2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

AVOT: Audio-Visual Object Tracking of Multiple Objects for Robotics

Justin Wilson[1] and Ming C. Lin[1,2]

***Abstract*— Existing state-of-the-art object tracking can run into challenges when objects collide, occlude, or come close to one another. These visually based trackers may also fail to differentiate between objects with the same appearance but different materials. Existing methods may stop tracking or incorrectly start tracking another object. These failures are uneasy for trackers to recover from since they often use results from previous frames. By using audio of the impact sounds from object collisions, rolling, etc., our audio-visual object tracking (AVOT) neural network can reduce tracking error and drift. We train AVOT end to end and use audio-visual inputs over all frames. Our audio-based technique may be used in conjunction with other neural networks to augment visually based object detection and tracking methods. We evaluate its runtime frames-per-second (FPS) performance and intersection over union (IoU) performance against OpenCV object tracking implementations and a deep learning method. Our experiments, using the synthetic Sound-20K audio-visual dataset, demonstrate that AVOT outperforms single-modality deep learning methods, when there is audio from object collisions. A proposed scheduler network to switch between AVOT and other methods based on audio onset maximizes accuracy and performance over all frames in multimodal object tracking.**

I. INTRODUCTION

Deep learning has enabled state-of-the-art techniques for image classification and object detection in images and video [39], [49], [50]. Object tracking classifies bounding boxes for each object in a video over time. These methods are useful for applications in autonomous driving [17], mobile robotics [52], person tracking [11], speaker recognition [48], [55], and 3D reconstruction [47]. For more granularity beyond bounding boxes, object segmentation provides pixel-level annotations [46], [58]. These existing object trackers achieve real-time performance and continue to improve on accuracy and the number of classes that they can detect.

However, occlusion, similar object categories, and smaller object sizes remain a challenge for visually based trackers [39]. Auditory cues can assist in these exacting areas, especially when similar and/or smaller objects are of a different material [4]. In this paper, we propose an audio-visual object tracker (AVOT) that augments visual only trackers with fused audio in a jointly trained end-to-end model. It is evaluated using synthetic Sound-20K dataset [63], consisting of tabletop sized objects of different geometry and materials. The data contains videos with multiple objects of various shapes (e.g. bottle, knife, etc.) and materials (e.g. steel, wood, etc.) colliding in a virtual scene. Colliding includes objects colliding within the scene, with each other, rolling, etc. We use videos with one, two, or three colliding objects.

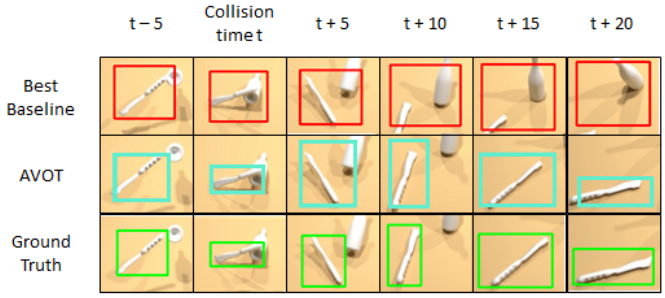

Fig. 1. An example failure case improved by our audio-visual object tracker. (Top row) best baseline, CSRT in this case, incorrectly latches to the wrong object after collision. (Middle row) our AVOT method continues to correctly track the object post-collision. (Bottom row) ground truth annotated by the experimenter. For clarity, we show the bounding box for only one of the objects being tracked, although the methods track both objects. Please see the Supplementary Video for more demonstration.

Other than speaker recognition, this is the first use of an audio-visual neural network for tracking tabletop sized objects and enhancing visual object trackers. The key contributions of this work include:

- An end-to-end, jointly trained audio-visual object tracker (AVOT) to enhance visual object tracking;
- Ground truth bounding box annotations for Sound-20K audio-visual dataset with 1, 2, and 3 object scenes;
- Scheduler for object detection re-initialization based on audio onset detection when using multimodal tracking.

Fusing audio with visual data, AVOT achieves 77.7% IoU post-collision tracking accuracy compared to 68.6% IoU using deep-learning visual tracking, SSD– [39], and 38.4% using CSRT [40] for virtual scenes with multiple objects based on our annotated Sound-20K dataset of 19 tabletop sized object classes of varying geometry and materials.

II. BACKGROUND AND RELATED WORK

While object detection methods must search over the entire search space to first detect an object, tracking algorithms can be much faster by leveraging knowledge from previous frames to reduce the search space. However, this can make error recovery difficult. Tracking is also more complex because unlike object detection bounding boxes which are class specific, tracking is object specific. It assigns identifiable bounding boxes to each object and attempts to maintain each assignment over all frames. So, tracking not only detects but also maintains bounding box assignment for each object, over all frames. More granular than bounding boxes, segmentation may also be performed for pixel-level annotations. For an

[1]Department of Computer Science, UNC Chapel Hill, Chapel Hill, NC 27599, USA {wilson, lin}@cs.unc.edu
[2]Department of Computer Science, Univ. of Maryland, College Park, MD 20740, USA lin@cs.umd.edu

Object Detection/Tracking Datasets

Dataset	# Class	# Img/Vid
COCO [37]	80	330K img
DAVIS [10]	384	10.5K img
ILSVRC [51]	1000	1.4M img
KITTI [42], [58]	8	10.9K img
OTB [60]	100	100 vid
PASCAL VOC [15]	20	21K img
Sound-20K [63]	55+	20K vid
VOT2018 [32]	35	147K img
YouTube–VOS [61]	7800+	4K+ vid

TABLE I

IN CONTRAST TO OTHER DATASETS, THE SOUND-20K DATASET CONTAINS THE LARGEST AUDIO-VISUAL DATA FOR OBJECT INTERACTIONS IN A VIRTUAL SCENE AND PROVIDES AN EXCELLENT BASELINE FOR ASSESSING THE ACCURACY OF OUR AVOT METHOD AGAINST OTHERS.

attribute and performance comparison, object tracking benchmark [60] provides attribute and performance comparisons between various methods and evaluation criteria.

The majority of object detection and tracking methods are visually based, even though some datasets are generated from videos with audio. Table I lists commonly used datasets for object detection and tracking evaluation. We add ground truth bounding box annotations to the Sound-20K dataset and use it as a baseline for assessing the accuracy of our AVOT method. While the general methodology of research areas such as speaker detection and person tracking leverage both audio and visual information, their implementations are specifically aimed at tracking human speakers (e.g. face detection is part of their pipeline). Our method aims to be applicable in a broader context and does not make assumptions about the targets. It currently can track up to nineteen object-material classes. Next we discuss object detection, tracking, and audio-visual techniques in more detail, as compared to our work.

A. Object Detection

In addition to overall classification of an image, researchers are interested in also detecting and classifying the specific objects within an image. This can be achieved by using object detection methods to locate and label each object with a class-specific bounding box. As is similar in image classification, object detection techniques require large amounts of training data but in its case, more annotations for each example. Because, in the case of object detection, training data requires both class labels and bounding box coordinates for each object. For example, the PASCAL Visual Object Classes (VOC) dataset contains images, object annotations, and segmentations for twenty different classes. Other available datasets are mentioned in Table I. Unfortunately, only a few datasets make available the video and accompanying audio, making audio-visual methods more time-consuming to explore. We contribute our Sound-20K ground truth annotations to aid future audio-visual research in this area.

Video object detection: object detection can be performed not only on images but on video as well. Here, additional contextual information is available such as sound and image sequence. This temporal memory has allowed video detection to achieve start-of-the-art performance and speeds by learning lightweight scene features for mobile [38] and shifting channels along the dimension of time [36]. However, video also introduces new challenges such as motion blur, defocus, and various poses. Temporal coherence can also be used to overcome these defects with flow-guided feature aggregation [65], for instance. Finally, in addition to scene features, time shifting, and temporally coherent features, temporal propagation for on demand detection has also yielded efficiency gains [12].

B. Object Tracking

Object tracking differs from object detection in that the labels and bounding boxes are dependent. In other words, tracking attempts to establish correspondences of the same object over multiple frames, for example, one particular car in traffic over time. While object tracking has been studied for decades, numerous factors remain a challenge, such as illumination variation, occlusion, and background clutters [60]. Given the sequential nature of the task and method, tracking can be fast and efficient but also accumulate error and drift. Moreover, it is not easy for object trackers to recover from failure or an incorrect assignment to another object. Approaches such as frame skipping, Siamese trackers, and deep learning are a few of the existing techniques being used to perform object tracking and segmentation quickly and accurately.

Frame skipping: this baseline approach *detects* every N-th frame and *tracks* frames in between. Advantages of frame skipping are realized in terms of speed and precision by efficiently tracking on a majority of the frames while still allowing for correction by performing detection every N keyframes. This can be referred to as a fixed scheduler. Dynamic schedulers have also been considered. For instance, Detect or Track [41] uses a scheduler network to determine whether to detect or track at certain frames. In our research, we propose an audio-based scheduler procedure that can alter tracking during audio onset detection in videos (Alg. 1).

Siamese trackers: Siamese neural networks take two inputs and, with shared weights, predict if the two inputs belong to the same output class. Fully-convolutional Siamese approaches can be used for object tracking [7], [22], [35], [62], [66] and unlike batch processing, online Siamese methods can perform tracking on streaming video with access only to current and previous frames [59]. To improve initialization of online adaption-based deep networks such as these, offline meta-learning has been applied [45]. Asymmetric Siamese networks have also been studied and learn a linear template to search test images by cross-correlation [57].

Deep learning: last but not least, object tracking performed using deep learning. Faster R-CNN [50] is a real-time, state-of-the-art object tracker and four staged end-to-end neural network. First, a convolutional feature map of the image is obtained by extracting from a convolutional layer of a pre-trained CNN (e.g. ImageNet [33], ResNet [23],

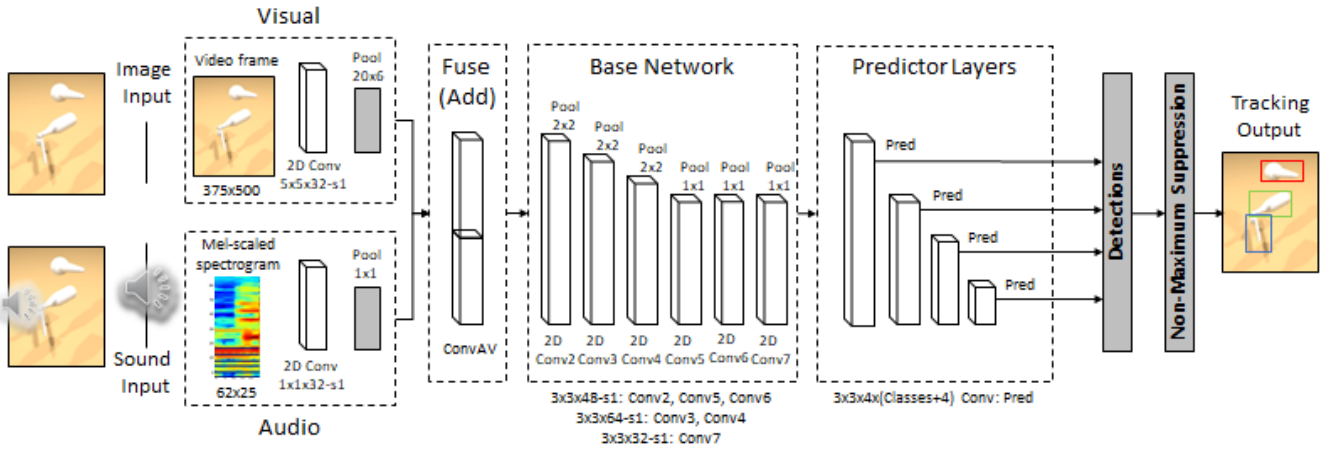

Fig. 2. Audio-Visual Object Tracker (AVOT) neural network architecture. AVOT is a feed-forward convolutional neural network that classifies and scales a fixed number of anchor bounding boxes to track objects in a video. Here, we define an object based on its geometry and material. Convolutional layers from the visual and audio inputs are fused using an add merge layer before being input into a base network of convolutional layers similar to standard classification networks. The base is then followed by predictor layers for detection, as is done in SSD, however designed and optimized for our audio-visual dataset and task. The single best detection for each object is then selected using non-maximum suppression.

MobileNets [28], DenseNet [29], etc.). The second stage is a Region Proposal Network, which are reference bounding boxes uniformly placed across the image. In this stage, specific regions are identified and adjusted based on the convolutional feature map from the first step. The third stage applies Region of Interest (RoI) Pooling to extract features from the convolutional map for each region. The fourth and final step then uses those features to classify the content in the bounding box (e.g. bottle, table, etc., background) and adjust the classified bounding box to a better fit, predicting $\Delta x_{center}, \Delta y_{center}, \Delta_{width}, \Delta_{height}$ from an anchor.

Single Shot MultiBox Detector (SSD) [39] is another real-time, state-of-the-art object tracker. SSD is slightly better than YOLO [49] in terms of speed while improving upon accuracy with additional feature layers on top of a base network [1]. Furthermore, SSD is slightly better than Faster R-CNN in terms of accuracy while eliminating object proposals with multiple feature maps of differing resolution. Although SSD uses similar default boxes, it applies them to several feature maps of different resolutions. In addition to a single unified framework for training and prediction, SSD input images are smaller at 300 x 300, compared to 512 x 512 for Faster R-CNN and 448 x 448 for YOLO [39]. This enables faster processing over other single shot, region proposal, and pooling techniques. This permits a wider range of computer vision applications to leverage this architecture. We use SSD as both a baseline and base network for our AVOT tracker.

C. Audio-Visual Methods

Audio-visual techniques have been used for speech separation [14], object and geometry classification [56], [63], [64], and audio-visual correspondence learning [3], [4]. Most directly related to audio-visual object tracking is speaker recognition [48], [55], tracking from audio-visual data using a linear prediction method [2], and object detection and tracking with audio and optical signals [27]. For speaker recognition, a face tracking algorithm and microphone array are used to estimate speaker position. These methods fuse audio and visual data by leveraging time delays in audio and motion changes in visual. While both modalities, in theory, can distinguish these changes, one may be more adept to do so. Also, the fusion of the two can decrease uncertainty and increase reliability [44]. Finally, audio can also come from contact microphones or acoustical sensors to capture touch sounds and optical signals for gesture recognition [27]. In our approach, we leverage audio from impact sounds of objects and images from video.

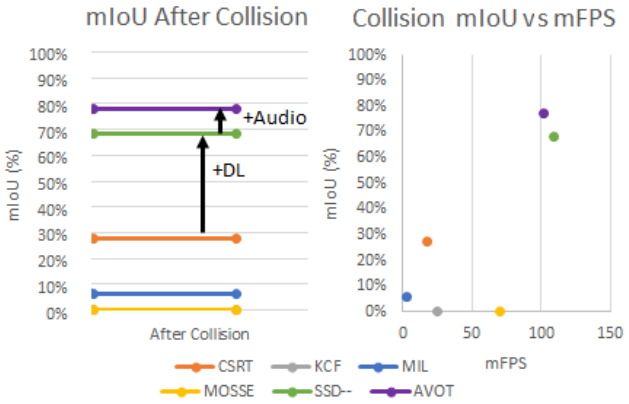

Fig. 3. Existing object trackers performance decline when objects collide in a moving two object Sound-20K virtual scene whereas AVOT improves with audio onset. Post-collision (i.e. when there's audio), deep learning (DL) methods achieve nearly 40% higher in accuracy over other methods and AVOT further outperforms SSD– by another 10% in *mean Intersection over Union* (mIoU) with an added benefit of audio-visual input. A scheduler network gated on audio can be used to achieve the best run performance and/or the highest accuracy across all cases using multimodal trackers.

III. Technical Approach

Unlike visually based object trackers, our method defines each object by its geometry and material. With audio-visual

[1]ImageNet VGG-16 was used as a base, but other neural networks should also produce good results.

data, the same shape (e.g. bottle) with different materials (e.g. steel vs. wood) are distinguishable and are therefore considered to be different objects. Our work also considers colliding objects. While a challenge for visually based tracking methods (Fig. 3), they provide auditory cues for an audio-visual object tracker. Scheduling between trackers can then be enabled based on audio availability.

Given the location of an object in the first frame of video, the object tracking task is to quickly and accurately estimate its position in all successive frames [54]. More specifically, for each video frame in a sequence $F = f_1, f_2, ..., f_N$ where N is number of frames, obtain bounding boxes $B = b_1, b_2, ..., b_M$ where M is the number of objects.

A. AVOT Neural Network Architecture

Similar to existing object tracking architectures, AVOT is a feed-forward convolutional neural network that classifies and scales a fixed number of anchor bounding boxes to fit each object in an image. We define an object based on its geometry and material. AVOT leverages audio and visual data for a more granular definition of an object to distinguish between objects with the same appearance but different materials.

Audio input: Audio frames from Sound-20K [63] videos match the image frame rate of 33 frames per second. As a result, each jpeg image has a corresponding 29 ms audio wav file. The audio is converted to mel-scaled spectrograms and serve as the audio input given their performance in CNNs for other tasks [30]. They are computed using a short-time Fourier transform with a 512 sample Hann window and 12.5% overlap. A Hanning (Hann) window was selected for its suitability for a variety of signals, good frequency resolution, and reduced spectral leakage. Each spectrogram is individually normalized and downsampled to a size of 62 frequency bins by 25 time bins (Fig. 4). Binning provides for appropriate fusion with image dimensions and weight matching to the logarithmic perception of frequency [56].

Image input: image dimensions are 500 x 375 pixels. Since SSD evaluated input sizes 300 x 300 and 512 x 512 (YOLO 448 x 448), our images are augmented but input dimensions unmodified as they fall within range of previous work. For data augmentation, we use common image transformations and sampling strategy similar to SSD and YOLO. Random cropping can be especially useful for creating zoomed in and out training examples to aid the classification of small objects in PASCAL VOC and Sound-20K. Each training image randomly samples from a data augmentation sequence to make the model more robust to object size and shape [39]. We use a reduced layer variation of VGG16 [53] as the base network leading up to our detection prediction layers. Images were extracted from video using ffmpeg with CRF scale set to 0 (lossless) and libx264 set to vcodec [63]. Each image is fused with its corresponding audio via an add-merge layer.

Architecture: Fig. 2 illustrates the layers of our multi-modal object tracker neural network. The early visual layer is based on [34] and audio layer based on impact [56] and environmental sound [30] classification. Convolutional layers

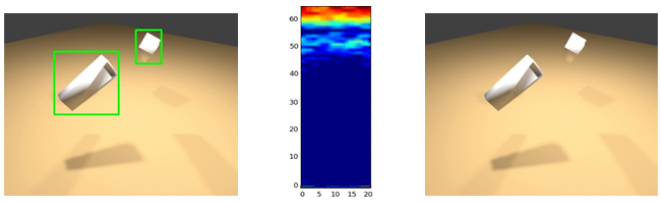

Fig. 4. AVOT needs ground truth boxes (left), input audio from the scene video converted to a mel-scaled spectrogram (center), and input image (right) for each object during training. We predict shape offsets and confidences for all object categories where an object is defined by its geometry and material.

from the visual and audio inputs are fused using an add merge layer. A multiply-merge layer was also considered and resulted in a similar training loss, however, at 1.5x the number of training epochs. Fused features are then input into a base network. Given our relatively small annotated audio-visual dataset, our base network is a reduced version of the standard image classification architecture [33]. The base is then followed by predictor, or also referred to as feature or classifier layers. Upper and lower feature maps are used for detection, as is done in SSD, to promote consistency and capture fine details respectively. The single best detection for each object is then selected using non-maximum suppression.

B. AVOT Dataset

Ground truth annotations were manually labeled by the experimenter for 18 objects. Each object is unique by geometry and material. The dataset is comprised of 17 three second videos of 103 image and audio frames each. This resulted in a total of 1,752 audio and visual segments. Videos contained one, two, and three colliding objects per scene. Our training and test datasets are split 80% and 20% respectively. The test dataset randomly samples frames from each video that are held out from training and used only for evaluation. For example, a video with 100 frames will have 20 frames randomly selected for test and the remaining 80 frames used for training. Fig. 5 shows loss by epoch for our AVOT tracker compared to a variation of visually based SSD.

C. Implementation Details

All models were implemented with Tensorflow [1] and Keras [13]. AVOT was run with early stopping at a maximum of 100 epochs, 100 steps per epoch, and batch size of 16 (Fig. 5). Training was performed using an Adam optimizer [31] and loss as defined by the weighted sum of localization loss (Smooth L1) and confidence loss (Softmax). We use a reduced variation SSD for predictor layers. AVOT anchor box scaling factors were set to 0.08, 0.16, 0.32, 0.64, and 0.96 and aspect ratios 0.5, 1.0, and 2.0 [39]. Here, we do not use SSD aspect ratios 1/3 or 3 given a smaller number of target classes. There are five scaling factors for four predictor layers because the last scaling factor is used for the second aspect ratio box of the last predictor layer. Although fewer layers, detections are still based on small 3 x 3 kernels at each feature map offset [39].

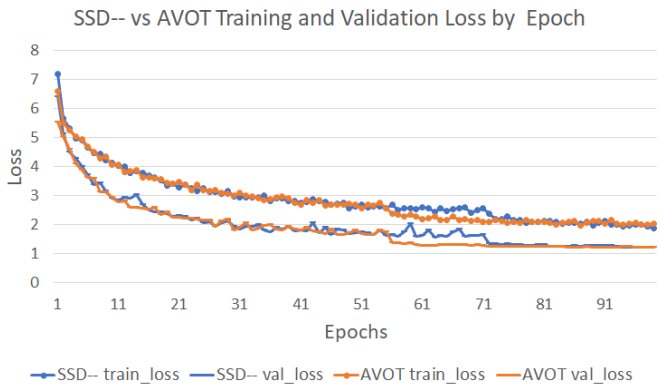

Fig. 5. The training (circle) and validation (line) loss for SSD– (blue) and AVOT (orange). Multimodal AVOT loss seems to decrease more consistently than visual only SSD with reduced layers, denoted as SSD–.

Initialization: our AVOT neural network uses he_normal initialization [21]. For evaluation, we also initialize baseline methods with he_normal rather than fine tune on pre-trained networks. Recent research suggests equivalent performance between random initialization for training instead of pre-trained weights [24]. Furthermore, given a smaller dataset of Sound-20K with ground truth annotations, we have reduced the layers of baseline implementations to avoid overfitting.

Non-maximum suppression (NMS) [43]: object trackers may produce more than one overlapping bounding box that are greater than the confidence and IoU thresholds for the same object. NMS is a post-process that selects the bounding box with the greatest confidence and suppresses remaining bounding boxes that overlap this maximum by some threshold. Here, NMS confidence and IoU threshold are set to 0.5 [39].

Scheduler network: Impact sounds from objects colliding emulate a type of scheduler network that can improve detections post collision. For added efficiency, only visual inputs can be processed leading up to audio onset. After, both audio and visual inputs can be used. In the case of our synthetic dataset, there is no audio prior to collision which makes audio onset easier to detect than videos with noise.

Algorithm 1 AVOT Scheduler Network

```
procedure AUDIOVISUALOBJECTTRACKER
    trackers ← initializeBoundingBoxes
    top:
    trackers ← update
    if Audio then RunScheduler
    goto top
    procedure RUNSCHEDULER
        AVOT detects objects
        re-initialize object trackers
        goto top
```

mIoU / mFPS Object Tracking Accuracy by Method

Method	2 Objects	3 Objects
AVOT (Ours)	**58.3%** / 101.6	**66.1%** / 101.0
CSRT [40]	46.9% / 17.1	30.1% / 4.7
KCF [26]	13.5% / 24.9	1.7% / 38.6
MIL [6]	43.0% / 2.5	21.6% / 1.6
MOSSE [8]	7.6% / 70.4	1.0% / **74.5**
SSD– [39]	55.5% / **108.7**	65.9% / **103.8**

TABLE II

MULTIPLE NETWORK MODELS WERE EVALUATED ON ACCURACY AND TIME USING MEAN INTERSECTION OVER UNION (MIOU) AND MEAN FRAMES PER SECOND (MFPS). **OURS IS AVOT**. FAILURE CASES FOR BASELINE METHODS WITHOUT AUDIO TEND TO CLASSIFY TO THE CORRECT GEOMETRY BUT WRONG MATERIAL. BY EXPLOITING BOTH VISUAL AND AUDIO DATA, AVOT ACHIEVES THE HIGHEST LEVEL OF TRACKING ACCURACY, WITH NEARLY COMPARABLE BEST RUNTIME PERFORMANCE, OVER EXISTING VISUAL TRACKING METHODS.

IV. EXPERIMENTS AND RESULTS

Evaluation was performed using ground truth annotations on the Sound-20K audio-visual dataset. This dataset is comprised of synthetic videos of multiple objects colliding in a scene. Training took roughly 30 minutes running on Ubuntu 16.04.6 LTS with a single Titan X GPU. We use Intersection over Union (IoU) for accuracy between ground truth and predicted object bounding boxes. As a general rule of thumb, a true positive prediction occurs when $IoU \geq 0.5$, according to the PASCAL VOC challenge. We measure the speed in mean frames per second (mFPS) with a batch size of 16 using a Titan X and cuDNN v7.4.2.

OpenCV implementations: online Multiple Instance Learning (MIL) [6], Kernelized Correlation Filters (KCF) [26], Discriminative Correlation Filter with Channel and Spatial Reliability (CSRT) [40], and an adaptive correlation filter known as Minimum Output Sum of Squared Error (MOSSE) [8] are a few trackers available in OpenCV [9]. We selected to evaluate these as baselines due to their advantages in terms of accuracy and/or speed. For these methods, appearance is learned from first frame bounding boxes that are initialized with ground truth coordinates.

A. Our Results vs. Baselines

Given our limited number of training examples in our audio-visual dataset, we used a reduced layer implementation of SSD (labeled in Table II as SSD–) for a baseline and base network for AVOT. Our AVOT neural network outperforms SSD– and other baseline methods after collision (Fig. 3). As shown in Fig. 3, AVOT was able to achieve the highest level of accuracy of 80% in *mean Intersection over Union* (mIoU)– about 10% more accurate than SSD. While these results are AVOT only, we further propose a scheduler network (**Algorithm 1**) to switch between AVOT and other methods based on audio onset to maximize accuracy and performance over all frames in multimodal object tracking.

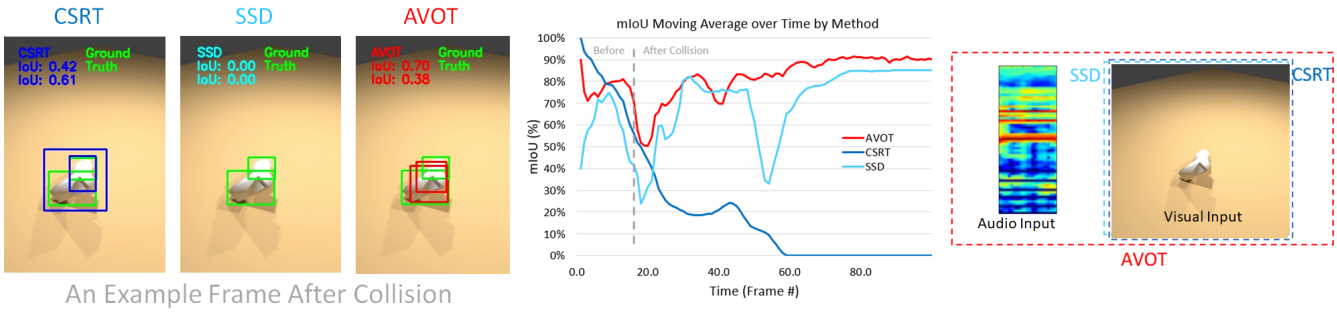

Fig. 6. We compare CSRT and SSD to our AVOT method for multi-object tracking. Two colliding objects with the same geometry but different materials are tracked free-falling in a virtual scene from Sound-20K [63]. CSRT is unable to maintain tracking post-collision and while SSD recovers, it temporarily loses tracking at the time of occlusion. Audio-visual AVOT maintains tracking across all frames. Please see the Supplementary Video for more demonstrations.

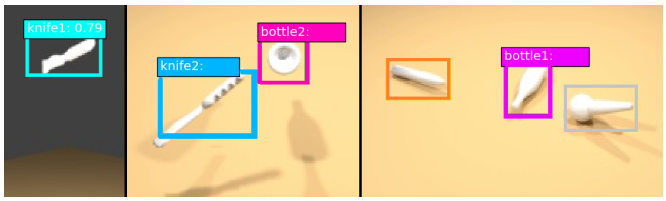

Fig. 7. Examples of AVOT applied to virtual scene of Sound-20K with predicted bounding box. These are exemplary screenshots of AVOT performing object tracking before and after collisions for one, two, and three object virtual scenes. Notice alphanumeric labels (e.g. bottle1 and bottle1) to differentiate the same geometry with different materials.

B. Maximization Activation

We analyzed activation maximizations to visualize the spectrogram audio and visual input which would produce the highest activation for a given volume class. They demonstrate features being learned by both modalities for the object tracking task. Please see the Supplementary Video for demonstration.

V. CONCLUSION

We present AVOT, an end-to-end trained neural network for object tracking using audio and visual data from videos. To distinguish between similar objects with different materials, we define an object based on its geometry and material. This more granular categorization benefits from a multimodal learning approach using audio and visual data, where audio is typically available from the sources of video but are currently underutilized. By fusing audio with visual data, our audio-visual object tracker (AVOT) outperforms single-modality methods when audio is present from impact, collision, and rolling sounds while maintaining real-time performance. We evaluated against Sound-20K and make our audio-visual data along with ground truth bounding box annotations available for future research in this area.

Future work: we will expand the size of our training set by annotating more objects in the Sound-20K dataset, increase the number of object classes that we are predicting, evaluate alternative fusion methods, and perform sensitivity analysis on scaling factors and aspect ratios. We would also like to augment our audio data and experiment with a variation of our object tracker with audio only.

ACKNOWLEDGMENT

This work is supported in part by the U.S. National Science Foundation and the Elizabeth Stevinson Iribe Chair Professorship.

REFERENCES

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu. Y, and Zheng, X. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org. https://www.tensorflow.org/

[2] Anusha, N. and Roy, L. Object Tracking from Audio and Video data using Linear Prediction method. National Institute of Technology, Rourkela Master's thesis. May 2015

[3] Arandjelovic, R. and Zisserman, A. Look, Listen and Learn. IEEE International Conference on Computer Vision. CoRR. 2017

[4] Arandjelovic, R. and Zisserman, A. Objects that Sound. CoRR. 2017

[5] Aytar, Y., Vondrick, C., and Torralba, A. 2016. Soundnet: Learning sound representations from unlabeled video. Advances in Neural Information Processing Systems. pp. 892–900

[6] Babenko, B., Yang, M.-H., Belongie, S. Visual tracking with online Mulitple Instance Learning. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2009

[7] Bertinetto, L., Valmadre, J., Henriques, J., Vedaldi, A., and Torr, P. HS. Fully-Convolutional Siamese Networks for Object Tracking. In European Conference on Computer Vision workshops, 2016

[8] Bolme, D., Beveridge, J., Draper, B., and Lui, Y. M. Visual Object Tracking using Adaptive Correlation Filters. CVPR. 2010

[9] Bradski, G. The OpenCV Library. Dr. Dobbś Journal of Software Tools. 2000

[10] Caelles, S., Pont–Tuset, J., Perazzi, F., Montes, A., Maninis, K.–K., and Van Gool, L. The 2019 DAVIS Challenge on VOS: Unsupervised Multi-Object Segmentation, CoRR, 2019

[11] Checka, N. and Wilson, K. Person tracking using audio-video sensor fusion. MIT Artificial Intelligence Laboratory journal. Volume 2002. 2001

[12] Chen, K., Wang, J., Yang, S., Zhang, X., Xiong, Y., Loy, C., and Lin, D. Optimizing Video Object Detection via a Scale-Time Lattice. 2018

[13] Chollet, F. and others. 2015. Keras. https://github.com/keras-team/keras

[14] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W., and Rubinstein, M. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. CoRR. 2018

[15] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results

[16] Everingham, M. and Eslami, S. M. A. and Van Gool, L. and Williams, C. K. I. and Winn, J. and Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision. Volume 111, number 1, pp. 98–136. January 2015

[17] Geiger, A. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3354–3361. 2012

[18] Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. Proc. IEEE ICASSP, 2017

[19] Grabner, H., Grabner, M., Bischof, H. Real-time Tracking via On-line Boosting. In Proceedings British Machine Vision Conference (BMVC), Volume 1, pp. 47 – 56. 2006

[20] Grabner, H., Bischof, H. On-line Boosting and Vision. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Volume 1, pp. 260 – 267. 2006

[21] He, K., Zhang, X., Ren, S., and Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. CoRR. 2015

[22] He, A., Luo, C., Tian, X., and Zeng, W. Towards a better match in siamese network based visual object tracker. In European Conference on Computer Vision workshops, 2018.

[23] He, K., Zhang, X., Ren, S., Sun, J. Deep Residual Learning for Image Recognition. CoRR. 2015

[24] He, K., Girshick, R.,Dollár, P. Rethinking ImageNet Pre-training. 2018

[25] Held, D., Thrun, S., and Savarese, S. Learning to Track at 100 FPS with Deep Regression Networks. Proceedings of the European Conference on Computer Vision (ECCV). 2016

[26] Henriques, J. a. F., Caseiro, R., Martins, P., and Batista, J. High-Speed Tracking with Kernelized Correlation Filters. PAMI, 2015

[27] Holz, D. Object detection and tracking with audio and optical signals. United States Patent US9465461 and US9465461B2

[28] Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. CoRR. 2017

[29] Huang, G., Liu, Z., Weinberger, K. Densely Connected Convolutional Networks. CoRR. 2016

[30] Huzaifah, M. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. CoRR, 2017

[31] Kingma, D. and Ba, J. Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR), 2015

[32] Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., and Čehovin, L. A Novel Performance Evaluation Methodology for Single-Target Trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016

[33] Krizhevsky, A., Sutskever, I., and Hinton, G. 2012. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25. pp. 1097–1105

[34] LeCun, Y. and Bengio, Y. Convolutional Networks for Images, Speech, and Time Series. The Handbook of Brain Theory and Neural Networks. MIT Press, pp. 255 – 258, 1998

[35] Li, B., Yan, J., Wu, W., Zhu, Z., and Hu, X. High performance visual tracking with siamese region proposal network. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[36] Lin, J. Gan, C., and Han, S. TSM: Temporal Shift Module for Efficient Video Understanding. CoRR. 2018

[37] Lin, T-Y, Maire, M., Belongie, S. J., Bourdev, L., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common Objects in Context. CoRR. 2014

[38] Liu, M., Zhu, M., White, M., Li, Y., Kalenichenko, D. Looking Fast and Slow: Memory-Guided Mobile Video Object Detection. CoRR. 2019

[39] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. SSD: Single Shot MultiBox Detector. CoRR. 2015

[40] Lukezic, A., Vojir, T., Cehovin, L., Matas, J., Kristan, M. Discriminative Correlation Filter with Channel and Spatial Reliability. International Journal of Computer Vision, Volume 126. 2016

[41] Luo, H., Xie, W., Wang, X., and Zeng, W. Detect or Track: Towards Cost-Effective Video Object Detection/Tracking. CoRR. 2018

[42] Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. MOT16: A Benchmark for Multi-Object Tracking. CVPR, 2016

[43] Neubeck, A. and Van Gool, L. Efficient Non-Maximum Suppression. Proceedings of the 18th International Conference on Pattern Recognition – Volume 03. pp. 850 – 855. ICPR, 2006

[44] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Multimodal Deep Learning. Proceedings of the 28th International Conference on Machine Learning, ICML 2011

[45] Park, E. and Berg, A. C. Meta-Tracker: Fast and Robust Online Adaption for Visual Object Trackers. CoRR. 2018

[46] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., and Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[47] Prisacariu, V., Kahler, O., W. Murray, D., and D. Reid, I. Real-Time 3D Tracking and Reconstruction on Mobile Phones. IEEE Transactions on Visualization and Computer Graphics. pp. 557 – 570. 2015.

[48] Qian, X., Brutti, A., Lanz, O., Omologo, M., and Cavallaro, A. Multi-speaker tracking from an audio-visual sensing device. IEEE Transactions on Multimedia (2019)

[49] Redmon, J., Divvala, S. K., Girshick, R. B., and Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. CoRR. 2015

[50] Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. pp. 91–99. 2015

[51] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), Volume 115, Number 3, pp. 211–252, 2015.

[52] Schulz, D., Burgard, W., Fox, D., Cremers, A.B. Tracking multiple moving objects with a mobile robot. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001

[53] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. NIPS, 2015

[54] Smeulders, A. W., Chu, D. M., Cucchiara, R., Calderara, S., Dehghan, A., and Shah, M. Visual tracking: An experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.

[55] Spors, S., Rabenstein, R., Strobel, N. Joint audio-video object tracking. International Conference on Image Processing. pp. 393 – 396 volume 1. 2001

[56] Sterling, A., Wilson, J., Lowe, S., and Lin. M. C. ISNN: Impact Sound Neural Network for Audio-Visual Object Classification. Computer Vision – ECCV, pp. 578–595. 2018

[57] Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P. HS. End-To-End Representation Learning for Correlation Filter Based Tracking. pp. 2805–2813. CVPR. 2017

[58] Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B., Geiger, A., and Leibe, B. MOTS: Multi-Object Tracking and Segmentation. CoRR. 2019

[59] Wang, Q., Zhang, L., Bertinetto, L., Hu, W., and Too, P. HS. Fast Online Object Tracking and Segmentation: A Unifying Approach. 2018

[60] Wu, Y., Lim, J., Yang, M.-H. Object Tracking Benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 37, Issue 9, September 2015

[61] Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., and Huang, T. Youtube-vos: Sequence-to-sequence video object segmentation. In European Conference on Computer Vision, 2018.

[62] Yang, T. and Chan, A. B. Learning dynamic memory networks for object tracking. In European Conference on Computer Vision, 2018.

[63] Zhang, Z., Wu, J., Li, Q., Huang, Z., Traer, J., McDermott, J., Tenenbaum, J., and Freeman, W. T. Generative Modeling of Audible Shapes for Object Perception. The IEEE International Conference on Computer Vision (ICCV), 2017

[64] Zhang, Z., Li, Q., Huang, Z., Wu, J., Tenenbaum, J., and Freeman, W. T. Shape and Material from Sound. Advances in Neural Information Processing Systems (NIPS), pp. 1278–1288, 2017.

[65] Zhu, X., Wang, Y., Dai, J., Yuan, L., and Wei, Y. Flow-Guided Feature Aggregation for Video Object Detection. CoRR. 2017

[66] Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., and Hu, W. Distractor-aware siamese networks for visual object tracking. In European Conference on Computer Vision, 2018.