# Audiovisual annotation procedure for multi-view field recordings

Patrice Guyot, Thierry Malon, Geoffrey Roman Jimenez, Sylvie Chambon, Vincent Charvillat, Alain Crouzil, André Péninou, Julien Pinquier, Florence Sèdes, Christine Sènac

# Audiovisual Annotation Procedure for Multi-view Field Recordings

Patrice Guyot(✉), Thierry Malon, Geoffrey Roman-Jimenez, Sylvie Chambon, Vincent Charvillat, Alain Crouzil, André Péninou, Julien Pinquier, Florence Sèdes, and Christine Sénac

IRIT, Université de Toulouse, CNRS, Toulouse, France
{patrice.guyot,thierry.malon,geoffrey.roman-jimenez,
sylvie.chambon,vincent.charvillat,alain.crouzil,andre.peninou,
julien.pinquier,florence.sedes,christine.senac}@irit.fr

**Abstract.** Audio and video parts of an audiovisual document interact to produce an audiovisual, or multi-modal, perception. Yet, automatic analysis on these documents are usually based on separate audio and video annotations. Regarding the audiovisual content, these annotations could be incomplete, or not relevant. Besides, the expanding possibilities of creating audiovisual documents lead to consider different kinds of contents, including videos filmed in uncontrolled conditions (i.e. fields recordings), or scenes filmed from different points of view (multi-view). In this paper we propose an original procedure to produce manual annotations in different contexts, including multi-modal and multi-view documents. This procedure, based on using both audio and video annotations, ensures consistency considering audio or video only, and provides additionally audiovisual information at a richer level. Finally, different applications are made possible when considering such annotated data. In particular, we present an example application in a network of recordings in which our annotations allow multi-source retrieval using mono or multi-modal queries.

**Keywords:** Audiovisual · Annotation · Multi-view · Multi-modal
Field recording · Multimedia · Ground truth

## 1 Introduction

Production of audiovisual documents is a fast-growing phenomenon which is founded on an increasing number of recording devices, for instance smartphones. In comparison to the data conceived in a controlled domain (e.g. TV, radio, music studio, motion capture studio, etc.), many recordings are generally produced in an uncontrolled context. They will be further referred to as *field recordings*.

Moreover, different audiovisual documents may correspond to the same scene, for instance a public event that is filmed by different points of view. These multi-view scenes contain lots of information and provide new opportunities for high-level automatic queries.

https://doi.org/10.1007/978-3-030-05710-7_33

In the context of automatic analysis, the aim of the different tasks (e.g. detection, classification) is to reduce the quantity of information embedded in audiovisual documents towards some particular semantic concept. For example, a video with a car in the foreground contains lots of information (type of car, ground, objects in the background, weather, localization, etc.) that could be reduced to the concepts *car* or *nice weather*.

In order to produce a model and to evaluate the performances of the algorithms on a set of data, researchers generally build a manual annotation that expresses this semantic information. The result of such manual annotation is generally called *ground truth*. As it usually refers to information provided by direct observation, it requires researchers to develop objective criteria. The ground truth depends on the definition of a space in which the data are projected in the most appropriate manner within a specific context. This task is not always straightforward: for instance, in the context of Music Information Retrieval, the evaluation of *musical artist similarity* requires the development of an objective measurement, meanwhile artist similarity relies on an elusive concept [5]. Thus, it appears that the term *ground truth* is sometimes misleading because it does not reflect an objective *truth* [1]. In that respect, we will use the term *reference* which seems more accurate to designate the manual annotations.

Audiovisual documents are in essence based on two modalities: audio and video. Yet in the context of audiovisual documents, the annotations are generally mono-modal (audio or video), while the perception of an audiovisual content is multi-modal and thus leads to a richer interpretation. Moreover, the different modalities influence each other, making the mono-modality annotation difficult in a multi-modality context.

This paper addresses the issue of producing multi-modal annotations in an audiovisual context. We propose a low-cost procedure to manually annotate multi-view field recordings. This paper is organized as follows. We first present the relative works about annotation and perception of audiovisual contents. Section 3 presents a specific procedure to solve the multi-modal issues of audiovisual annotations. This procedure is usable in mono or multi-view contexts. Finally, different applications of this procedure are described in Sect. 4.

## 2 Related Works

### 2.1 Audio and Video Ground Truth: From Precise to Weak Annotations

The challenge of multimedia modeling, developed intensively during the 2000s, has produced multiple campaigns for information retrieval, for instance with video [20] or audio events [15]. In this framework, vast amount of data have been manually annotated. These annotations are usually precise and time consuming. For example, audio events, such as speaker turns in case of speaker diarisation, or music notes in the case of Music Information Retrieval, are usually annotated at a millisecond scale [3]. Moreover, as these annotations are hardly objective, an

agreement between annotators is usually needed [21]. For these different tasks of annotation, different softwares have been proposed (see [19] for a comparison).

Lately, Deep Learning based approaches [9] outperform the state of the art in many domains. However, they require a large amount of data. Because a precise annotation of these data is almost impossible, recent datasets include only *weak annotations*. These weak annotations really differ from a precisely annotated ground truth, as they may be incomplete, not relevant and heterogeneous. For example, the Audioset dataset [6] provides a large set of audio data extracted from videos, but the annotations were tagged by YouTube users on the audiovisual content. In the area of vision, the AVA dataset [7] provides precise spatio-temporal annotations of persons conducting actions, but the sound of the video is not taken into account.

Finally, most of the research works are usually mono-modal based (only audio or video stream). The issue of merging audio and visual information to richer concepts is rarely addressed by the different scientific communities. In that scope, the softwares used for manual annotations seem to deal with multi-modality as a juxtaposition of mono-modal annotations.

### 2.2 Multi-modal and Multi-view

The modeling of multi-modal (or cross-modal) inputs is very challenging, for example when studying discourses containing speech and non-linguistic signs [8]. Some applications rely on a precise interaction between image and sound. For instance, the detection of talking heads has been addressed [11]. In this context, various works deal with the fusion of audio and video modalities, for example with early, intermediate or late fusion [18]. Besides, other applications deal with other modalities, for instance image and texts [17].

The issue of annotating a multi-modal dataset in the case of audiovisual content is clearly addressed in [10]. Whereas this study aims at automatically detecting overt aggression in public places, the authors state that "problems with automatically processing multi-modal data start already from the annotation level". The complexity of the interactions between modalities forced the authors to produce three different types of annotations: audio, video, and multimodal. The combination of these three annotations increase the performances of an automatic detector based on a machine learning approach. However, the processing of these annotations is time-consuming and sensitive. Firstly, this procedure necessitates at least three different kinds of playbacks (audio, video and audiovisual) to perform the annotation. Secondly, in order to process independent annotations with limited influence among modalities, three different annotators at least are required.

Furthermore, increasing amounts of scenes are filmed simultaneously from different points of view. In particular, in the context of video surveillance, different cameras are usually used [22]. The framework of Motion Capture also provides interesting databases that include different views [16]. The reflective markers placed on a human body allow the recording of the absolute position of each part of the body, which can be directly used as a reference. However,

these applications usually remain in the field of laboratory studies and are hard to deploy in a real-life context.

The context of field recordings is generally more challenging [2] due to the number of overlapping events and objects. Considering audio, many events overlap and produce a mixture. Moreover, the movement of the audio sources (for instance a passing car) makes it difficult to position the starting and ending boundaries of the events. Same kind of difficulties arise considering images, with occlusion, superposition, illumination and size of objects.

Different works review datasets from the perspective of multi-modal and multi-view features [12,13,16]. However, as observed in [13], these datasets are often limited by different criteria including presence of audio, realism for real life applications, and number of overlapping and disjoint views.

### 2.3 Audio-Vision

The relationship between sound and image has been investigated for a long time, in particular in the context of cinema. A reference book [4] details the different possibilities of using sounds in videos.

Focusing on the area where the action takes place, the first distinction has to be drawn between sounds of the scene that could be heard by the film's characters, and sounds that could not. The first category is called *diegetic* sounds. The second category consists of non-diegetic sounds that are added in a post-production step, for example in the case of voice-over. More precisely, the diegetic sound source can be on-screen or off-screen. Furthermore, the source of the sound can be at times visualized in the image. Otherwise, if the sound source is not visible, the sound is called *acousmatic*. Figure 1 summarizes these different interactions.

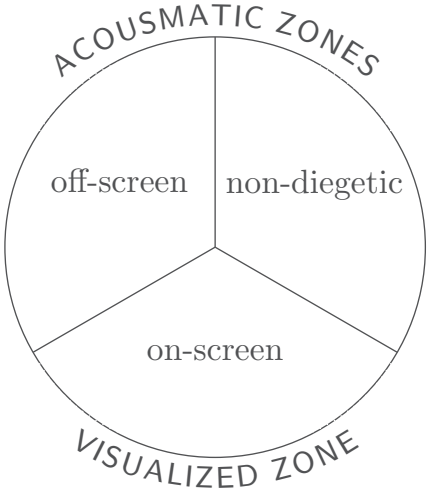


**Fig. 1.** The audiovisual scene (adapted from [4]).

In a multi-view sequence, the source of a sound may be either visualized or off-screen according to the different points of view. If the source of a sound is

ambiguous or not visible in the current video, a different viewpoint may disclose it. We speak about *causal identification* when the source of a sound can be identified, whether it is visible or not in the current viewpoint.

As these different types of interaction are precisely depicted in a movie script, they are quite unusual in research papers. To our knowledge, research datasets do not provide information about on-screen and off-screen sounds. However, it seems that these different types of interaction have an influence on the perception and the understanding of the audiovisual content.

Finally, the audio and video parts interact in different ways to create an audiovisual perception. One of the clearest examples of the influence between audio and video lies in the McGurk effect [14], that demonstrates an interaction between hearing and vision in speech perception. Its most known implementation consists of a video of a human face saying a pseudo-word (*ga-ga*) with a voice over saying another one (*ba-ba*), leading to the perception of a third one (*da-da*). When annotating, this kind of phenomenon could occur in the same way and would lead to three different annotations (audio, video and audiovisual).

## 3 Audio/visual Annotation

### 3.1 Problematic

Audio and video annotations are usually based on different paradigms, but are both based on predefined categories to annotate, such as *car* or *speech*. Audio annotations usually consist in determining the start and the end of audio events and tag each event with a category (*engine noise*, *speech*, *horn sound*, etc.). Considering video, a usual procedure of annotation is to set a bounding box on each object of interest in each frame of the video and tag each object with some categories (*car*, *person*, *clothes*, etc.).

Procedures of annotation usually consider the audio and video streams as if they were disconnected and each media is annotated separately. In that process, a valuable information may be lost. In this article, we argue that the whole information embedded in an audiovisual content is greater than the sum of its audio and video parts. For example, if we separately annotate *speech* events (audio only) and *person* objects (video only), we cannot deduce if a visible person is the speaker or not.

In that context, some issues are clearly observable with the Audioset dataset (see Sect. 2.1). Most of the tags seem to have been set according to the video part, which usually dominates the audiovisual content. As a consequence, a video of a cat annotated as *cat* will also be annotated *cat* in the audio annotation, even if the cat remains silent in the video. To address these issues, we intend to merge the audio and video modalities into audiovisual objects. Practically, we aim to create an audiovisual object based on a moving bounding box and a corresponding audio event. Surprisingly, this task proved to be very difficult and many issues appeared and are detailed below.

A first challenge is about matching and merging one visual object and one audio event. First of all, we have considered a systematic fusion of events from the

two modalities considering that this fusion could match segments from audio and video streams in the case of temporal overlapping. Unfortunately, this matching may introduce some wrong annotations when the audio annotation corresponds to an off-screen source. For example, Fig. 2 shows a car at the foreground. At the same moment the sound track is overpowered by an off-screen motorbike.


**Fig. 2.** Image bounding boxes around the cars. If a passing car is clearly visible at the foreground, a motorbike behind the camera overpowers the corresponding soundtrack.

A second challenge lies in the case of defining several annotations with temporal overlapping. The context of field recordings induces an audio mixture. Depending on his expertise, a human annotator may not be able to set precisely the starting and ending boundaries of the different audio events of this mixture. In this case, the matching between a specific audio event and the potential corresponding visual object can be impossible. In the same way, considering visual annotations, when annotating a group of objects, the annotator might be unable to draw bounding boxes around each element. Depending on the scale of the image or the mixing of objects in the image, the annotator may annotate each element separately, the entire group as a single object, or a mixture of single elements and rest of the group.

In this context, the issue consists of matching several audio and visual annotations with ensuring their relevancy. When many visual objects may have produced some sound events, the separation of the sound sources in the audio signal may be impossible. Let us consider the Fig. 3 that represents audio segments (time boundaries) and video annotations. In that scene, the passing of two consecutive vehicles have been auditory annotated as a single audio event. The solution for creating an audiovisual object from these annotations relies on the segmentation of the audio event in two parts to create two audiovisual objects. We have tested many possibilities to obtain the boundaries of the audio events but none of them was satisfying whatever the situation.

In a more sophisticated way, we could directly build audio, video and audio-visual annotations from the audiovisual stream. However, the completion of this task is not straightforward. Indeed, the influence of the audiovisual content may influence the annotation of the mono-modal streams. For instance, an annotator would more likely create an audio event for a moving car than for a stopped one, even if they both produce a motor noise.

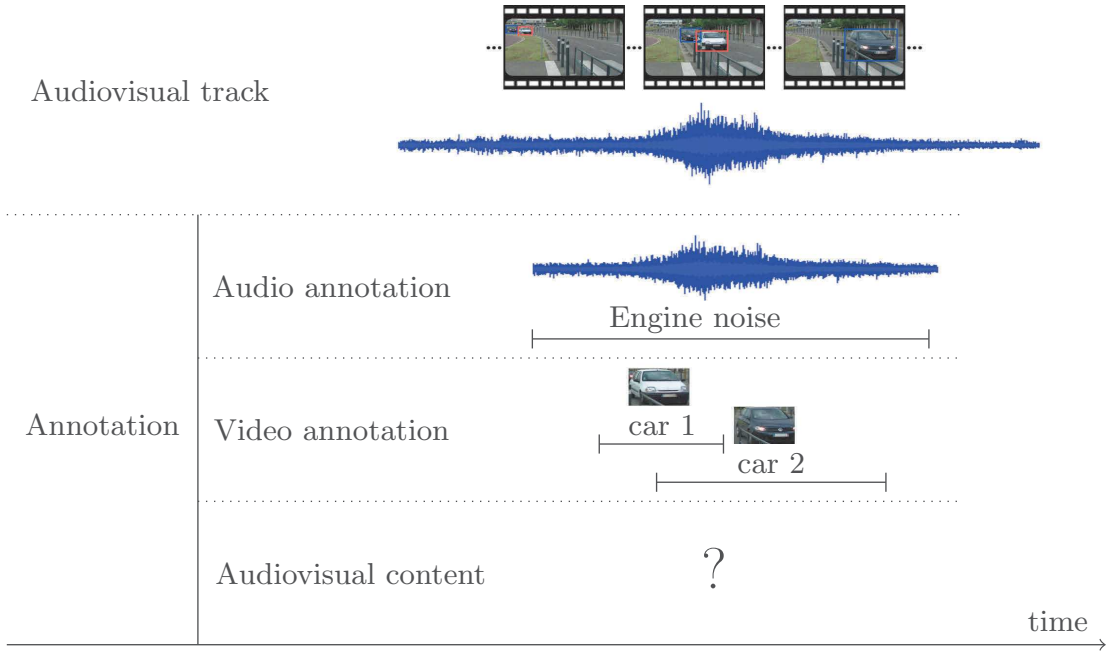


**Fig. 3.** Audiovisual annotation of the passing of two cars. In the audio modality, the passage of cars is heard as a unique lengthy sound. On the contrary, the video annotation clearly exhibits two different vehicles. Consequently, the automatic fusion of these two modalities to create audiovisual object(s) is very difficult to define.

### 3.2 Procedure of Annotation

We present here a procedure to obtain audio and visual annotations, as well as audiovisual information. It aims at satisfying the following goals:

- Audiovisual added value: the annotations must embed multi-modal information that allows a better understanding of the scene and an added value in comparison of the whole set of mono-modal annotations.
- Mono-modal used: the audio and visual annotations must be usable in a mono-modal context. Therefore, additional information from other modality are not to be considered when creating mono-modal annotations.
- Low additional cost: the audiovisual annotation must be objective and straightforward, and must not generate a heavy additional cost.

To address these different constraints, we propose the following two-steps protocol, which is designed to be processed manually.

**Step 1: Mono-Modal Annotations.** In this step, the audio and video annotations are processed separately. Optimally, the annotations have to be processed by different persons, without access to the other modality. For example, the annotator of the audio stream works only with audio. These two annotations can be processed in parallel.

For each modality, a unique identifier is set for each object in the scenes. The objects visible at different moments of the video (or on different videos in a multi-view context) must bear the same identifier. Similarly, the same identifier is set for each annotation of the same audio event in the case of a clearly unique event, for example a big explosion recorded by various devices.

At the end, an audio annotation contains description of audio events that are made up of *time boundaries*, *categories* and *identifier*. Visual annotations allow the description of objects on the basis of *time*, *spatial coordinates of bounding boxes*, *categories*, and *identifier*.

**Step 2: Multi-modal Links.** In a second step, audio and visual modalities are linked with each other. Links between audio and video identifiers are created in case of causal identification (see Sect. 2.3). In a multi-view case, an audio event could be associated with an off-screen object that is visible on another view. This process is detailed in next section.

In this step, the mono-modal annotations (audio or video) cannot be modified regarding the other modality, even if they appear to be wrong in the multi-modal context (see the McGurk effect in Sect. 2.3). These annotations were valid from a mono-modal annotation point of view and remain as they stand.

### 3.3 Implementation of Multi-modal Links

Considering the mono-modal annotations, we focus on audio and video annotations that temporally overlap. These annotations may refer to the same audio-visual document, or to different documents in the multi-view case.

Each of the audio annotations is considered in terms of sound source. If a causal identification is possible (see Sect. 2.3), we link audio to video annotations. A link means that audio annotations are enriched with the list of the linked visual objects considered as the source of the audio event. When an audio annotation is linked to several visual objects, the sources of the audio event can be all of the objects or some of them indifferently.

Table 1 summarizes the different annotation links between audio and video. Note that we only link audio event to video object (not video object to audio event) because of the unbalanced relationship between audio and video.

We detail below some concrete examples of links between audio and video annotations. In these examples, we focus on vehicles. However, as our procedure is generic, it can be applied on different kinds of events and objects.

*Passing Vehicle:* the audio and video events are linked if they undoubtedly originate from the same vehicle. If any doubt exists, for instance if the source of

**Table 1.** Annotation link procedure depending on the presence of audio and video annotations and the possibility of causal identification. $A$ corresponds to an annotation of a single audio event (e.g. engine noise, speech, etc.). $V$ corresponds to an annotation of a single visual object (e.g. car, person, etc.). $\{A_i\}$ corresponds to a set of audio annotations. $\{V_j\}$ corresponds to a set of video annotations. A link between annotations is denoted by $\rightarrow$.

| | (1) | (2) | (3) | | (4) | | (5) | | (6) | |
|---|---|---|---|---|---|---|---|---|---|---|
| Audio annotation | $\{A_i\}$ | $\{\emptyset\}$ | $A$ | | $A$ | | $\{A_i\}$ | | $\{A_i\}$ | |
| Video annotation | $\{\emptyset\}$ | $\{V_j\}$ | $V$ | | $\{V_j\}$ | | $V$ | | $\{V_j\}$ | |
| Causal identification | No | No | No | Yes | No | Yes | No | Yes | No | Yes |
| Annotation link | — | — | — | $A \rightarrow V$ | — | $A \rightarrow \{V_j\}$ | — | $\{A_i \rightarrow V\}$ | — | $\{A_i \rightarrow \{V_j\}\}$ |

the audio event could be another vehicle that is not visible, the events are not linked (see Table 1 column 3 and Fig. 2).

*Slammed Door:* if a sound event occurs from the interaction of several visually annotated objects, we link the audio event to the each visual objects (see Table 1 column 4). For instance, in case of the closure of a car door with annotations for two objects (car and person), the audio event slammed door is linked to each of the two objects.

*Passing Vehicle and Horn:* in the case of multiple audio events that obviously originate from the same visual object, we link all audio events to the object (see Table 1 column 5). Thus, if an object *car* has been annotated visually and two audio events *engine* and *horn* are produced by the car, then the two audio events are linked to the visual object.

*Passing of Multiple Vehicles:* in the case of multiple vehicles passing with a different number of audio events (see Fig. 3), we link the audio events to all visual objects (see Table 1 column 6). However, if the audio source is not obvious (for instance a car horn when different vehicles are present), we do not link the audio event to any visual object.

## 4 Applications

We present hereafter different applications that are made possible by our procedure of annotation. In the case of mono-modal request, the annotated corpus can be used for different purposes. The audio annotations can be used in audio detection tasks (see [15] for examples). Similarly, the video annotations provide a framework for objects detection (see [20] for example). Using the bounding

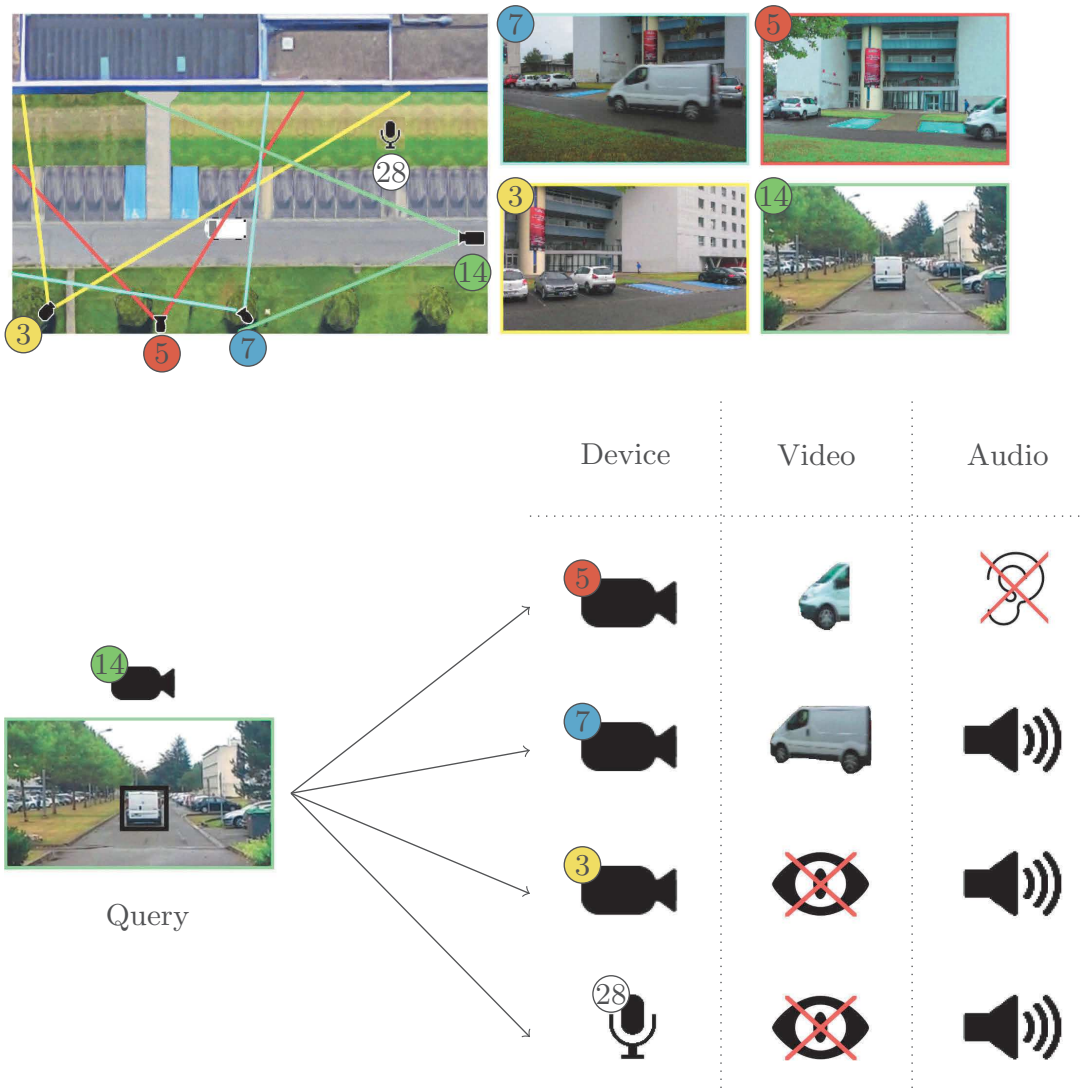


**Fig. 4.** Within a network of recording devices, our multi-modal annotation procedure allows to retrieve either visual objects only (camera 5), multi-modal objects (camera 7), or audio events only (camera 3, microphone 28) from audio or video queries. Note that camera 3 records audio and video, but the audible object is off-screen.

boxes drawn on each object, object re-identification based on image appearance can be driven.

In a context of surveillance with a network of recording devices (cameras recording video, microphones recording audio, smart-phones recording both video and audio...), our annotations allow users to perform different kinds of requests. Figure 4 illustrates this application in the context of the ToCaDa dataset [13]. Several devices are set around a scene: devices 3 and 7 record both audio and video, whereas devices 5 and 14 only record video stream. Finally, microphone 28 only records audio. From an audiovisual document, we may perform queries that can be either video only (for example by clicking the bounding box containing the vehicle on the video from camera 14) or audio only (for example by clicking on the represented audio event from the same video) in order to

retrieve the object ID. All the audio events and video objects associated to the same ID are returned as results. These results can either be audio, visual, or audiovisual.

In a more complex application, this framework also allows multi-modal queries that aim to retrieve audiovisual objects, for example a vehicle with distinct sound and appearance.

## 5 Conclusion

In this paper, we propose a simple procedure to produce audiovisual annotations in different contexts such as multi-view dataset. Our approach aims to produce audio, visual, and audiovisual information. It is based on separate annotations on the audio and video modalities, followed by an audiovisual matching. In this way, an audiovisual annotation is produced, as well as audio and video annotations that remain relevant in a mono-modal context.

This procedure is simple. With respect to mono-modal annotations, our method does not extend the time of processing significantly. It can be deployed at a large scale, but, unlike *weak annotations*, maximizes the relevance of the annotation. Moreover, in the context of multi-view annotations, the required uniqueness of annotation identifiers allows for creating possibly relevant annotations not only with on-screen objects but also with off-screen objects. Finally, the resulting annotations produce a valuable approximation of what should be a *ground truth*.

## References


1. Aroyo, L., Welty, C.: Truth is a lie: crowd truth and the seven myths of human annotation. AI Mag. **36**(1), 15–24 (2015)
2. Auer, E., et al.: Automatic annotation of media field recordings. In: ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010), pp. 31–34. University de Lisbon (2010)
3. Bird, S., Liberman, M.: A formal framework for linguistic annotation. Speech Commun. **33**(1–2), 23–60 (2001)
4. Chion, M.: Audio-Vision: Sound on Screen. Columbia University Press, New York (1994)
5. Ellis, D.P., Whitman, B., Berenzweig, A., Lawrence, S.: The quest for ground truth in musical artist similarity. In: ISMIR, Paris, France (2002)
6. Gemmeke, J.F., et al.: Audio set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780. IEEE (2017)
7. Gu, C., et al.: AVA: a video dataset of spatio-temporally localized atomic visual actions. CoRR abs/1705.08421 (2017)
8. Iedema, R.: Multimodality, resemiotization: extending the analysis of discourse as multi-semiotic practice. Vis. Commun. **2**(1), 29–57 (2003)
9. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)

10. Lefter, I., Rothkrantz, L.J.M., Burghouts, G., Yang, Z., Wiggers, P.: Addressing multimodality in overt aggression detection. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS (LNAI), vol. 6836, pp. 25–32. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23538-2_4
11. Li, D., Dimitrova, N., Li, M., Sethi, I.K.: Multimedia content processing through cross-modal association. In: Proceedings of the Eleventh ACM International Conference on Multimedia, pp. 604–611. ACM (2003)
12. Liu, A.A., Xu, N., Nie, W.Z., Su, Y.T., Wong, Y., Kankanhalli, M.: Benchmarking a multimodal and multiview and interactive dataset for human action recognition. IEEE Trans. Cybern. **47**(7), 1781–1794 (2017)
13. Malon, T., et al.: Toulouse campus surveillance dataset: scenarios, soundtracks, synchronized videos with overlapping and disjoint views (regular paper). In: ACM Multimedia Systems Conference (MMSys), Amsterdam, 12 June 2018–15 June 2018. ACM Multimedia Systems, June 2018
14. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature **264**(5588), 746 (1976)
15. Mesaros, A., et al.: DCASE 2017 challenge setup: tasks, datasets and baseline system. In: DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events (2017)
16. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley MHAD: a comprehensive multimodal human action database. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV), pp. 53–60. IEEE (2013)
17. Pereira, J.C., et al.: On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 521–535 (2014)
18. Pinquier, J., et al.: Strategies for multiple feature fusion with hierarchical hmm: application to activity recognition from wearable audiovisual sensors. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 3192–3195. IEEE (2012)
19. Rohlfing, K., et al.: Comparison of multimodal annotation tools-workshop report. Gesprächforschung-Online-Zeitschrift zur Verbalen Interaktion **7**, 99–123 (2006)
20. Russakovsky, O., et al.: ImageNet Large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015)
21. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G.: Semantic annotation and retrieval of music and sound effects. IEEE Trans. Audio Speech Lang. Process. **16**(2), 467–476 (2008)
22. Wang, X.: Intelligent multi-camera video surveillance: a review. Pattern Recogn. Lett. **34**(1), 3–19 (2013)