# Toulouse Campus Surveillance Dataset: scenarios, soundtracks, synchronized videos with overlapping and disjoint views

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

# Toulouse Campus Surveillance Dataset: scenarios, soundtracks, synchronized videos with overlapping and disjoint views

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

IRIT, Université de Toulouse, CNRS, Toulouse, France

{first name}.{last name}@irit.fr

## ABSTRACT

In surveillance applications, humans and vehicles are the most important common elements studied. In consequence, detecting and matching a person or a car that appears on several videos is a key problem. Many algorithms have been introduced and nowadays, a major relative problem is to evaluate precisely and to compare these algorithms, in reference to a common ground-truth. In this paper, our goal is to introduce a new dataset for evaluating multi-view based methods. This dataset aims at paving the way for multidisciplinary approaches and applications such as 4D-scene reconstruction, object identification/tracking, audio event detection and multi-source meta-data modeling and querying. Consequently, we provide two sets of 25 synchronized videos with audio tracks, all depicting the same scene from multiple viewpoints, each set of videos following a detailed scenario consisting in comings and goings of people and cars. Every video was annotated by regularly drawing bounding boxes on every moving object with a flag indicating whether the object is fully visible or occluded, specifying its category (human or vehicle), providing visual details (for example clothes types or colors), and timestamps of its apparitions and disappearances. Audio events are also annotated by a category and timestamps.

## KEYWORDS

Video and audio dataset, forensics, multi-source, synchronization, annotations, overlapping fields of view, disjoint fields of view.

Thierry Malon, Geoffrey Roman-Jimenez, Patrice Guyot, Sylvie Chambon, Vincent Charvillat, Alain Crouzil, André Péninou, Julien Pinquier, Florence Sèdes, Christine Sénac IRIT, Université de Toulouse, CNRS, Toulouse, France
{first name}.{last name}@irit.fr

https://doi.org/10.1145/3204949.3208133

## 1 INTRODUCTION

Cameras used in video surveillance are of multiple types and models. Thus, they may or may not include soundtrack, GPS coordinates, timestamps, and can present significant differences in pixel resolution and image quality. Using multiple cameras for surveillance is quite unusual, and, when several cameras are used, the fields of view of the different cameras may or not overlap. As analyzing multiple videos simultaneously can bring more relevant information about the scene, there has been an increasing interest, in particular in the domain of surveillance, in developing pattern recognition tools enabling to automatically extract and to summarize all the relevant information in a user-friendly form. In order to have a better understanding of the scene, it can be useful to recognize and to track an object (human or vehicle) in different views over time. Numerous algorithms have been proposed in the literature to perform automatic object detection and tracking across a video sequence [3, 11, 14]. However, in the case of a scene filmed by multiple cameras across several places (streets, train stations, buildings), global object tracking remains a challenging task because of the possible multiple points of view and the configurations of the cameras (video resolution, light exposition, color). Besides, audio surveillance is a very active field [5], proposing numerous algorithms for detection of single or compound audio events (glass breaking, motor noise, explosion, aggression) that could be used in complementarity with video to recognize a given object. More generally, the use of multiple sources of information (video, soundtrack, GPS locations) would help the automatic recognition over time of visual, audio or audio-visual objects and their tracking across various places.

During the last decade, public datasets become more and more available, helping for evaluation and comparison of algorithms and so, contributing to improvements of human and vehicle detection and tracking. However, most of the datasets focus on a specific task and do not allow to evaluate approaches mixing multiple sources of information. Only few datasets provide synchronized videos with overlapping fields of view and rarely provide more than 4 different views while more and more approaches can benefit from more views. Moreover, soundtracks are almost never provided while they are a rich source of information as voices and motor noises can help to recognize, respectively, a person or a car.

| Properties | 3DPeS [2] | VIRAT [13] | MuHAVi [17] | Human3.6M [9] | Proposed dataset |
|---|---|---|---|---|---|
| # of cameras | 8 static | 16 static | 8 static | 4 static | 25 static |
| Soundtrack | No | No | No | No | Yes |
| # of microphones | 0 | 0 | 0 | 0 | 25+2 |
| Overlapping FOV | Very partially | 2+2 | 8 | 4 | 17 |
| Disjoint FOV | 8 | 12 | 0 | 0 | 4 |
| Synchronized | No | No | Partially | Yes | Yes |
| Pixel resolution | $704 \times 576$ | $1920 \times 1080$ | $720 \times 576$ | $1000 \times 1000$ | Mostly $1920 \times 1080$ |
| # visual objects | 200 | Hundreds (N/A) | 14 | 11 | 30 |
| # action types | 0 | 23 | 17 | 15 | 0 |
| # bounding boxes | 0 | $\approx$ 1/object/second | 0 | $\approx$ 1/object/frame | $\approx$ 1/object/second |
| In/Outdoor | Outdoor | Outdoor | Indoor | Indoor | Outdoor |
| With scenario | | | $\times$ | $\times$ | $\times$ |
| Realistic | $\times$ | $\times$ | | | $\times$ |

**Table 1: Comparison of the properties of several datasets.**

In consequence, we produced a large dataset composed of synchronized videos of the same scene recorded from multiple viewpoints with both overlapping and non-overlapping fields of view. Soundtracks are also included. The dedicated applications of this dataset are objects detection and matching, 4D scene reconstruction, sound event detection and (meta-)data modeling and querying.

In Section 2, after presenting existing datasets, we introduce our own and conduct a comparison. We also briefly present our two detailed scenarios (scripts) implying vehicles and humans coming and going around a building. These scenarios were then played in real conditions and recorded. The way these videos were recorded and synchronized is detailed in Section 3. For each video, we provide a large amount of annotations detailed in Section 4. Finally, in Section 5, the conclusion and perspectives, we describe potential applications that are suited to be evaluated on this dataset.

## 2 EXISTING DATASETS

Our dataset stands at the intersection of multiple fields. On the one hand, image related topics cover video surveillance, multi-view, object detection, recognition and tracking, as well as action and event detection. On the other hand, audio related fields cover salient sound detection and recognition as well as multi-source processing. To our knowledge, there exists no dataset suitable for both video and audio methods evaluation. Thus, we will review visual datasets and audio datasets separately.

### 2.1 Visual datasets

We only reference works that the are closest to ours. For an exhaustive survey, the reader can find details in [4]. The HumanEva dataset [16] consists in an indoor dataset composed of 4 people moving over a scene while being filmed by 7 synchronized cameras (4 color cameras + 3 black and white cameras) with largely overlapping fields of view. People are filmed one by one and their body is fully visible. The Utrecht Multi-Person Motion (UMPM) benchmark [18] provides 4 color cameras with 30 persons. Video sequences are more challenging than video sequences of HumanEva dataset as they contain several persons walking at the same time and occluding each other. These two datasets are designed for articulated human motion recognition.

Purely action oriented datasets can be found in Multicamera Human Action Video (MuHAVi) dataset [17] where 14 actors are performing 17 different action classes (such as “kick”, “punch”, “gunshot collapse”) while 8 cameras capture the indoor scene and in [12] with 20 actors, 18 action categories divided into 4 groups (micro or intense action with or without an object), and 5 cameras with both indoor and outdoor scenes, as well as different illumination settings. Likewise, Human3.6M [9] contains videos where 11 actors perform 15 different classes of actions while being filmed by 4 digital cameras. Its specificity lies in the fact that 1 time-of-flight sensor and 10 motion cameras were also used to estimate and to provide the 3D pose of the actors on each frame. Both background subtraction and bounding boxes are provided at each frame. In total, more than 3.6M frames are available. In all these cases, actions are performed in unrealistic conditions as actors follow one by one a scenario consisting in performing actions one after the other.

The Video Image Retrieval and Analysis Tool (VIRAT) dataset [13] provides a large amount of surveillance videos with a pixel resolution of $1920 \times 1080$. In this dataset, 16 scenes were recorded for hours and, at the end, only 25 hours with significant activities were kept. Moreover, only two pairs of videos present overlapping fields of view. Moving objects have been annotated by workers with bounding boxes, as long as some buildings or areas. Three types of events are also annotated, namely single person events, person and vehicle events and person and facility events, leading to 23 classes of events. Most actions were performed by general population with minimal scripted actions, resulting in realistic scenarios with frequent incidental movers and occlusions.

**Figure 1: A subset of all the synchronized videos for a particular frame of the first scenario. First row: cameras located in front of the building. Second and third rows: cameras that face the car park. A car is circled in red to highlight the largely overlapping fields of view. When a frame is missing in a video, it is replaced with a black screen (camera 24).**

The 3D People Surveillance Dataset (3DPeS) [2] comprises 8 cameras with disjoint views and 200 different people. Each person appears, on average, in 2 views. More than 600 video sequences are available. Thus, it is well-suited for people re-identification. Cameras parameters are provided, as well as a coarse 3D reconstruction of the surveilled environment.

## 2.2 Audio datasets

The AudioSet dataset [7] consists of YouTube videos annotated in audio events. This dataset holds numerous video references (2 million of 10s videos) and a wide range of audio classes (632). However, the audio events are not temporally annotated. Moreover, the dataset contains misinterpreted sounds resulting in sounds labeled with a wrong class, for example a "cracking whip" labeled as "gunshot, gunfire".

Other datasets contain only audio files. The Freesound project [6] holds nearly 400 thousand of recordings updated by users with tags. Even if several datasets are hosted on Freesound [10, 15], most of the sounds are not annotated. The Urbansound dataset [15] is composed of 1302 audio files of complex field recordings that are temporally annotated.

Finally, to our knowledge, there is no audio dataset which is specifically dedicated to surveillance applications. Moreover, the existing corpora are either poorly annotated, either unrealistic or too complex. Therefore, in the case of a realistic sound mixture, the acoustic properties of the sound are hard to model [8]. The use of audio data from different recording points offers unprecedented opportunities for a relevant pattern of audio events.

## 2.3 Limitations of video surveillance datasets

Existing multi-view surveillance datasets are often limited for several of the following reasons:

(1) unrealistic actions due to acting in constrained scenes,
(2) few or no overlapping in the fields of view of the cameras,
(3) lack of diversity in terms of events or objects,
(4) lack of ground truth annotations,
(5) absence of audio track,
(6) poor pixel resolution or image quality,
(7) composed of only short videos consisting of a single action,
(8) poor or no synchronization between the different viewpoints.

These limitations make it difficult to combine different approaches on a single dataset. Table 1 compares our dataset to others that we have found being the closest accorded to the intended applications. The characteristics used in this comparison are, in particular, the pixel resolution, the number of overlapping and disjoint views, the audio track availability, the number of objects to be detected and the realism of the scene in the context of surveillance. In the next section, we will develop all the aspects of this new dataset.

# 3 PROPOSED TOCADA DATASET

The Toulouse Campus surveillance Dataset, named ToCaDa, contains two sets of 25 temporally synchronized videos corresponding to two scripted scenarios. Figure 1 shows a subset of the 25 views of the same frame. With the help of about 50 persons (actors and camera holders), these videos were shot on July 17th 2017 at 9:50 a.m. and 11:04 a.m. respectively. Among the cameras:

- 9 were located inside the main building and shot from the windows at different floors. All these cameras are focusing the car park and the path leading to the main entrance of the building with large overlapping fields of view.
- 8 were located in front of the building and filmed it with large overlapping fields of view too (these 9+8=17 overlapping view cameras can be seen on Figure 2).
- 8 cameras were arranged further, scattered around the university campus, see Figure 3. Each of their views is disjoint from all the others.


**Figure 2: The main building which concentrates 17 cameras with overlapping fields of view. 8 other cameras are located out of this field of view, see Figure 3.**

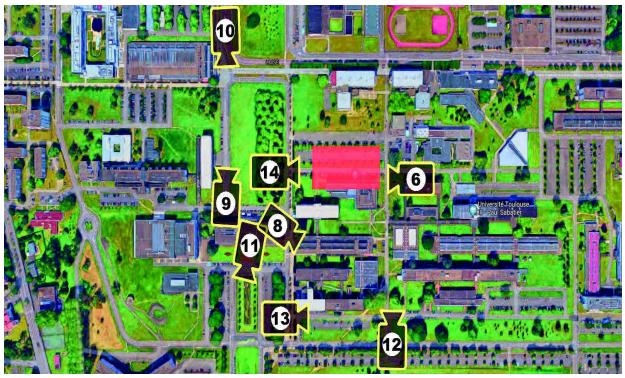


**Figure 3: The positions of the different cameras within a parcel of Toulouse Campus (Paul Sabatier site). These 8 cameras have disjoint views. The red area corresponds to Figure 2.**

About 20 actors were asked to follow two realistic scenarios by performing scripted actions, like driving a car, walking, entering or leaving a building, or holding an item in hand while being filmed. In addition to ordinary actions, some suspicious behaviors are present. More precisely:

- In the first scenario, a suspect car ($C$) with two men inside ($D$ the driver and $P$ the passenger) arrives and parks in front of the main building (at the sight of the cameras with overlapping views). $P$ gets off of the car $C$ and enters the building. Two minutes later, $P$ leaves the building holding a packet and gets in $C$. $C$ leaves the parking and gets away from the university campus (passing in front of some of the disjoint fields of view cameras).
- In the second scenario starts similarly with a suspect car ($C$) and two men inside ($D$ the driver and $P$ the passenger) which arrives and parks in front of the main building (again at the sight of the cameras with overlapping views). $P$ gets off of $C$ and enters the building. One minute later, a woman complains to $D$ about his bad parking. $C$ quickly goes away and stops in the field of view of the camera 8. Approximatively one minute later, $P$ leaves the main building holding a packet, and runs away. $P$ meets $C$ a little further (in the field of view of the camera 8), gets in $C$, and $C$ quickly leaves the university campus (passing in the fields of view of most cameras).

In total, about 30 different moving objects are present, relatively or not to the above scenarios (suspicious behaviors). Concerning sound events, 80 different sound objects can be heard in the videos. About 10 objects present both a visual and an audio component, resulting in a total of 100 audio-visual objects.

Typical video size is around 400 to 600 Mb. The pixel resolution of the original videos varies from $640 \times 480$ to $1920 \times 1080$ but most of the videos have a $1920 \times 1080$ pixel resolution. We provide different resolutions for each video: $1920 \times 1080$, $960 \times 540$ and $640 \times 360$.

To take full advantage of the videos, it seems helpful to temporally synchronize them so that, for a given time, all the frames of the different cameras match.

Camera holders used their own mobile devices to record the scene, leading to a large variety of resolutions, image quality, frame rates and video duration. In order to coordinate this heterogeneous disposal, three foghorns were blown:

(1) The first one stands for a warning 20 seconds before the start, to let enough time to start shooting.
(2) The second one is the actual starting time, useful to temporally synchronize the videos.
(3) The third one indicates the ending time.

All the videos were collected and were manually synchronized using the second and the third foghorn blows as starting and ending times. Indeed, the second one can be heard at the beginning of every video. All the videos of scenario 1 last 4:48 and all the videos of scenario 2 last 5:40. We used kdenlive[1] to cut the original videos and to produce the synchronized ones.

Due to the wide variety of devices used during the shooting of the two scenarios, issues were encountered on some cameras, leading to videos where a few seconds are lacking. To ensure temporal synchronization between videos, black frames were added on the missing intervals of time as it can be seen with camera 24 on Figure 1.

Limits of our dataset are the following.

**Few simultaneous actions:** there are rarely more than two actions occurring at the same time. People are not very challenging to detect and do not appear in groups.

**Few salient sound events:** our videos do not contain screams, horns, explosions or gun shots. Sound events are mainly categorized as motor type sounds.

The video files with audio soundtracks, the detailed scenarios with the list of actions and corresponding times, the files containing ground truth annotations (that will be detailed in the next section) and the list of the irregularities concerning the videos (lacking times or different pixel resolution) can be found on the following link: https://doi.org/10.5281/zenodo.1219421 Additional soundtracks recorded with specific audio devices are also provided.

[1] https://kdenlive.org

## 4 ANNOTATIONS

Ground truth annotations are stored in json files. Each file corresponds to a video and shares the same title but not the same extension, namely <video_name>.mp4 annotations are stored in <video_name>.json. Both visual and audio annotations are stored in each file.

By annotating, our goal is to detect the visual objects and the salient sound events and, when possible, to associate them. Thus, we have grouped them into the generic term audio-visual object. This way, the appearance of a vehicle and its motor sound will constitute a single coherent audio-visual object and is associated to a same ID. An object that can be seen but can not be heard is also an audio-visual object but with only a visual component, and similarly for an object that can only be heard.

We have developed a program for navigating through the frames of the synchronized videos and for identifying audio-visual objects by drawing bounding boxes at particular frames and/or specifying starting and ending times of a salient sound. Each audio-visual object is associated to a unique ID. Regarding bounding boxes, the coordinates of top-left and bottom-right corners of the bounding boxes are given, as in Figure 4. Bounding boxes were drawn such that the object is fully contained inside the box and as tight as possible. For this purpose, our annotation tool allows the user to draw an initial approximate bounding box and then to adjust its boundaries at a pixel-level.

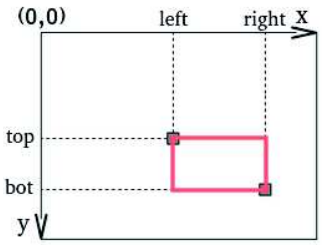


**Figure 4: Coordinates of the bounding boxes top-left and bottom-right corners are given according to this reference.**

As drawing one bounding box for each object on every frame requires a huge amount of time, we have drawn bounding boxes on a subset of frames, so that the intermediate bounding boxes of an object can be linearly interpolated using its previous and next drawn bounding boxes. In average, we have drawn one bounding box per second for humans and two for vehicles due to their speed variation. For objects with irregular speed or trajectory, we have drawn more bounding boxes.

Each bounding box also contains a flag "fully_visible" set to 0 if the object is occluded, even partially, and set to 1 otherwise. On a given frame, an object can be mostly occluded but still present, in particular when it is located near a border of the video. In such case, drawing a bounding box is not relevant. Also, an object can appear and can disappear several times. Instead of considering it as several different objects, we have defined a list visible_times that contains the time segments (defined by a tuple of times "from" and "to") between which an object becomes visible (even partially) and totally disappears from the screen. The visual category of each object is given, consisting in a general term (human, bike, car), along with a list of details like gender and clothes in the case of a human. An example of the json structure of the visual component of an object in a particular video is given in Listing 1.

Regarding the audio component of an audio-visual object, namely the salient sound events, an ID is also given. If the sound comes from an object whose visual component was annotated, the visual and audio IDs have to be the same. An audio category (voice, motor sound) is also given, as long as a list of details and time bounds. In case of an object presenting different sound categories (a car with door slams, music and motor sound for example), one object is created for each category and the same ID is given. In Listing 2, an example of the audio component corresponding to the visual component from Listing 1 is given.

```
{
    "id": 11,
    "category": "motorbike",
    "details": [ "man", "black clothes" ],
    "visible_times" : [
        { "from": 13.8, "to": 18.2 },
        { "from": 29.72, "to": 32.28 }
    ],
    "tbbox": [
        {
            "fully_visible": 0,
            "time": 14,
            "bbox": {
                "right": 536,
                "top": 164,
                "left": 518,
                "bot": 217
            }
        },
        {
            "fully_visible": 1,
            "time": 15,
            "bbox": {
                "right": 550,
                "top": 167,
                "left": 533,
                "bot": 223
            }
        }
    ]
}
```

**Listing 1: json file structure of the visual component of an object in a video, visible from 13.8s to 18.2s and from 29.72s to 32.28s and associated with id 11.**

```
{
    "id": 11,
    "category": "MOTOR",
    "details": ["motorbike", "very noisy"],
    "audible_times" : [
        { "from": 11.24, "to": 21.32 }
    ],
}
```

**Listing 2: json file structure of an audio event in a given video. As it is associated to id 11, it corresponds to the same audio-visual object as the one of Listing 1.**

## 5 CONCLUSIONS AND PERSPECTIVES

We have introduced a new dataset composed of two sets of 25 synchronized videos of the same scene with 17 overlapping views and 8 disjoint views. Videos are provided with their associated soundtracks. We have annotated the videos by manually drawing bounding boxes on moving objects. We have also manually annotated audio events. Our dataset offers simultaneously a large number of both overlapping and disjoint synchronized views and a realistic environment. It also provides audio tracks with sound events, high pixel resolution and ground truth annotations. The originality and the richness of this dataset come from the wide diversity of topics it covers and the presence of scripted and non-scripted actions and events. Therefore, our dataset is well-suited for numerous pattern recognition applications related to, but not restricted to, the domain of surveillance. We describe below, some multidisciplinary applications that could be evaluated using this dataset:

**3D and 4D reconstruction:** the multiple cameras sharing overlapping fields of view along with some provided photographs of the scene allow to perform a 3D reconstruction of the static parts of the scene, see Figure 5, and to retrieve intrinsic parameters and poses of the cameras using a Structure-from-Motion algorithm [1]. Beyond a 3D reconstruction, the temporal synchronization of the videos could enable to render dynamic parts of the scene as well and to obtain a 4D reconstruction.

**Figure 5: 3D reconstruction of the main building from the overlapping views using Structure-from-Motion [1].**

**Object recognition and consistent labeling:** evaluation of algorithms for human and vehicle detection and consistent labeling across multiple views can be performed using the annotated bounding boxes and IDs. To this end, overlapping views provide a 3D environment that could help to infer the label of an object in a video knowing its position and label in an other video.

**Sound event recognition:** the audio events recorded from different points and manually annotated provide opportunities to evaluate the relevance of consistent acoustic models by, for example, launching the identification and indexing of a specific sound event. Looking for a particular sound by similarity is also feasible.

**(Meta-)data modeling and querying:** the multiple layers of information of this dataset, both low-level (audio/video signal) and high-level (semantic data available in the ground truth files) enable to handle information at different resolutions of space and time, allowing performing queries from heterogeneous information.

## REFERENCES


[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. 2009. Building Rome in a day. In *International Conference on Computer Vision*. 72–79.

[2] D. Baltieri, R. Vezzani, and R. Cucchiara. 2011. 3DPeS: 3D People Dataset for Surveillance and Forensics. In *ACM Workshop on Human Gesture and Behavior Understanding*. 59–64.

[3] A. Bedagkar-Gala and S. K. Shah. 2014. A survey of approaches and trends in person re-identification. *Image and Vision Computing* (2014), 270–286.

[4] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. 2013. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding* (2013), 633–659.

[5] M. Crocco, M. Cristani, A. Trucco, and V. Murino. 2016. Audio surveillance: A systematic review. *Comput. Surveys* (2016).

[6] F. Font, G. Roma, and X. Serra. 2013. Freesound technical demo. In *ACM international conference on Multimedia*. 411–412.

[7] J. F. Gemmeke, D. PW Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 776–780.

[8] J-R. Gloaguen, A. Can, M. Lagrange, and J-F. Petiot. 2017. Creation of a corpus of realistic urban sound scenes with controlled acoustic properties. In *Meeting of the Acoustical Society of America and the 8th Forum Acusticum*.

[9] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2014).

[10] Z. Kons and O. Toledo-Ronen. 2013. Audio event classification using deep neural networks.. In *Interspeech*. 1482–1486.

[11] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan. 2015. Crowded scene analysis: A survey. *IEEE transactions on circuits and systems for video technology* (2015), 367–386.

[12] W. Li, Y. Wong, A. Liu, Y. Li, Y. Su, and M. S. Kankanhalli. 2017. Multi-Camera Action Dataset for Cross-Camera Action Recognition Benchmarking. *e-print arXiv (1607.06408)* (2017).

[13] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. 2011. A large-scale benchmark dataset for event recognition in surveillance video. In *Computer Vision and Pattern Recognition*. 3153–3160.

[14] H. S. Parekh, D. G. Thakore, and U. K. Jaliya. 2014. A survey on object detection and tracking methods. *International Journal of Innovative Research in Computer and Communication Engineering* 2 (2014), 2970–2979.

[15] J. Salamon, C. Jacoby, and J. P. Bello. 2014. A Dataset and Taxonomy for Urban Sound Research. In *ACM International Conference on Multimedia*.

[16] L. Sigal, A. O. Balan, and M. J. Black. 2009. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated HumanMotion. *International Journal of Computer Vision* (2009).

[17] S. Singh, S. A. Velastin, and H. Ragheb. 2010. MuHAVi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods. In *International Conference on Advanced Video and Signal Based Surveillance*. 48–55.

[18] N. P. van der Aa, X. Luo, G. J. Giezeman, R. T. Tan, and R. C. Veltkamp. 2011. UMPM benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *IEEE International Conference on Computer Vision Workshops*. 1264–1269.