AudAlign

Project Proposal

Project Advisor:

Dr. Muhammad Ahmad Raza

Group Members:

Zain Mehmood	21L-1825
Minal Anwar	21L-1848
Ali Nawaz Chaudary	21I-0801

National University Of Computer and Emerging Sciences
Department of Computer Science
Lahore, Pakistan

Abstract

Multimedia content extensively uses carefully picked sounds to enhance the quality and user experience. These sounds are mostly created manually by artists and synced by video editors which is a time consuming task done repeatedly in content production. AudAlign aims to address the problem of synchronizing and adding sounds produced by interaction of two objects in a video. The sound relates to how an object's interaction sounds in the real world. Objects in video can either be dynamic bodies or a moving object interacting with a static object. To solve this problem, object detection will be used to trigger the appropriate sound which relates to the object in question, object segmentation will be majorly used to detect boundaries of the object which will aid in collision detection and synchronization timing. Other metrics such as object's velocity and surface area of collision will augment the sound played. Synchronizing and associating sounds automatically is expected to help reduce the contemporary manual labor of sound synchronization in multimedia content.

1. Introduction

Since digital platforms have made multimedia content more widely available, creating engaging audio-visual experiences has become crucial to holding an audience's attention. Sound is an essential component of this, as it adds reality and depth to the experience for the viewers. Traditionally, sounds in multimedia content, such as movies and videos, are precisely created and synchronized by sound designers and video editors. This manual process of syncing sounds with visual elements is time-consuming and labor-intensive, particularly when dealing with complex interactions between multiple objects.

With the increasing demand for efficient content production, there is a growing need for automated solutions that can streamline this process. The goal of this project, AudAlign, is to solve the problem of integrating and synchronizing sounds that arise from different objects. For example, when a moving object collides with another, or when a dynamic object interacts with a static one, the sound produced in reality can be replicated digitally. AudAlign approach to achieve this by utilizing advanced object detection and segmentation techniques, to detect the collision which will trigger appropriate sounds that correspond to the objects in question. This ensures that the audio component is not only synchronized but also accurately represents the real-world interaction dynamics.

The core of AudAlign’s approach lies in using machine learning models to automate sound synchronization. Object detection is employed to identify the objects involved in the interaction, while object segmentation defines their boundaries to enhance collision detection and timing precision. Additionally, the system takes into account various physical metrics, such as the objects' velocity and the surface area of collision, to determine the characteristics of the sound to be played. By automating these processes, AudAlign aims to significantly reduce the manual labor involved in sound synchronization, thus accelerating the content creation pipeline and enhancing the overall quality of multimedia productions.

AudAlign will transform sound synchronization in multimedia by offering a smooth and automatic solution for aligning sounds with visual cues.

2. Goals and Objectives

Our project’s’ objectives include:

- Detection & Segmentation of Table like surface, ping pong ball, Table tennis Racket
- Detecting collision of Two moving objects and Single moving object
- Calculating physical metrics such as velocity, surface area and depth during collision
- Selecting and augmenting appropriate sound based on objects
- Synchronizing the sound on collision with precise timing

3. Scope of the Project

There are hundreds of varying sound classes produced by interaction of different objects, thus we have limited our scope to a selected few objects. Our project will mainly focus on synchronizing sounds of a table tennis match as it provides non periodic collisions with varying intensities of sounds. Our project also plans to implement consistent depth estimation for accurate collision detection due to which we will be focusing on fixed camera videos. In addition to this, A web application or user interface will be built to allow the users to select various options from sounds synchronized, fine tune and verify the automatic synchronization and quality.

4. Initial Study and Work Done so Far

In our literature review, we observed that due to advancement in deep generative models, the problem of audio-visual synchronization and synthesis is being worked on actively. The current sound synthesis and audio-video synchronization primarily focuses on synthesizing synchronized audio tracks for critical scenarios in soundless videos [1]. Many different technologies and techniques have been used to solve the audio-visual synchronization, but they focus on areas such as speech-oriented synching [2]. We aim to synchronize and augment the sound produced by interaction of two objects.

Since this is an area in which no major attempts have been made, we plan to solve this problem using a series of steps that involves a combination of established object detection and segmentation models, namely YOLO (You Only Look Once) [3] and a novel technique to estimate the depth in a video using a single perspective [4]. This will help us prevent the detection of overlapping objects with various depths as collisions. So far, we have been exploring the relevant technologies that can be used in our project and collecting Datasets for training the segmentation models along with audio libraries for finding the right sound.

References

[1] S. Ghose and J. J. Prevost, "AutoFoley: Artificial Synthesis of Synchronized Sound Tracks for Silent Videos With Deep Learning," in IEEE Transactions on Multimedia, vol. 23, pp. 1895-1907, 2021, doi: 10.1109/TMM.2020.3005033.

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In Proc. ECCV, 2020b.

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91

[4] Luo, X., Huang, J. B., Szeliski, R., Matzen, K., & Kopf, J. (2020). Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, *39*(4), 71-1.