks visual detail. A two-stage approach can overcome these issues by *learning a transformation to the learned manifold.*[1]. The key idea of learning a transformation at the manifold level is that it potentially allows to reconstruct the overall structure, as well as the details, of a visual modality because the decoder is specific to that modality. This transformation can be effectively learned because manifold data lies on a higher dimensional space which can enforce sparsity. We use a VQ-VAE framework for learning the manifold of a given visual modality, in our case depth, and segmentation maps.

[1] We refer to manifold as the space defined by the encoded representation of some data type (visual data in our case).

Audio Transformation. The purpose of the AT-net is to encode an audio sample to the closest manifold sample in the visual domain. The AT-net consists of three components: audio encoder, domain transformation MLP, and visual manifold decoder. The AT-net is simply trained with the L_2 loss at the manifold level. In more detail, given a training (*i.e.* audio, visual) pair, first the visual sample is encoded onto the visual manifold using the VQ-VAE encoder. Then, the encoding on the manifold for this data sample acts as the target for training the AT-net. We train the AT-net on the continuous (not quantized) latent representation of the visual data.

3 Results

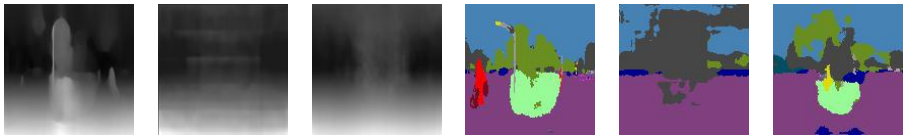
We evaluate our method on the publicly available datasets: the "Omni Audiotory Perception" (OAP) [26] and "Multimodal Audio Visual Dectection" (MAVD) [24]. Depth maps and semantic segmentation maps computed from RGB frames using the methodology in [26] which uses `Monodepth2` [11] and `Deeplabv3` [4].

Network specifics. For the VQ-VAE we adopt a similar configuration as [25]. For the AT-net, the audio encoder is a Resnet18 [12]. The domain transformation MLP consists of three dense layers, and the visual decoder consists of series of strided transposed convolutional layers which upsamples the output to the desired manifold size. We use the Adam optimizer [14] with learning rate of 1×10^{-4} for both the VQ-VAE and AT-net training.

Evaluation. For depth estimation, we use error metrics as defined in Eigen et al. [8]. Specifically, we use: absolute relative distance (ABS_{rel}), squared relative distance (SQR_{rel}), RMSE linear ($RMSE_{\text{lin}}$), and RMSE logarithmic ($RMSE_{\log}$). In addition, we also evaluate AUC_{crr} of relative correct predictions below thresholds $\tau \in \{0, 0.01, \dots 0.3\}$. For the semantic segmentation evaluation we report performance results in terms of mean Intersection over Union (mIoU), alongside the individual IoU of each one of ground truth classes. Classes that scores an IoU of below 0.1% for all the considered methods are removed from the evaluation.

Comparison. We compare our method with the method proposed in Vasudevan et al. [26], using both 2ch and 8ch audio input. We include multimodal and unimodal configurations, but do not include the super-resolution task. We also compare against ECHO2DEPTH [10] (audio input only) for the depth estimation task. For our method we report performance with respect to the spatial resolution of the data manifold being 8×8.

Omni auditory perception dataset. In Tab. 1 we report the performance comparison for the task of depth maps estimation. In Tab. 2 we report the performance comparison for the task of semantic segmentation maps estimation. For this experiment we consider only significant classes—classes which are above the threshold of 0.1% IOU for all methods. We show qualitative results for this dataset in Fig. 2. Our method performs best.



(a) GT (b) [26] (c) Proposed (d) GT (e) [26] (f) Proposed

Fig. 2. Qualitative results for the scene 149 in the Omni auditory perception dataset.

End-to-End vs Two Stage. We note that in Tab. 1 we further compare against a version of our method that is trained end-to-end. This variant performs *significantly* worse, demonstrating the importance of two-stage training.

Table 1. OAP depth map estimation results.

Method	$ABS_{\text{rel}} \downarrow$	$SQR_{\text{rel}} \downarrow$	$RMSE_{\text{lin}} \downarrow$	$RMSE_{\log} \downarrow$	$AUC_{\text{err}} \uparrow$
[10] ECHO2DEPTH	.731	6.13	7.00	1.19	.077
[26] 2ch	.478	2.78	5.12	.527	.068
[26] 8ch	.674	4.51	5.29	.602	.047
[26] 2ch +seg	.304	1.32	4.68	.446	.089
[26] 8ch +seg	.372	1.45	4.59	.471	.076
Proposed	**.241**	**1.01**	**4.12**	**.361**	**.118**
Proposed (End-to-End)	.442	2.31	4.92	.526	.072

Table 2. OAP semantic segmentation results reported in IoU [%].

Method	*AVG*	Road	Side.	Build.	Fence	Veget.	Terr.	Sky	Car
[26] 2ch	18.83	**81.90**	0.58	26.63	0.09	10.88	0.03	24.94	5.59
[26] 8ch	17.66	74.94	0.47	27.35	0.29	14.05	0.06	18.69	5.48
[26] 2ch + depth	18.69	79.68	0.19	**28.88**	0.25	7.85	0.30	26.70	5.74
[26] 8ch +depth	18.60	75.27	0.58	24.17	0.29	**15.39**	0.07	26.96	**6.11**
Proposed	**20.73**	74.67	**4.34**	11.79	**1.53**	15.11	**2.59**	**53.31**	2.49

Table 3. MAVD depth map estimation results.

Method	$ABS_{\text{rel}} \downarrow$	$SQR_{\text{rel}} \downarrow$	$RMSE_{\text{lin}} \downarrow$	$RMSE_{\log} \downarrow$	$AUC_{\text{err}} \uparrow$
[10] ECHO2DEPTH	.218	.561	2.34	.324	.14
[26] 2ch	.232	.740	2.69	.324	.37
[26] 8ch	.210	.669	2.54	.301	.46
[26] 2ch +seg	.223	.676	2.55	.313	.46
[26] 8ch +seg	.205	.634	2.49	.298	.15
Proposed	**.126**	**.290**	**1.51**	**.180**	**.191**

Table 4. MAVD semantic segmentation results reported in IoU [%].

Method	*AVG*	Road	Side.	Build.	Fence	Veget.	Terr.	Sky	Car
[26] 2ch	36.66	76.42	25.59	48.56	16.16	58.88	14.86	26.47	26.40
[26] 8ch	40.05	78.75	28.96	52.68	18.65	63.56	17.59	29.81	30.47
[26] 2ch +depth	41.68	76.12	29.10	49.18	26.86	59.92	25.68	39.70	26.94
[26] 8ch +depth	47.36	80.15	34.83	55.31	30.71	64.46	31.61	47.74	34.11
Proposed	**61.02**	**90.42**	**55.43**	**77.61**	**38.23**	**84.56**	**39.68**	**47.64**	**54.62**

Multimodal audio visual detection dataset. In Tab. 3 we report the performance comparison for the task of depth maps estimation. In Tab. 4 we report the performance comparison for the task of semantic segmentation maps estimation.

4 Conclusion

We presented a novel framework for estimating visual information from audio. The fundamental idea behind our method is learning the transformation between the audio and the visual domains at a visual manifold level from a VQ-VAE rather than using an end-to-end approach. We showed that this proposal results in superior performance compared to previous approaches to this problem.

References

1. Baek, F., Ha, I., Kim, H.: Augmented reality system for facility management using image-based indoor localization. Automation in construction **99**, 18–26 (2019)
2. Caicedo, J.C., Lazebnik, S.: Active object localization with deep reinforcement learning. In: Proceedings of the IEEE international conference on computer vision. pp. 2488–2496 (2015)
3. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Soundspaces: Audio-visual navigation in 3d environments. In: ECCV (2020)
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
6. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3133–3142 (2020)
7. Christensen, J.H., Hornauer, S., Stella, X.Y.: Batvision: Learning to see 3d spatial layout with two ears. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 1581–1587. IEEE (2020)
8. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
9. Fusco, G., Coughlan, J.M.: Indoor localization for visually impaired travelers using computer vision on a smartphone. In: Proceedings of the 17th International Web for All Conference. pp. 1–11 (2020)
10. Gao, R., et al.: Visualechoes: Spatial image representation learning through echolocation. In: ECCV (2020)
11. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3828–3838 (2019)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Janai, J., Güney, F., Behl, A., Geiger, A., et al.: Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision **12**(1–3), 1–308 (2020)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Kumar, S., Gil, S., Katabi, D., Rus, D.: Accurate indoor localization with zero start-up cost. In: Proceedings of the 20th annual international conference on Mobile computing and networking. pp. 483–494 (2014)
16. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)

18. Royer, E., Bom, J., Dhome, M., Thuilot, B., Lhuillier, M., Marmoiton, F.: Outdoor autonomous navigation using monocular vision. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1253–1258. IEEE (2005)
19. Royer, E., Lhuillier, M., Dhome, M., Lavest, J.M.: Monocular vision for mobile robot localization and autonomous navigation. International Journal of Computer Vision **74**(3), 237–260 (2007)
20. Sadeghi, H., Valaee, S., Shirani, S.: A weighted knn epipolar geometry-based approach for vision-based indoor localization using smartphone cameras. In: 2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM). pp. 37–40. IEEE (2014)
21. Scaramuzza, D., Achtelik, M.C., Doitsidis, L., Friedrich, F., Kosmatopoulos, E., Martinelli, A., Achtelik, M.W., Chli, M., Chatzichristofis, S., Kneip, L., et al.: Vision-controlled micro flying robots: from system design to autonomous navigation and mapping in gps-denied environments. IEEE Robotics & Automation Magazine **21**(3), 26–40 (2014)
22. Sun, M., Zhang, L., Liu, Y., Miao, X., Ding, X.: See-your-room: indoor localization with camera vision. In: Proceedings of the ACM turing celebration conference-China. pp. 1–5 (2019)
23. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 648–656 (2015)
24. Valverde, F.R., Hurtado, J.V., Valada, A.: There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11612–11621 (2021)
25. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
26. Vasudevan, A.B., Dai, D., Gool, L.V.: Semantic object prediction and spatial sound super-resolution with binaural sounds. In: European conference on computer vision. pp. 638–655. Springer (2020)
27. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1325–1334 (2018)