ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094965

FLOWGRAD: USING MOTION FOR VISUAL SOUND SOURCE LOCALIZATION

Rajsuryan Singh [1], *Pablo Zinemanas*[1], *Xavier Serra*[1], *Juan Pablo Bello*[2], *Magdalena Fuentes*[2,3]

[1] MTG, Universitat Pompeu Fabra, Barcelona, Spain
[2] MARL, New York University, New York, USA
[3] IDM, New York University, New York, USA

ABSTRACT

Most recent work in visual sound source localization relies on semantic audio-visual representations learned in a self-supervised manner and, by design, excludes temporal information present in videos. While it proves to be effective for widely used benchmark datasets, the method falls short for challenging scenarios like urban traffic. This work introduces temporal context into the state-of-the-art methods for sound source localization in urban scenes using optical flow to encode motion information. An analysis of the strengths and weaknesses of our methods helps us better understand the problem of visual sound source localization and sheds light on open challenges for audio-visual scene understanding. The code and pretrained models are publicly available at `https://github.com/rrrajjjj/flowgrad`

Index Terms— Sound source localization, audio-visual urban scene understanding, explainability.

1. INTRODUCTION

Vision and audition are complementary sources of information and their effective integration, i.e. the ability to localize sounds and connect them to visual objects, enables a rich understanding of a dynamic environment. Early attempts at modeling audio-visual perception exploited the synchrony between audio and visual events, e.g. lip movements aligned to speech, with probabilistic models [1, 2], and canonical correlation analysis [3]. With recent advances in deep learning, especially in computer vision, the field has pivoted to deep-neural-network-based methods. A notable difference between the two approaches is the shift from using the temporal correlation between audio and video to the semantic similarity between them as the primary source of information for localization. This has happened to the extent that most state-of-the-art methods, except for a very few examples [4, 5], completely disregard the temporal context available in videos [6, 7, 8, 9, 10]. These methods focus on learning semantic auditory and visual representations in a self-supervised manner that enables sound source localization (SSL) via the similarity between audio and visual embeddings. This approach has been effective for the widely used benchmark datasets[6, 7, 8, 9, 10], however recent work by Wu et al. has raised questions about the generalizability of these methods beyond these datasets [11]. They further point out the strong biases present in these benchmarks and demonstrate that the methods developed on these datasets fail to generalize to urban scenes.

Urban scene understanding has many potential applications in various sectors, including assistive devices for the hard-of-hearing, traffic monitoring, and autonomous driving. However, visual sound source localization (VSSL) in urban scenes is a challenging task, and state-of-the-art methods are not sufficient [11]. Benchmark datasets for VSSL, such as VGG-SS and Flickr, typically have only one sound source per image, whereas urban scenes often have multiple agents that may or may not be producing sounds. To address this issue, we investigate the use of temporal context in our approach.

We test our methods on Urbansas dataset [12], which is an audio-visual dataset for detecting sound events in urban environments. We only use Urbansas for evaluation because other VSSL benchmarks have a bias towards static sound sources in the center of the image, making the inclusion of motion information unnecessary, and RCGrad has already been evaluated on other VSSL benchmarks in [11]. Our baseline model for Urbansas is RCGrad [11], which is the state-of-the-art.

We propose the use of optical flow as a means to incorporate temporal information and we explore hard-coded as well as learning-based algorithms to combine it with RCGrad. First, we use optical flow as a heuristic to filter stationary objects from the predictions of RCGrad and observe a significant improvement in localization performance, especially in curbing false positives. Further, we add optical flow as a feature to the neural network in two ways: i) we add optical flow as an additional channel into the vision encoder, and ii) we train a separate optical flow encoder within the RCGrad framework.

2. METHOD

2.1. RCGrad

RCGrad [11] uses resnet-18 as the audio as well as the vision encoder. The vision encoder is pretrained on Imagenet while the audio encoder is randomly initialized. The model is then trained with a contrastive loss on VGG-Sound [13].

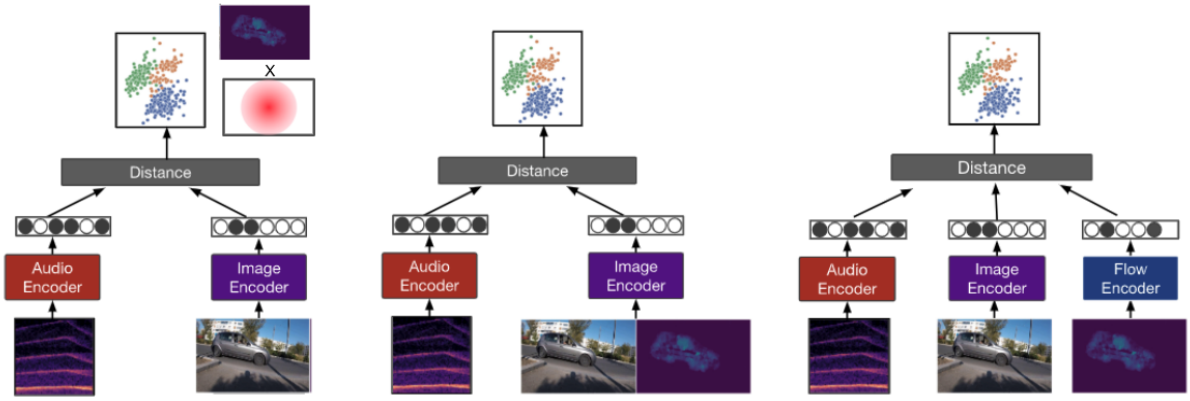

Fig. 1. Left: FlowGrad-H, element-wise multiplication of the RCGrad predictions with optical flow. Center: FlowGrad-IC, optical flow as an extra channel in the image encoder. Right: FlowGrad-EN, optical flow added through a third flow encoder.

Each training example is a randomly selected image from a 5-second video along with the corresponding audio. As is standard in the literature, the model uses separate audio and vision encoders optimized with audio-visual correspondence as the training objective. Localization is done using a modified version of Grad-CAM [14] wherein instead of back-propagating class labels, the audio embedding is back-propagated through the vision subnetwork to generate localization maps.

2.2. FlowGrad: Incorporating temporal context

Optical Flow as a heuristic. In the context of urban scene understanding, one major limitation of RCGrad is the attribution of sounds to parked vehicles. Since the representations are purely semantic and there is no temporal context, the model cannot distinguish between stationary and moving vehicles. As a result, parked vehicles often end up as false positives diminishing the performance. Optical flow, on the other hand, only has motion information. Anything that moves, be it vehicles, pedestrians, or tree leaves, have high values of activation. Hence, optical flow and RCGrad have complementary strengths that can be leveraged by taking an intersection of objects that have high activations for both. We execute this idea by simply doing an element-wise multiplication of the RCGrad predictions with the optical flow. This suppresses objects that are either not moving or that are moving but are not sounding, leaving us with sounding vehicles. We call this model *FlowGrad-H*, depicted on the left of Figure 1.

Optical Flow as an image channel. As effective as heuristics can prove to be, they are often rigid, brittle, and prone to a lack of generalizability. In an attempt to move away from the naive use of optical flow as a filter and towards using it to imbue the representations with temporality, we include it as an image channel. Here, the model can, at least in principle, take the motion information into account while making predictions, instead of motion being used as a filter post-hoc. The relationship between motion and sounds can hence be learned. To do so, we extended RCGrad to take in 4 channels (RGB and optical flow) as the input to the image encoder (see center of Figure 1). We initialize the model with the pre-trained version of RC-Grad, and for the weights of the optical flow channel we used the average of the weights of the RGB channels. We train the model on the unlabeled portion of Urbansas using contrastive loss. Following [11], during training we fit the model with a frame along with a 5-second audio clip around it, plus the corresponding optical flow calculated between consecutive frames at 8fps as an additional channel. This model is *FlowGrad-IC*.

Optical Flow encoder. In the above-mentioned method, the 4 channels of the image encoder are pooled in very early layers of the network. This may result in shallow integration of motion information. Moreover, since the model was initialized with weights pre-trained on audio and images, simply discarding the additional optical flow channel provides a trivial solution for minimizing the loss. To avoid this, we added a separate flow encoder with the same Resnet-18 architecture as the image and the audio encoders to RCGrad (see right of Figure 1). We initialized the weights as the average of RGB channels of the vision encoder, and modified the training loss to be the sum of all pairwise losses (audio-image, image-flow, and audio-flow). The localization is then done by backpropagating the audio embeddings through the image as well as the flow encoder to generate two localization maps. These maps are then multiplied element-wise to give the final localization map. We call this model *FlowGrad-EN*.

3. EXPERIMENTAL DESIGN

The Urbansas dataset is an audio-visual dataset developed for studying the detection and localization of sounding vehicles in the wild [12]. The dataset consists of labeled and unlabeled videos of urban traffic with stereo audio, to a total of

15 hours of video out of which 3 hours have been manually annotated, with both audio events and video bounding-box annotations, for sound event detection and source localization. We train on the unlabelled videos following the training protocol from [11] and we evaluated our models on the annotated portion of the dataset. For evaluation, we only consider frames that have both audio and video annotations and where the sounding vehicle is visible and identifiable giving us 5704 annotated image-audio pairs.

Baselines. We employ three baselines: i) *RCGrad* [11], which is the current state-of-the-art localization method in Urbansas; and ii) a vision-only object recognition topline method with temporal and class filtering (*vision-only+CF+TF*), which is a strong reference with temporal integration and information about the classes present but not sound; and iii) an *optical flow baseline*, which helps us understand how much of the data can be explained by motion only. We have replicated the results of RCGrad [11] using the pre-trained models from the official repository.

For the vision-only+CF+TF baseline we use a pre-trained YOLOR object detection model [15]. This model has been trained to predict bounding boxes around objects on the MS-COCO dataset [16], which is a large-scale dataset with just under a million annotated objects, nearly 10% of which correspond to vehicles. We use the pretrained *yolor_p6* model weights for inference, and we filter the results to the four vehicle classes present in Urbansas - car, motorcycle, bus, and truck. Further, we apply motion-based filtering. For each pair of consecutive frames (f and f+1), if a bounding box in f has an IoU greater than 0.95 with one in f+1, both the bounding boxes are discarded. This ensures that stationary objects are filtered out in the final predictions. For the optical flow baseline, we use the normalized optical flow directly as predictions, without any semantic filtering. This means that we consider that anything that is moving is producing sound. This method serves to demonstrate the correspondence, or a lack thereof, between moving and sounding objects.

Optical Flow. The optical flow is calculated using the Gunnar Farneback algorithm [17]. Images are sampled at 8 frames-per-second, converted to grayscale, and dense optical flow is estimated between the current and the next frame using the OpenCV implementation of the algorithm.

Metrics. Following [11], the localization maps are min-max normalized and we use consensus intersection over union (cIoU) and area under the curve (AUC) as performance metrics as in the literature [7, 8, 10, 11]. We binarize localization maps with a threshold of 0.5 to calculate the cIoU. The AUC is calculated for cIoUs at different thersholds.

4. RESULTS AND DISCUSSION

Results are presented in Table 1. The first observation is that the vision topline performs considerably better than the other methods.This is because YOLOR is a supervised object detection model that has information about classes and predicts precise bounding boxes around vehicles whereas the other models produce coarser and less precise heatmaps (see Figure 2), scoring lower in the IoU. The ground truth annotations are also bounding boxes generated using an object detection model [12] and this congruence between the ground truth and the predictions further inflates the IoU. This combined with motion-based filtering gives us a very strong supervised reference to pit our self-supervised models against.

model	*IoU* ($\tau = 0.5$)	*AUC*
Vision-only+CF+TF (topline)	**0.68**	**0.51**
Optical flow only (baseline)	0.33	0.23
RCGrad [11] (sota)	0.16	0.13
FlowGrad-H	**0.50**	**0.30**
FlowGrad-IC	0.26	0.18
FlowGrad-EN	0.37	0.23

Table 1. IoU and AUC results for the different models.

All models that use motion information outperform RCGrad, since predictions of stationary vehicles are eliminated overcoming RCGrad's major limitation. Using thresholded optical flow directly as localization maps outperforms vanilla RCGrad,which suggests that there is a high correlation between motion and sound in Urbansas. As can be seen in the first two rows of Figure 2, the optical flow baseline produces more precise localization heatmaps around moving vehicles, ignoring those that are parked, while the predictions of RCGrad focus on any visible car, including parked vehicles which are silent and hence are false positives.

Looking at the results in Table 1, we conclude that motion alone is not enough to explain sounding objects in urban settings, as the integration of motion; sound and semantics leads to the best performing unsupervised systems (FlowGrad-H and FlowGrad-EN). The best way of combining optical flow with the deep learning model seems to be as a post-processing heuristic (FlowGrad-H), followed by adding a flow encoder (FlowGrad-EN), and lastly, adding optical flow as an extra channel to the image encoder (FlowGrad-IC). A heuristic performing better than learning based methods is counterintuitive but it's been shown that if the dataset has strong biases, even trivial heuristics like a big-enough bounding box located in the middle of the image perform similar (and sometimes outperform) state-of-the-art methods [11] suggesting a strong sound-motion correspondence bias in Urbansas. With the integration of flow, the model is able to distinguish the parked vehicle (see Figure 2), and the localization maps are for the most part less diffused and this stringency is likely to contribute to the increased IoU numbers due to a decrease in the overall area of union. By the same token, the size of the predicted masks may also, at least in part, explain why FlowGrad-EN does not perform as well as the naive use of

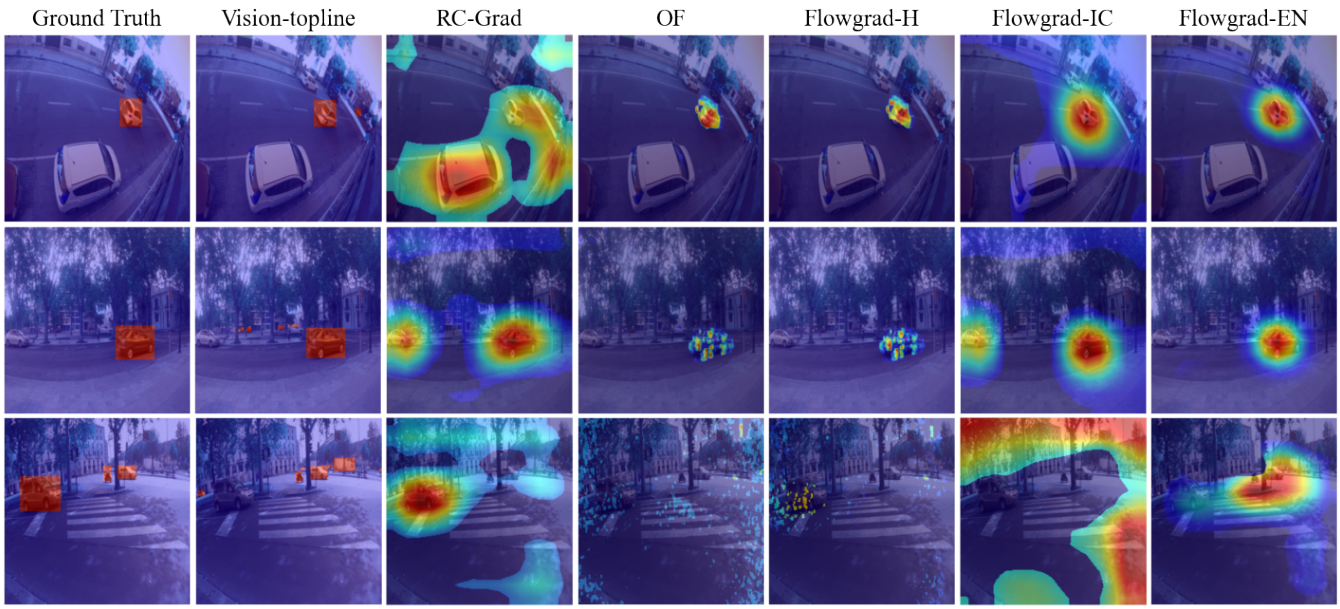

Fig. 2. Predictions of the baselines and the proposed models on selected examples. Optical flow proves to be effective in the first two examples but sounding vehicles parked at traffic signals are a limitation of the method as shown in the bottom row.

optical flow as a heuristic (FlowGrad-H). Optical flow generates very precise masks around objects minimizing the area of union and hence increasing the IoU while this method still produces diffused localization maps. This opens up the question once more, as discussed in [11], of whether bounding boxes combined with IoU are in fact a good way of evaluating localization models.

After examining the performance of our models on different scenes in Urbansas, we found that incorporating optical flow has its limitations and may even decrease performance in certain scenarios, such as those with shaky cameras. In cases where sound and motion do not correspond well, such as when parked vehicles produce sound, but are ignored by motion models, it is difficult for semantics or motion alone to describe the scene accurately. Therefore, we may need to use complementary information such as spatial sound or reasoning. Another approach we could explore to improve performance in these scenarios is to extend the temporal context window used in optical flow calculations. Currently, we use a short context window of 0.125 seconds, but increasing it to 5 seconds may provide the necessary information to attribute sounds to temporarily stationary vehicles. One way to do this is to aggregate optical flow over a 5-second window, similar to action recognition strategies, and use the resulting stack of optical flow as a feature. Alternatively, we could average the optical flow across the time window, as done in [4].

Trees, pedestrians, and other moving objects are also exceptions to the assumption. Moving tree leaves can often have high optical flow, but they have no contribution to the sounds whatsoever. However, in contrast to the previous example, using optical flow along with semantics and sound (as in FlowGrad) is a simple fix to this issue as the RCGrad predictions generally have very low activations for trees if the sounding object is an engine. The case with pedestrians is not as straightforward as it is for trees. They have characteristic sounds associated with them that are clearly audible, especially if they are close to the microphone. Most models we use for sound source localization (and certainly the ones investigated in this work) are class-agnostic and are trained in a self-supervised manner without any class labels. So RCGrad localizes pedestrians as sound sources as we have observed in some cases. Pedestrians also have high optical flow and hence cannot be filtered out by either method or a combination thereof. Since pedestrians are not labeled in the Urbansas dataset, they are evaluated as false positives. However, we think this is a limitation of the dataset rather than the method, and we will extend Urbansas' annotations in future work.

5. CONCLUSIONS AND FUTURE WORK

In this paper we investigated the correspondence of motion and sound to help visual sound source localization methods. Our proposed method (FlowGrad) and their variations, greatly outperform previous state of the art models for urban sound source localization, showing the importance of motion and temporal context for analyzing urban scenes. For future work, we plan to improve the quality of the optical flow estimation to make it more robust to lighting and camera instability, and explore the use of multiple frames from as input to the vision encoder as in [18].

6. REFERENCES

[1] John Hershey and Javier Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," *Advances in neural information processing systems*, vol. 12, 1999.

[2] John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola, "Learning joint statistical models for audio-visual fusion and segregation," *Advances in neural information processing systems*, vol. 13, 2000.

[3] Einat Kidron, Yoav Y Schechner, and Michael Elad, "Pixels that sound," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, vol. 1, pp. 88–95.

[4] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman, "Self-supervised learning of audio-visual objects from video," in *European Conference on Computer Vision*. Springer, 2020, pp. 208–224.

[5] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba, "The sound of motions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1735–1744.

[6] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[7] Relja Arandjelovic and Andrew Zisserman, "Objects that sound," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 435–451.

[8] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon, "Learning to localize sound source in visual scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4358–4366.

[9] Takashi Oya, Shohei Iwase, Ryota Natsume, Takahiro Itazuri, Shugo Yamaguchi, and Shigeo Morishima, "Do we need sound for sound source localization?," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[10] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman, "Localizing visual sounds the hard way," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16867–16876.

[11] Ho-Hsiang Wu, Magdalena Fuentes, Prem Seetharaman, and Juan Pablo Bello, "How to Listen? Rethinking Visual Sound Localization," in *Proc. Interspeech 2022*, 2022, pp. 876–880.

[12] Magdalena Fuentes, Bea Steers, Pablo Zinemanas, Martin Rocamora, Luca Bondi, Julia Wilkins, Qianyi Shi, Yao Hou, Samarjit Das, Xavier Serra, and Juan Bello, "Urban sound & sight: Dataset and benchmark for audio-visual urban scene understanding," *ICASSP 2022*, 2022.

[13] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[15] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao, "You only learn one representation: Unified network for multiple tasks," *arXiv preprint arXiv:2105.04206*, 2021.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[17] Gunnar Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[18] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon, "Less can be more: Sound source localization with a classification model," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3308–3317.