Suggesting Sounds for Images from Video Collections

Matthias Solèr[1], Jean-Charles Bazin[2], Oliver Wang[2], Andreas Krause[1] and Alexander Sorkine-Hornung[2]

[1]Computer Science Department, ETH Zürich, Switzerland
{msoler,krausea}@ethz.ch
[2]Disney Research, Switzerland
{jean-charles.bazin,owang,alex}@disneyresearch.com

Abstract. Given a still image, humans can easily think of a sound associated with this image. For instance, people might associate the picture of a car with the sound of a car engine. In this paper we aim to retrieve sounds corresponding to a query image. To solve this challenging task, our approach exploits the correlation between the audio and visual modalities in video collections. A major difficulty is the high amount of uncorrelated audio in the videos, i.e., audio that does not correspond to the main image content, such as voice-over, background music, added sound effects, or sounds originating off-screen. We present an unsupervised, clustering-based solution that is able to automatically separate correlated sounds from uncorrelated ones. The core algorithm is based on a joint audio-visual feature space, in which we perform iterated mutual kNN clustering in order to effectively filter out uncorrelated sounds. To this end we also introduce a new dataset of correlated audio-visual data, on which we evaluate our approach and compare it to alternative solutions. Experiments show that our approach can successfully deal with a high amount of uncorrelated audio.

Keywords: Sound suggestion, audio-visual content, data filtering

1 Introduction

Visual content interpretation is at the core of computer vision. Impressive results have been obtained for visual data over the last years, e.g., for classification and recognition [57,29], segmentation [23], tracking and 3D reconstruction [1,9]. In comparison, learning relationships between audio and visual data is still a largely unexplored area, despite exciting applications on joint audio-video processing [17,6,13,37].

A fascinating example of joint audio-visual learning occurs daily in our lives. When humans see an object, they can usually imagine a plausible sound that it would make, due to having learned the correlation between visual and audio modalities from numerous examples throughout their life. In this paper, we aim to mimic this process, i.e. given an input still image, suggest sounds by interpreting the visual content of this image. Before proceeding further, it worths

mentioning that the sounds associated to an image can be highly ambiguous or even inexistent. For example, returning a sound for an image of a flower would not make sense because a flower does not make sound. In the following, we only consider images for which a sound can be associated.

Being able to output a sound for a query image has many practical applications in computer vision and multimedia, for example automatic Foley processing for video production[1] [19], image sonification for the visually impaired [44], and augmenting visual content database with sounds and audio restoration [26].

One approach to generate sound from images is to synthesize audio using physics-based simulation [19,11], however, this is still an open problem for general objects. Instead, we define the sound retrieval task as learning the correlation between audio and visual information collected from a video collection.

Video provides us with a natural and appealing way to learn this correlation: video cameras are equipped with microphones and capture synchronized audio and visual information. In principle, *every* single video frame captured constitutes a possible training example, and the Internet provides us with a virtually inexhaustible amount of training data. However, in practice there exist a number of significant challenges. First and foremost videos often contain a very high amount of *uncorrelated* audio, i.e., audio that does not correspond to the visual content of the video frames. This is due to voice-over (commentaries, speech, etc), background music, added sound effects, or sounds originating off-screen. This high level of noise in the training samples causes difficulties when employing standard machine learning techniques. An additional challenge is that an object might be naturally associated with different sounds, so we would like to learn and capture a multi-modal solution space. Furthermore, evaluating the quality of suggested sounds is also not trivial due to the difficulty in acquiring ground truth image-sound correspondences.

So, given an image, how can we find a sound that could correspond to this image? As the input to the training phase of our method, we take a collection of unstructured, casually captured videos with corresponding audio recordings. A key observation in our method is that while the input videos may contain a significant amount of uncorrelated data, i.e., images whose audio is uncorrelated (commonly due to added voiceover or music), these uncorrelated examples will not share any common features. On the other hand, the true visual and audio examples will be recurring across multiple videos.

In the rest of this paper, we use the term "correlated" audio-visual examples to refer to pairs of video frames and their associated audio segments whose audio corresponds to the visual content of the video frame. These audio segments are called "clean" since they should correspond to uncorrupted audio segments, i.e., free from background noise, added music or voiceover for example.

Based on this assumption, we develop an unsupervised filtering approach that automatically identifies significant audio-visual clusters via a mutual kNN-based technique, and then removes the uncorrelated visual-audio examples from the video collection. This filtering is performed offline, once, and results in a clean

[1] `https://en.wikipedia.org/wiki/Foley_%28filmmaking%29`

collection of correlated audio-visual examples that can then be used to retrieve a sound corresponding to the input query image. This approach to output a sound for a query image from a video collection corrupted with uncorrelated data constitutes the main contribution of our work. Given a query image, we can then output a corresponding sound by looking up the most similar visual features in the filtered collection and returning its corresponding audio segment. This retrieval step is conducted online at interactive rates.

2 Related Work

Our paper deals with audio and visual content. We will first review works which are related to either modality, and then works on joint audio-visual content.

2.1 Visual Content

Visual Classification and Recognition Recent works have demonstrated impressive results for visual classification and object recognition (see review in [57]). A possible approach for our sound retrieval problem would be to classify the query image using one of these classification techniques (e.g., assign it a "phone" label), and then return the sound of this query label from examples in a certain dataset. If a clean sound database with ground truth labels was available, this could be done simply as a lookup problem using query labels derived from image or video classification [57,29,7,61,64]. While such sound databases exist nowadays (e.g., *freesound.org*), the main drawback of this approach is that building and expanding such a database requires a considerable amount of manual and time-consuming work to label and monitor all samples, and moreover, not all types of sounds are available in these databases. In contrast our goal is to develop a fully automatic and unsupervised approach that can mine data from arbitrary unstructured video collections.

Another option would be to identify a video frame from the collection that is visually similar to the input query (e.g., using distance between image appearance descriptors [15,49,58]), and then output the audio segment of that video frame. However, outputting the sound of a video frame selected from visual similarity (obtained via image labels in a classification framework or via image descriptors) might return uncorrelated audio (see our evaluation) since a potentially large amount of audio tracks in the video collection might not be correlated with the observed visual content.

Representative Images Doersch et al. [20] find visual cues (e.g., patches) representative for some cities by relying on a database of images with ground truth GPS labels. Instead, we aim for an unsupervised approach and our input is a collection of unlabeled videos.

Some other works are dedicated to creating clean image collections. For example, Elor et al. [3] aim to obtain a clean set of images (and their segmentation). In contrast, our work is dedicated to learning image sounds from videos, which

requires to consider both visual and audio modalities, in order to reliably retrieve an appropriate sound for a query image.

Another option could be to build on methods that identify canonical images from an image collection [60,14]. However, it is not clear how to extend these methods to multiple modalities in our application scenario.

2.2 Audio Content

Audio Synthesis An approach to output sound of a given image could be to synthesize the audio corresponding to the image, e.g., using physics-based simulation [19,11,68,69,10,54]. However most of these techniques are dedicated to specific objects (e.g., fire) and cannot be generalized due to the task complexity.

Audio Classification, Recognition and Retrieval Audio analysis has been studied for a long time [2], and many products are now available on the market, for example Siri and Google Now for speech recognition, or Soundcloud, Spotify, and Shazam for music data. Audio classification techniques [35] could be used to label sounds in a video, but this does not solve our problem since the query image has no sound. One option would be to detect and ignore the frames of the video collection with particular audio, for example tutorial speeches or background music, and then apply the above approach. However it would require the construction of a handcrafted list of heuristics and will not be able to deal with the numerous forms of uncorrelated audio.

2.3 Audio-Visual Content

Audio Source Localization in Images Some works [30,4,27] use the audio and visual signals of an input video to identify which pixels of the video "sound" by associating changes in audio with visual motion changes. Since these approaches are designed to return pixel locations in videos, it is not clear how they can be extended to learn the sound of an image and output a sound given a still image.

Audio Suggestion from Visual Content Audio suggestion from visual content is mainly performed in the context of music recommendation (e.g., for a series of pictures [62], picture slideshows [22,36] or videos [38]), and music synthesis [42,47,16,66] for artistic performances. Outputting (or recognizing) speech from videos has been studied in the particular context of lip reading [48,43,67]. Some methods are also dedicated to cross-modal learning [24]. In contrast to these works, we are interested in finding a real-world *sound* (e.g., rather than musical backing) corresponding to a query image. This application requires finer scale retrieval, and notably different datasets. Closely related to our goal is the recent work of Owens et al. [50] which synthesizes impact sounds of objects for silent videos. Our work and theirs were conducted independently and in parallel. Their method estimates the feature representation of audio given a video using a recursive neural network and selecting the most similar example from their

video collection. Their video collection only contains correlated, clean and non-overlapping sounds that were acquired by a user manually hitting the objects with a drumstick. In comparison, we consider an unstructured video collection which in practice contains a high amount of uncorrelated audio. Our proposed approach can process this corrupted collection and returns a clean version, which could be used as an input training set by these methods.

3 Proposed Approach

We first discuss preprocessing steps (visual and audio descriptors, low audio frames), the core filtering step, and finally describe the retrieval step.

3.1 Preprocessing

We first normalize the videos such that they all have the same resolution, frame rate and audio sampling rate. We remove low audio frames and then compute the descriptors of the remaining visual and audio data, as discussed in the following.

Low Audio Pre-Filtering In practice, only few frames per video actually contain sound related to the visual content. Many of the other frames have a rather low volume or just contain background noise, which we first filter out. This pre-filtering is an effective way to reduce the amount of data to be subsequently processed. To filter out the audio segments of such frames, we use the Root Mean Square (RMS) audio energy computed over short windows [56], and simply keep the audio segments with RMS scores above the median RMS of each video. For reference, on the video collection used in this paper (Sec. 4.1), about 65% of the frames are filtered out.

Audio Descriptors To describe audio data, we employ the popular Mel-Frequency Cepstral Coefficients (MFCCs) [45], which are commonly employed descriptors in audio analysis, description and retrieval [39,51,65]. We compute the audio descriptor of a video frame by the MFCC feature over a temporal window centered at that frame. Since sounds might have different durations, we consider multiple temporal windows, and then concatenate these (multi-scale) MFCC features [18]. In practice, we used 5 windows of lengths 1, 3, 5, 7, 9 times of a frame duration (i.e., 40ms to 360ms for 25fps videos).

Visual Descriptors Recent works showed that reliable visual descriptors for mid to high level image information can be obtained by deep learning [53,21,58,28]. These works also showed that such feature descriptors effectively generalize to different tasks and image classes that they were originally trained on. Based on these impressive results, we compute the visual descriptor of each frame of the video collection by Caffe [28] using a network trained for image classification.

Joint Audio-Visual Features Every video frame and its associated audio segment from the video collection (after the low audio pre-filtering) constitutes a training sample. We combine both the audio and visual features to create a weighted joint audio-visual feature space. Let s_i be the i-th audio-visual sample and $\mathcal{S} = \{s_1, \ldots, s_n\}$ be the set of n audio-visual samples. Each s_i consists of a pair of corresponding (normalized) visual and audio descriptors, respectively written f_i^V and f_i^A. The distance between two audio-visual samples in the combined space is defined as a weighted sum of both modalities [5]:

$$d(s_i, s_j) = w_V d(f_i^V, f_j^V) + w_A d(f_i^A, f_j^A) \tag{1}$$

where the adjustable weights w_V and w_A allow us to tune the clustering procedure described in the following.

3.2 Clustering of Correlated Audio-Visual Samples

After having a reduced set of frames with audible audio segments, we now aim to identify the correlated pairs of frames and audio segments, and filter out the uncorrelated ones. We achieve this by finding significant clusters of correlated samples in the audio-visual feature space. Straightforward clustering in our joint audio-visual feature space, e.g., using mean-shift or similar techniques, is not practical, as we need to deal with a significant amount of outliers. We instead propose to use mutual kNN, which has been shown to be particularly effective for identifying the most significant clusters [41,40,3].

Contrary to conventional kNN, two nodes of a mutual kNN graph are connected if and only if their k-nearest neighbor relationship is *mutual*, i.e., if they are in each others' k-nearest neighbor set [8]. The clusters are then obtained by computing the connected components of the mutual kNN graph. In our application, we construct a graph where there is a node per audio-visual example, and an edge between two nodes if their feature descriptors are part of the k-nearest neighbors of each other, where the distance is defined at Eq. 1. We ignore small clusters as noisy examples can randomly create small mutual neighbor clusters [40].

The above mentioned weights for the audio and visual features in combination with the mutual kNN procedure allow for an iterative approach to remove uncorrelated samples. Let each mutual kNN cluster $\mathcal{C}_l$ over the set $\mathcal{S}$ be defined as:

$$\mathcal{C}_l = \{\mathcal{S}_l \subseteq \mathcal{S} | s_i \in \mathcal{N}(s_j) \text{ and } s_j \in \mathcal{N}(s_i), \forall (s_i, s_j) \in \mathcal{S}_l\} \tag{2}$$

where $\mathcal{N}(s_i)$ is the set of the k nearest neighbors of s_i according to the feature descriptor distance at Eq. 1. When iteratively exploring the clustering and with dynamic weights, we obtain a set of clusters such that

$$\mathcal{C}_l^t = \{\mathcal{S}_l \subseteq \mathcal{C}_l^{t-1} | s_i \in \mathcal{N}_{w^t}(s_j) \text{ and } s_j \in \mathcal{N}_{w^t}(s_i), \forall (s_i, s_j) \in \mathcal{S}_l\} \tag{3}$$

where $\mathcal{C}_l^t$ is a cluster obtained at iteration t, with $\mathcal{C}_l^0 = \mathcal{S}$ and $1 \leq t \leq T$. To emphasize that the relative weights (w_V and w_A) might evolve along time at

Eq. 1, we write $\mathcal{N}_{w^t}(s_i)$ the set of the k nearest neighbors of s_i obtained with the weights $w^t = (w_V^t, w_A^t)$ at iteration t.

If the influence weights w^t are fixed then the final clustering is obtained after one iteration, i.e., $T = 1$. If the influence weights can evolve, then the clustering at the current iteration is performed on the clusters of the previous iteration in a hierarchical manner. Depending on the weights and number of iterations, Eq. 3 may correspond to one of the following instances:

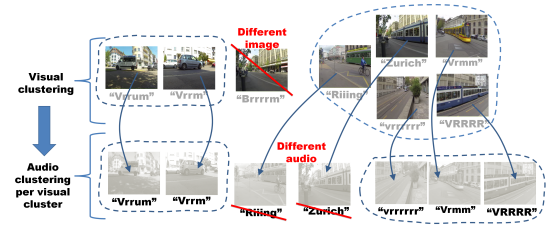
- $T = 1$ and $w_V = w_A = 1$ represent a one-step mutual-kNN on the joint audio-visual space. It simultaneously considers both visual and audio modalities. We call it "joint". The intuition for this joint space clustering is that it could potentially retrieve multiple sounds for visually similar images (e.g., for a same object), as their different joint audio-visual features will tend to cluster in multiple clusters.
- $T = 2$, with $w_V^1 = 1$ and $w_A^1 = 0$ at the first iteration, and then $w_V^2 = 0$ and $w_A^2 = 1$ at the second iteration. It represents a hierarchical two-step mutual kNN approach where the first step provides visual clusters, and the second step clusters the audio for each visual cluster. We call it "V+A".
- $T = 2$, with $w_V^1 = 0$ and $w_A^1 = 1$ at the first iteration, and then $w_V^2 = 1$ and $w_A^2 = 0$ at the second iteration. It represents the inverse of the above strategy: i.e., first audio clustering, and then clustering the visual features associated to each audio cluster. We call it "A+V". The intuition is to favor objects which are the most common for a given sound, thus is related to the notion of finding objects with unique sounds.

Concretely, the "joint" strategy performs one mutual kNN on the visual and audio descriptors, and returns a set of clusters. Ideally, each computed cluster contains examples that are similar both visually and audio. This is illustrated in Fig. 1b. For clarification of the two-step process, let's consider "V+A". In the first step, we compute visual clusters, and in a second step, for each visual cluster, we compute the clusters on audio features associated to this visual cluster, as illustrated in Fig. 1a. In the first step, we want to find clusters containing similar looking objects or scenes. For this, we apply a first mutual kNN on the visual features. After the first iteration, we obtain clusters containing elements with mostly similar visual features, but varying audio features. Therefore, given a mutual kNN visual cluster, we apply a second mutual kNN to detect the audio features uncorrelated to this visual cluster and remove them.

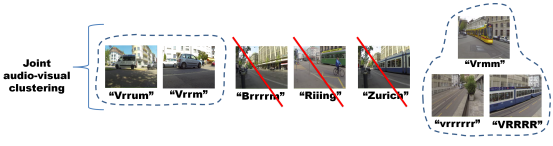
Interestingly, experiments will show that "joint" is outperformed by the strategies "V+A" or "A+V", each of these strategies "V+A" and "A+V has its own strengths, and the target application context determines the most relevant strategy.

3.3 Audio Retrieval

The above filtering provides a clean collection of correlated visual-audio examples. Given a query image, we now output a sound for this image by retrieving



(a)



(b)

Fig. 1: Illustration of the filtering step. (a): two-step mutual kNN hierarchical approach on visual and then audio feature space ("V+A"). (b): mutual kNN approach on the joint audio-visual feature space ("joint"). For illustrative purpose and a better understanding, instead of showing the audio signal, we write an onomatopoeia of the sound. In (a), the audio (resp. image) is not used in the visual (resp. audio) clustering step, and thus the sound onomatopeias (resp. the images) are greyed out.

an appropriate sound from this collection. The filtered clean collection allows for an efficient and robust procedure for the retrieval step: we compute the visual feature of the query image, look for the nearest neighbor of this visual feature in the clean collection, and output its corresponding audio segment. In practice, by storing the timestamp and video index of each example of the clean collection, we can access that video at that time, which allows to output its associated sound segment with any preferred duration set by the user.

Depending on the application or context, we might want to return a list of $M > 1$ sounds (rather than a single one). For this, we can return the top M sounds, i.e., the M nearest neighbors, for a user to select the preferred sound via an interactive interface. In case the user wants various sounds, we increase the variety of the sound outputs by retrieving $M' > M$ examples from the filtered collection (i.e., more than the number proposed to the user) and then suggesting a canonical subset of M of these retrieved audio segments via spectral clustering on the audio features [59].

4 Results

Evaluating whether a suggested sound corresponds to the input image is a difficult task. For example, given a door image, we expect to hear a "door" sound, but how can we decide if the suggested sound is indeed a sound of a door or not? The

Fig. 2: Representative images of the different categories of the introduced dataset.

class information is not sufficient due to extra sounds, background noise, etc, as mentioned earlier. Several video datasets exist for various purposes [61,55,32,25] but they either have no sound or do not provide ground truth correlated audio-visual data. Therefore, to evaluate the results, we create a dataset with "clean" sounds and known categories (Sec. 4.1). We analyze different aspects of filtering, and further evaluate our approach with a user study.

We invite the readers to refer to our project website to access our dataset as well as representative results of sound suggestions from query images.

4.1 Dataset

We manually collected several videos with minimal background noise and clean object sounds using a GoPro camera. The key advantage of this collection is that it provides a good estimate of the ground truth and also permits us to automatically measure and compare the accuracy and robustness of our approach.

We recorded videos for 9 different categories of image-sound pairs: *keyboard* typing, *washing* dishes, *door* opening and closing, walking on *stairs*, *vacuum cleaner*, drink *toasting*, using a *binder*, *trams* and *cars*. To reproduce real case scenarios, the videos were acquired from multiple viewpoint perspectives, both indoor and outdoor, with some overlap in visual and audio modalities (e.g., potentially similar appearance of street in *trams* and *cars*). Representative images from the different classes are shown in Fig. 2. Each category contains between 10 and 20 videos, with durations between 15 and 90 seconds, providing a set of about 150,000 examples of correlated audio segments and video frames.

We split up the samples into training and test sets such that no video has corresponding samples in both sets at the same time, and apply cross-validation over multiple splits. The training and test sets respectively contain around 80% and 20% of all the samples.

4.2 Experiments

Implementation Details The framework is implemented in Matlab (mutual kNN, etc) along with VLFeat modules (e.g., for nearest neighbor queries via approximate Nearest Neighbor). The parameters of the filters are set by cross-validation.

Our experiments are conducted on a desktop computer equipped with an Intel Core i7-960 at 3.2Ghz, 24GB RAM and a NVidia GTX 980Ti graphics card. Computing the visual descriptors using Caffe [28] takes about 5s per minute of video. The low audio pre-filtering step (Sec. 3.1) takes about 40ms per minute of audio. After this step, we have a collection of around 54,000 audio-visual examples. The mutual kNN approaches on this collection take about 15 minutes. The retrieval step for a query image over the filtered collection takes about 0.1s, which allows interactive use.

In terms of memory consumption, each visual-audio example is described by a 4096-dimension visual descriptor and $5 \times 13 = 65$ dimension multi-scale MFCC audio descriptor. Using single-precision floating-point format (4 bytes), each GB can store about 64,000 visual-audio samples. The additional memory footprint of k-d trees is comparatively small.

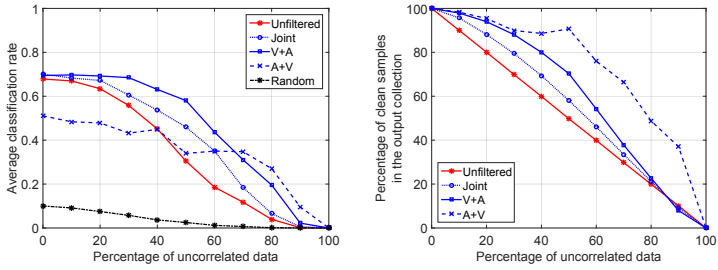
Comparison To evaluate our filtering approach, we compare it to an "unfiltered" approach that outputs audio segments from the video collection without the correlation filtering step (Sec. 3.2), i.e., it returns the sounds of the most visually similar images in the original collection. As a general sanity check, we also apply a random selection of audio segments from the video collection ("random"). For fair comparison, we prefilter out the silent frames for all methods (Sec. 3.1).

Dataset Corruption To measure the robustness of the methods to uncorrelated data, we corrupt the collection with uncorrelated audio examples. We do this by replacing a certain amount of audio segments by other audio segments randomly selected from a video set. We define the corruption ratio as the percentage of frames from the input video collection whose audio has been corrupted (also called percentage of uncorrelated visual-audio examples).

Evaluation Fig. 3a shows the overall classification rate on our dataset for the different approaches. It is measured as the number of correctly estimated classes over the number of tested frames: the estimated class is obtained by the classes associated to the $m = 5$ output audio segments and then majority vote; in case of tie, then random class.

Note that the aim of this paper is not to categorize images/videos into classes: here, we are using the classification rate to measure whether the retrieved sound corresponds to the input image by checking if the retrieved sound comes from the expected class in a relatively clean dataset.

First, the best classification rate when there is no uncorrelated data is around 70%. The gap between this rate and the ideal 100% accuracy is mainly due to similarity in the visual and audio descriptors: some classes share descriptor similarities in audio and/or visual space (e.g., trams and cars in a street environment). The classification rate of the random choice baseline starts at around 11% as excepted since the dataset contains 9 classes, and its classification rate continues diminishing when the percentage of uncorrelated data increases. "A+V" starts at around 52% and gives the lowest performance for a small percentage of uncorrelated audio. The performance of the unfiltered baseline starts relatively



(a) (b)

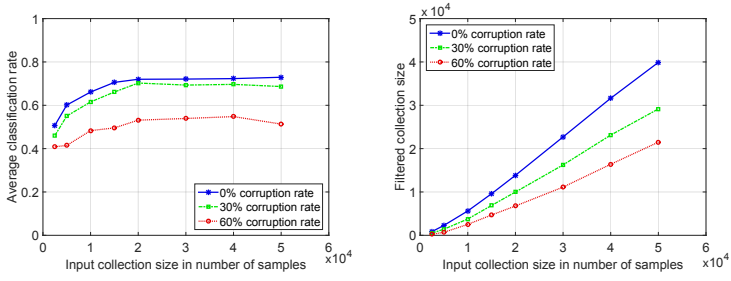
Fig. 3: Comparison of methods with respect to the amount of uncorrelated data. It indicates that different weighting strategies are appropriate for different contexts. The "V+A" strategy allows us to retrieve sounds of a query image reliably up to a high degree of corruption (see (a)). In contrast, the "A+V" strategy manages to better filter out noise overall (see (b)).

high at around 68% but quickly drops down when the percentage of uncorrelated data is higher than 20%.

The "V+A" strategy provides the best overall performance: it starts at around 70%, remains relatively stable up to about 50% of uncorrelated data, and then continues providing the highest performance up to about 65% of uncorrelated data. This suggests that this strategy is robust to large amounts of uncorrelated audio, and is a relevant method for such application context. For higher amount of data corruption, it is slightly outperformed by "A+V" and the classification rate of the filtered approaches drops to the level of the unfiltered baseline. With such an extreme amount of data corruption, the assumption that the dominant audio-visual correlation is the correct sound might not hold anymore for several classes, i.e., the very numerous uncorrelated examples might lead to another dominant but wrong correlation. For instance, this occurs in the example of the volcano videos with helicopter sounds that is discussed later.

Filtering Accuracy In addition, we also measure the quality of the filtered collection, i.e., what is the percentage of clean examples contained in the output filtered collection according to the percentage of uncorrelated data in the input collection. The results are shown in Fig. 3b. The best performance is obtained by the "A+V" strategy. For example even when the input collection contains 70% of uncorrelated audio (i.e., only 30% of the audio correlates to the visual signal), it manages to provide a filtered collection with 66% of correlated audio-visual examples. It shows that it can successfully identify and filter out the uncorrelated examples, and thus can provide a cleaner version of the input collection, even in presence of a high amount of uncorrelated data. "A+V" can then be considered the strategy of choice for the application context of obtaining a clean filtered collection.

It is true that some correlated examples also have been filtered out. We apply a conservative filter approach on purpose in order to increase the probability of obtaining a clean collection, i.e., composed of only (or at least mainly composed of) clean correlated audio. In turns, this allows us to retrieve correct sound for a query image. To further evaluate this, we conducted additional experiments compiled in Fig. 4. Fig. 4a suggests that the accuracy of our approach is rather stable with respect to the input collection size, once a certain amount of examples (around 20,000) is available. Fig. 4b illustrates that the filtered output collection size grows with respect to the input collection size in a dominant linear manner. On the whole this scalability analysis suggests that our approach can provide a larger (filtered) collection of correlated audio-visual examples from a larger input collection (potentially corrupted by uncorrelated audio data).



(a) (b)

Fig. 4: Evaluation of the classification rate (a) and output collection size (b) with respect to the input collection size (i.e., before filtering) with different rates of uncorrelated data. These experiments indicate that the accuracy (classification rate) is stable after a certain collection size (a) and the output collection size increases with the input collection size (b).

User Study For further evaluation, we also conducted a user study. 15 participants were shown an image and a 2-second audio segment, and then were asked to respond to the statement "This audio segment corresponds to the visual content of this image" by choosing a response from a five-point Likert scale: strongly agree (5), agree (4), neither agree nor disagree (3), disagree (2), or strongly disagree (1). We tested 10 (randomly selected) images per class, over the 9 classes of our dataset, which results in 90 images per participant. For each image, we prepared 3 different possible image-audio pairs for questioning by varying the audio track according to the following three methods: ground truth (i.e., the sound associated to this image in the clean dataset), "V+A" mutual kNN, and the unfiltered version. To mimic practical scenarios, we corrupted 60% of the audio samples of the training dataset for mutual kNN and the unfiltered version (the ground truth is untouched).

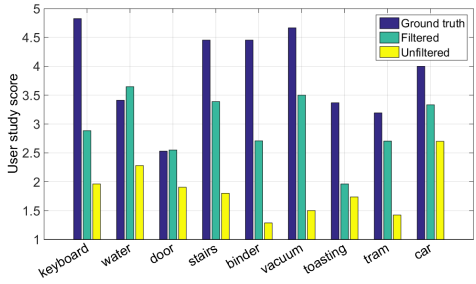


Fig. 5: Evaluation by user study. Overall, our audio-visual filtering approach successfully competes with method retrieving sound from the most visually similar video frames (see "unfiltered").

Fig. 5 shows that our filter approach is constantly better than the unfiltered approach overall, in agreement with Fig. 3a. 19% and 43% of the responses respectively obtained without and with our filtering step were "agree" or "strongly agree". It suggests that our filtering approach can deal with uncorrelated data and provide a better quality of the suggested sound.

Limitations and Future Work As demonstrated by the experiments, our approach is able to handle datasets containing uncorrelated audio and improves on naive classification-based solutions. Our approach can constitute a baseline in follow-up comparisons. Exciting research opportunities exist to further improve the performance for highly uncorrelated data. We assume that sounds that most often co-occur with specific objects are likely caused by the object visible. This assumption holds for many cases, but it can still fail when the majority of videos of an object contain a common uncorrelated sound. For example, when learning the sound of volcanoes, most of the videos we used were recorded from helicopters and contained a continuous helicopter sound in the background. Therefore our method learned to associate volcano images with helicopter sounds. Also, we intentionally applied a conservative approach to increase the probability of obtaining a clean collection. The downside is that it might limit the variety in the output collection and thus in the retrieved audio sets.

We additionally observed some limitations in our descriptors. For example, *trams* and *cars* videos were often confused due to their visual similarity (street scene). While we used existing feature descriptors trained for image classification, learning a descriptor designed for this specific task of audio retrieval could possibly improve result quality. Our visual descriptors are computed globally over the whole image. To deal with objects covering only a small part of the image, different kinds of descriptors should be used and/or in combination with object localization [57] or using saliency information [52]. This would enable, for example, an image to be associated with multiple different sounds, each derived from one or more objects in the scene. The user could also interactively specify which detected objects or which parts of the image to consider.

An extension of our work is to explore how to apply the audio-visual correlation for the converse target problem, that is given an audio segment, suggest pictures or videos. Beyond multimedia entertainment application, it could also be used to augment audio database with visual contents.

Our implementation allows the user to choose the duration of the suggested sounds (Sec. 3.3). Different sounds can have different durations, for example short toasting sound and longer passing car sound. Therefore an interesting research direction is to learn the duration of sounds and automatically output the sound with the appropriate duration.

An exciting direction is to investigate the use of motion descriptors [33,31] in addition to visual appearance descriptors. This would eventually permit to learn and cluster the different "motion sounds" of an object, for example the sounds of a door opening or closing, or the sounds of a car speeding up or slowing down.

Over the course of this work, we often wondered "what is a good sound corresponding to an input image?". In this paper, we used the notion of "significant" sounds [40,41] (i.e., sounds that are the most common for visually similar images) and conducted evaluation on several aspects of applications (Sec. 4.2). However, the answer to this question can be different for other applications. For example, one might rather be interested in finding "discriminative" sounds [20] (e.g., the hoot of an owl which is unique to that animal) or even "stereotype" sounds (e.g., an old creaking door which is potentially not accurately reflecting the daily life reality) [12,34]. These might be valid answers to the question and would require specific approaches to be explored in follow-up work.

5 Conclusion

In this paper, we investigated the problem of suggesting sounds for query images. Our approach takes a single image as input and suggests one or multiple audio segments issued from a video collection. One of the main challenges when solving this problem is the high amount of uncorrelated audio in the video collection. Therefore our main contribution is an approach to filter the data by using both audio and video modalities. The main goal of the filtering step is to filter out the audio segments which do not correlate with the image contents.

We conducted experiments that show that our filtering approach can successfully identify and filter out uncorrelated data, which in turn provides a filtered collection of correlated audio-visual examples. In addition to the application of sound suggestion from query images, this filtered clean collection could also be used as a knowledge prior for various tasks related to video classification or action recognition [29,7,61,64] or to build semantic audio database [63,46].

Moreover the user study results indicate that the sounds retrieved by our approach mainly correspond to the image content. Therefore we believe that our approach opens up new possibilities in the context of audio generation in accordance with visual content.

References

1. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building Rome in a day. In: ICCV (2009)
2. Anusuya, M.A., Katti, S.K.: Speech recognition by machine: A review. International Journal of Computer Science and Information Security (2009)
3. Averbuch-Elor, H., Wang, Y., Qian, Y., Gong, M., Kopf, J., Zhang, H., Cohen-Or, D.: Distilled collections from textual image queries. Computer Graphics Forum (EGSR) (2015)
4. Barzelay, Z., Schechner, Y.Y.: Harmony in motion. In: CVPR (2007)
5. Bazin, J.C., Malleson, C., Wang, O., Bradley, D., Beeler, T., Hilton, A., Sorkine-Hornung, A.: FaceDirector: continuous control of facial performance in video. In: ICCV (2015)
6. Berthouzoz, F., Li, W., Agrawala, M.: Tools for placing cuts and transitions in interview video. TOG (SIGGRAPH) (2012)
7. Brezeale, D., Cook, D.J.: Automatic video classification: A survey of the literature. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews (2008)
8. Brito, M., Chavez, E., Quiroz, A., Yukich, J.: Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. Statistics & Probability Letters (1997)
9. Cao, C., Bradley, D., Zhou, K., Beeler, T.: Real-time high-fidelity facial performance capture. TOG (SIGGRAPH) (2015)
10. Cardle, M., Brooks, S., Bar-Joseph, Z., Robinson, P.: Sound-by-numbers: motion-driven sound synthesis. In: SIGGRAPH/Eurographics Symposium on Computer Animation (SCA) (2003)
11. Chadwick, J.N., James, D.L.: Animating fire with sound. TOG (SIGGRAPH) (2011)
12. Chen, H., Gallagher, A.C., Girod, B.: What's in a name? First names as facial attributes. In: CVPR (2013)
13. Chu, W., Chen, J., Wu, J.: Tiling slideshow: An audiovisual presentation method for consumer photos. IEEE MultiMedia (2007)
14. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: International Conference on World Wide Web (2009)
15. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. ECCV Workshop on Statistical Learning in Computer Vision (2004)
16. Dannenberg, R., Neuendorffer, T.: Sound synthesis from real-time video images. In: International Computer Music Conference (ICMC) (2003)
17. Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G.J., Durand, F., Freeman, W.T.: The visual microphone: passive recovery of sound from video. TOG (SIGGRAPH) (2014)
18. Dieleman, S., Schrauwen, B.: Multiscale approaches to music audio feature learning. In: International Society for Music Information Retrieval Conference (ISMIR) (2013)
19. van den Doel, K., Kry, P.G., Pai, D.K.: FoleyAutomatic: physically-based sound effects for interactive simulation and animation. In: SIGGRAPH (2001)
20. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes Paris look like Paris? TOG (SIGGRAPH) (2012)
21. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Learning (ICML) (2014)

22. Dunker, P., Popp, P., Cook, R.: Content-aware auto-soundtracks for personal photo music slideshows. In: ICME (2011)
23. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The PASCAL visual object classes challenge: A retrospective. IJCV (2015)
24. Fried, O., Fiebrink, R.: Cross-modal sound mapping using deep learning. In: International Conference on New Interfaces for Musical Expression (NIME) (2013)
25. Galasso, F., Nagaraja, N.S., Cardenas, T.J., Brox, T., Schiele, B.: A unified video segmentation benchmark: Annotation, metrics and analysis. In: ICCV (2013)
26. Godsill, S., Rayner, P., Cappé, O.: Digital audio restoration. Springer (2002)
27. Izadinia, H., Saleemi, I., Shah, M.: Multimodal analysis for identification and segmentation of moving-sounding objects. IEEE Transactions on Multimedia (2013)
28. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R.B., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia (2014)
29. Karpathy, A., Toderici, G., Shetty., S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
30. Kidron, E., Schechner, Y.Y., Elad, M.: Pixels that sound. In: CVPR (2005)
31. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: BMVC (2008)
32. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T.A., Serre, T.: HMDB: A large video database for human motion recognition. In: ICCV (2011)
33. Laptev, I.: On space-time interest points. IJCV (2005)
34. Lea, M.A., Thomas, R.D., Lamkin, N.A., Bell, A.: Who do you look like? Evidence of facial stereotypes for male names. Psychonomic Bulletin and Review (2007)
35. Lee, H., Pham, P.T., Largman, Y., Ng, A.Y.: Unsupervised feature learning for audio classification using convolutional deep belief networks. In: NIPS (2009)
36. Li, C., Shan, M.: Emotion-based impressionism slideshow with automatic music accompaniment. In: International Conference on Multimedia (2007)
37. Liao, Z., Yu, Y., Gong, B., Cheng, L.: AudeoSynth: music-driven video montage. TOG (SIGGRAPH) (2015)
38. Lin, Y., Tsai, T., Hu, M., Cheng, W., Wu, J.: Semantic based background music recommendation for home videos. In: International Conference on MultiMedia Modeling (MMM) (2014)
39. Logan, B.: Mel frequency cepstral coefficients for music modeling. In: International Symposium on Music Information Retrieval (2000)
40. Maier, M., Hein, M., von Luxburg, U.: Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters. Theoretical Computer Science (2009)
41. Maier, M., Hein, M., von Luxburg, U.: Cluster identification in nearest-neighbor graphs. In: International Conference on Algorithmic Learning Theory (ALT) (2007)
42. Matta, S., Kumar, D., Yu, X., Burry, M.: An approach for image sonification. In: International Symposium on Control, Communications and Signal Processing (2004)
43. Matthews, I., Cootes, T.F., Bangham, J.A., Cox, S.J., Harvey, R.: Extraction of visual features for lipreading. TPAMI (2002)
44. Meijer, P.: An experimental system for auditory image representations. IEEE Transactions on Biomedical Engineering (1992)
45. Mermelstein, P.: Distance measures for speech recognition: Psychological and instrumental. In: Pattern Recognition and Artificial Intelligence (1976)

46. Miotto, R., Lanckriet, G.R.G.: A generative context model for semantic music annotation and retrieval. IEEE Transactions on Audio, Speech & Language Processing (2012)
47. Nayak, M., Srinivasan, S., Kankanhalli, M.: Music synthesis for home videos: an analogy based approach. In: IEEE Pacific Rim Conference on Multimedia (PCM) (2003)
48. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: International Conference on Machine Learning (ICML) (2011)
49. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV (2001)
50. Owens, A., Isola, P., McDermott, J., Torralba, A., Adels, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR (2016)
51. Paulus, J., Müller, M., Klapuri, A.: Audio-based music structure analysis. In: International Conference on Music Information Retrieval (2010)
52. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: CVPR (2012)
53. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: CVPR Workshop (2014)
54. Ren, Z., Yeh, H., Lin, M.C.: Example-guided physically based modal sound synthesis. TOG (SIGGRAPH) (2013)
55. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: CVPR (2015)
56. Rubin, S., Berthouzoz, F., Mysore, G., Li, W., Agrawala, M.: UnderScore: musical underlays for audio stories. In: ACM UIST (2012)
57. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: ImageNet large scale visual recognition challenge. IJCV (2015)
58. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations (2014)
59. Shi, J., Malik, J.: Normalized cuts and image segmentation. TPAMI (2000)
60. Simon, I., Snavely, N., Seitz, S.M.: Scene summarization for online image collections. In: ICCV (2007)
61. Soomro, K., Zamir, A.R., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild. CRCV-TR-12-01 (2012)
62. Stupar, A., Michel, S.: Picasso - to sing, you must close your eyes and draw. In: International Conference on Research and Development in Information Retrieval (SIGIR) (2011)
63. Turnbull, D., Barrington, L., Torres, D.A., Lanckriet, G.R.G.: Semantic annotation and retrieval of music and sound effects. IEEE Transactions on Audio, Speech & Language Processing (2008)
64. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV (2013)
65. Wenner, S., Bazin, J.C., Sorkine-Hornung, A., Kim, C., Gross, M.: Scalable music: Automatic music retargeting and synthesis. CGF (Eurographics) (2013)
66. Wu, X., Li, Z.: A study of image-based music composition. In: ICME (2008)
67. Zhao, G., Barnard, M., Pietikäinen, M.: Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia (2009)
68. Zheng, C., James, D.L.: Harmonic fluids. TOG (SIGGRAPH) (2009)
69. Zheng, C., James, D.L.: Rigid-body fracture sound with precomputed soundbanks. TOG (SIGGRAPH) (2010)