The Terrascope Dataset: A Scripted Multi-Camera Indoor Video Surveillance Dataset with Ground-truth

Christopher Jaynes, Amit Kale, Nathaniel Sanders, Etienne Grossmann
Ctr. for Visualization and Virtual Environments
and Dept. of Computer Science
University of Kentucky
jaynes@cs.uky.edu

(Proceedings of the IEEE Workshop on Visual Surveillannce and Performance Analysis for Tracking and Surveillar

Abstract

This paper introduces a new video surveillance dataset that was captured by a network of synchronized cameras placed throughout an indoor setting and augmented with ground-truth data. The dataset includes ten minutes of footage of individuals who are moving throughout the sensor network. In addition, three scripted scenarios that contain behaviors exhibited over a wide-area, such as "gathering for a meeting" or "stealing an object" are included to assist researchers who are interested in wide-area surveillance and behavior recognition. In addition to the video data, a face and gait database for all twelve individuals observed by the network of cameras is supplied. Hand-segmented ground-truth foreground regions are provided for every 500th frame in all cameras and for many sequential frames in two overlapping views. The entrance and exit time of each individual in each camera for one of the scenarios is provided in an XML database. We believe that the dataset will help provide a common development and verification framework for the increasing number of research efforts related to video surveillance in multiple, potentially non-overlapping, camera networks.

1. Introduction

Video surveillance involves many central and open problems of computer vision, including multi-camera calibration, object tracking, recognition, and event detection. Solutions to these problems can facilitate applications of video surveillance networks to many areas of our lives. Beyond the obvious security uses, potential applications include environmental monitoring and animal behavior studies, assistive technologies for the elderly, and intelligent environments that efficiently distribute computational resources, manage infrastructure, and support everyday activities.

Due to the rich challenges in the area as well as the potential for impact, the video surveillance domain and its host of problems have been the focus of computer vision research for more than two decades. This attention from the research community has only increased over the past few years and there is no reason to believe that the number of research results produced each year will diminish. As the number of competing algorithms grows, it is important that we are able to empirically study their performance trade-offs in a scientific manner. These studies can then justify attempts to compose these algorithms into complete systems that will lead to new applications.

Given this challenge, a number of researchers have conducted careful comparisons between specific algorithms [1, 2], and others have gone beyond the traditional performance analysis of a single algorithm to discover characteristics [3]. A community focused on these efforts has formed and holds annual conferences devoted to the subject viz. the PETS workshop series. Perhaps the most valuable outcome of these meetings has been the introduction of a variety of testbed datasets (described in Section 1.1) that can then be used to study various video surveillance algorithms. As new research problems emerge, new controlled datasets must also be produced. Here we introduce a new dataset that focuses on an emerging subproblem in the video-surveillance domain; multi-camera wide area surveillance.

As the availability of digital video cameras increases dramatically, researchers are interested in exploiting very large-scale networks of cameras that may be distributed over a wide-area to track individuals, recognize behaviors and match subjects as they move from one view (potentially disjoint) to the next. The "Terrascope" project [4] and others like it [5, 6] are developing algorithms specific to the problem of potentially hundreds of sensors with unknown configuration deployed over a large area. Although many of the problems of single-view or stereo surveillance remain, the new and unique challenges must be reflected in new datasets that are capable of supporting rigorous experimentation.

The *Terrascope* dataset, described here, is intended to provide ground truth information on surveillance in an indoor office environment. The dataset was collected from a network of nine time-synchronized digital cameras. The dataset includes 10 minutes of "natural activity" of participants. In addition, three different "scenarios," ranging from 2 to 3 minutes in length are included that contain individuals carrying out scripted activities. For each of the twelve different individuals observed by the network, a sample image of their face and a sample sequence containing the subject's gait is also provided. Ground truth frames are also provided in which subjects and objects in motion were hand-segmented and labeled with a unique identifier. Entrance and exit events are stored in an XML database that encodes the time of the event as well as the identity of the subject. The data is freely available to the research and educational community.

The database supports analysis of (multicamera) tracking algorithms and background subtraction algorithms for the single camera case. Moreover, by providing identity labels for the actors moving in the multicamera network the database provides a natural testbed for evaluation of cross-camera tracking and recognition algorithms [7]. A biometric database that includes frontal face images and a video gait sample for each individual seen in the network supports valuation of face and gait recognition algorithms.

The sequences also display different activities, some of them occuring repeatedly or being visible in more than one camera, providing an interesting dataset for activity recognition algorithms such as [8]. Finally, the slight overlap of some cameras is meant to support algorithms that make use of multiple views of the same scene.

1.1. Related Work

In recent years there has been considerable interest in the visual surveillance of wide area scenes. The growth in the development of the field has not been matched with complementary systematic performance evaluation of developed techniques. Comparing algorithms is especially difficult if they have been tested on different datasets under widely varying conditions. In recent years there have been several attempts at systematic comparison of different vision algorithms. We mention a few of them here. Examples include the FERET dataset for a systematic evaluation of face recognition algorithms [1] and the USF dataset for gait recognition algorithms [2]. Another example is a skin database consisting of approximately 2100 frames consisting of hand labeled skin regions which can be used for assessment of skin detection algorithms [9].

The PETS workshop series is unique in that all participants are testing algorithms on the same datasets. We briefly describe some of the datasets available for performance evaluation of vision algorithms. The PETS 2001 dataset [1] provides multiview (two camera) outdoor video sequences of people and vehicles in motion while the PETS 2002 dataset consists of people moving in front of a shop window and was designed for people-tracking and counting and hand posture classification. More recently, a number of video clips including people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and leaving a package in a public place were recorded as a part of the CAVIAR project [10]. Our dataset is unique in the sense that it allows for performance evaluation of not only low-level vision algorithms but also provides for evaluation of high level vision tasks including human identification and detection of abnormal activity.

2. Video sequences

The dataset consists of annotated video sequences complemented with a mugshot and gait database for all of the involved individuals. The video sequences were produced by nine synchronized cameras capturing the activities of twelve members of a laboratory. Video sequences are stored as a sequence of still frames encoded in PNG format. One view from each of the nine cameras is shown in Figure 1. More details regarding frame capture can be found in Section 2.1.

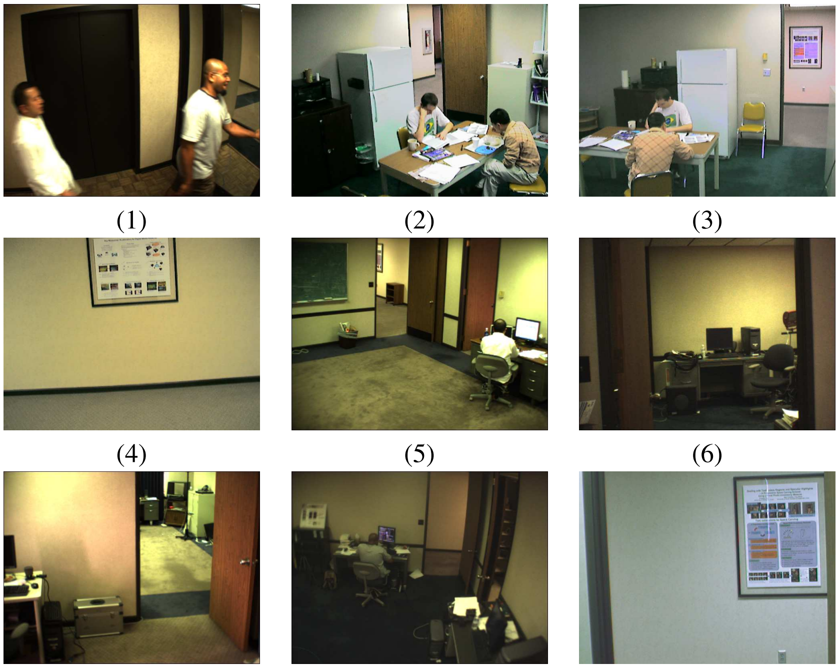
The nine cameras were kept fixed and acquired both the 10 minute natural activity video as well as three 2-3 minute scenarios. The persons visible in the dataset all formally consented to participate. They were instructed to follow the general line of the scripts described in Sections 2.2.1-2.2.3, while otherwise acting normally. The scripts contained the following general behaviors:

1. Group Meeting: Many people gather around a table in a room and conduct a meeting.
2. Group exit and intrusion: A group assembles at an elevator to leave the building, while elsewhere, a subject substitutes a suitcase for another.
3. Suspicious Behavior/Theft: All Subjects leave work at end of day and one subject returns searching for a particular item and accesses a computer.

In addition to these scenarios, a 10 minute, un-annotated video sequence of natural activities, involving the participants seen in the three scripted scenarios is provided. This data contains several views of individuals moving from one scene to the next as they carry out their daily routine within the surveillance space. The natural activity data can be used to train algorithms that require training data captured under similar (but different) conditions.

The dataset also contains mugshots and profile video sequences of each individual walking. This data is provided

[1] ftp://pets2001.cs.rdg.ac.uk/



(1) (2) (3)

(4) (5) (6)

(7) (8) (9)

Figure 1: Example Images

for the purpose of performing face and gait recognition and is detailed in Sequence 3.

Finally, the annotation data, which will be detailed in Section 3, consists of hand-segmented images and event logs. In general, every 500th frame of each scenario was hand-segmented so that the region occupied by each person was filled with his/her color. In addition, frames from two cameras in Scenario 1 that had significant overlap of the meeting area, were hand-segmented sequentially for 50 frames to provide dense ground truth in a cluttered scene.

2.1. Camera Deployment and Framegrabbing

The setting of the scenarios is an office space covering several labs, hallways, and a meeting room. Figure 2 shows an overhead view of the space and of the camera placement. Camera labels in the Figure correspond to those in table 1. Illumination was generally uncontrolled and was normal to office lighting conditions: mostly electric, with lights being turned on and off by participates at different times in the data. Some effects of natural lighting are visible in Scenario 3.

Camera intrinsics were set by hand to capture reasonable images of each area under observation. Once set, intrinsics remained fixed on each camera. Each camera was then attached to its own PC using firewire as interconnect. This allows us to capture uncompressed video at 30 frames per second[2]. Each PC is connected to the local-area network using 100 megabit Ethernet. We synchronize each PC's clock using the Network Time Protocol (NTP) before acquiring data. We then instruct each PC to start capturing data at the same time in the near future (within 5 minutes). Experiments show that the time to start acquiring video on two machines synchronized in this way will differ, on average, by .02 seconds – less than the time to acquire one frame.

During frame capture, the local clock time for each camera was written to a simple text log file. The log shows the clock time at the start of capture for each frame as well as the time when the frame capture was complete. In this way, the synchronized "global" time is available for all frames in the network.

2.2. Sequence description

We now describe each scenario individually.

2.2.1 Scenario 1: Group Meeting

Most of the action in this scenario happens in cameras two and three. People are waiting for a meeting to begin. Some people walk by at various times and finally the leader of the meeting arrives and the meeting begins. During the meeting

[2] Camera number 4 achieved lower frame rates: ∼27 fps in Scenario 1, ∼24.8fps in Scenario 2, ∼29.2 fps in Scenario 3 and ∼28.5 fps in the natural sequence.

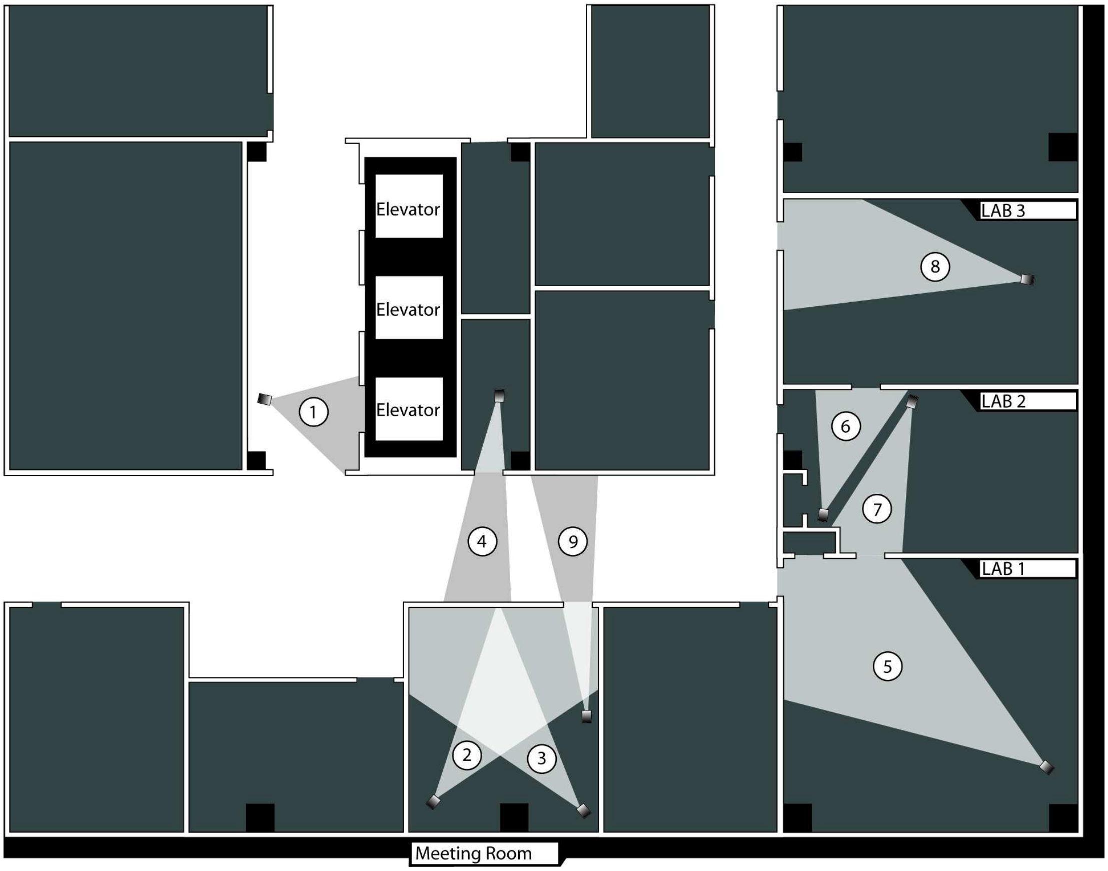


Figure 2: Schematic view of the data collection space. Camera placement was fixed throughout the different data collection scenarios.

No.	Description	Camera Type
1	Elevator entrance area	Point Grey
2	Meeting room, view 1	Sony DFW-VL500
3	Meeting room, view 2	Sony DFW-VL500
4	Meeting room entrance	Sony DFW-VL500
5	Lab 1, view 1	Point Grey
6	Lab 2-3 entrance	Point Grey
7	Lab 1-2 entrance	Point Grey
8	Lab 3	Point Grey
9	Meeting entrance closeup	Unibrain Fire-i400

Table 1: Summary of Camera locations and types.

someone arrives late. After that, a woman receives a call on her cell phone and leaves to take the call. Participants interact with objects on the table in various ways. People at the meeting hand objects to each other and move objects to different places on the meeting table. One behavior that is not the simple transfer or displacement of an object occurs when members of the meeting hold, in unison, a document that they are discussing. Also at this meeting, people are conversing and show a variety of facial expressions, hand expressions, and postures.

2.2.2 Scenario 2: Group Exit and Intruder

People get up, turn off the lights behind them and go to the elevator, as if leaving work. In the meantime, person number 11 arrives by the elevator, a suitcase in his hand. In the absence of the others, he explores the rooms systematically, turning lights on when arriving and off when leaving. Finally, he finds another suitcase, substitutes it with his own and carries it away, back to the elevator.

Six people appear in this scenario. One person displays a clear "searching" behavior, while the others are less prominent. A significant part of the images has dim lighting or high contrast between dark and lit places, creating a challenge for methods that claim robustness to illumination changes across views.

2.2.3 Scenario 3: Suspicious Behavior/Theft

This activity is distributed across all the cameras in the system and involves the case of loitering and theft. A group of people exit their offices and walk to the elevator. One individual from the group turns back, exiting the field of view of the elevator camera from the same place that he enters. The identity of the person can be verified by his gait, which can be observed by Camera 4, and his face in Camera 3 shown in Figure 2. He then checks if anyone is there in the room. He then proceeds to the neighboring room and picks up a hard drive. He then enters the next room and tampers with the computer of a coworker and exits.

2.2.4 Natural video sequences

These ten-minute sequences display a wide range of activities and behavior: reading, eating, walking, standing, sitting, rising, talking, carrying objects, typing and various gestures can be observed in the twelve participating individuals. Moreover, most activities can be observed repeatedly from different directions.

3. Complementary data

In addition to the video data described above, a major research-enabling component of the proposed dataset lies in its complementary information, which we now detail.

3.1. Segmented images

This data provides ground truth image segmentation of individuals and moving objects, as well as identification of each individual and object.

In frames numbered 1, 501, 1001, etc, the image region occupied by a person or moving object was flood-filled with a unique color associated with that person or object. The "active objects", meaning objects that move in the scene, were also flood-filled with their respective colors. Figure 3 shows a few labeled frames from each scenario and Table 2 gives the color code of each individual and object. The exact time-stamp of capture of every frame is available as well.

Note that since the hand-labeling was done by many people, there may exist some variability in the accuracy of the region boundaries. This is unavoidable, since region boundaries are often ambiguous.

Person ID	**R**	**G**	**B**
001	255	0	0
002	0	255	0
003	0	0	255
004	255	255	0
005	0	255	255
006	255	0	255
007	128	128	0
008	0	128	128
009	128	0	128
010	255	128	0
011	0	255	128
012	255	128	128
File1	0	64	0
Cup1	0	32	0
File2	0	64	64
Cup2	0	32	32
Briefcase1	0	64	32
Briefcase2	255	255	255
Writingpad	64	32	32
Can	32	16	16

Table 2: Color encoding of different individuals and objects in the database.

3.2. Event labeling

High-level analysis is concerned with the events that occur in a video sequence. In order to enable performance evaluation of event-detection methods, we have labeled entry and exit events in each scenario.

Entry is defined as the first frame in which the person is "maximally visible" in the sense of maximum surface visibility in the camera. Exit is defined as the first frame in which the person is no longer visible in the camera. An XML file which records the camera number, frame index, person index and event(ENTRY/EXIT) is also provided. The structure of the XML file (first three lines) is as follows

```
<event>
  <camera>2</camera>
  <frame>1801</frame>
  <person>3</person>
  <type>EXIT</type>
</event>

<event>
  <camera>3</camera>
  <frame>1803</frame>
  <person>2</person>
  <type>ENTER</type>
</event>
```

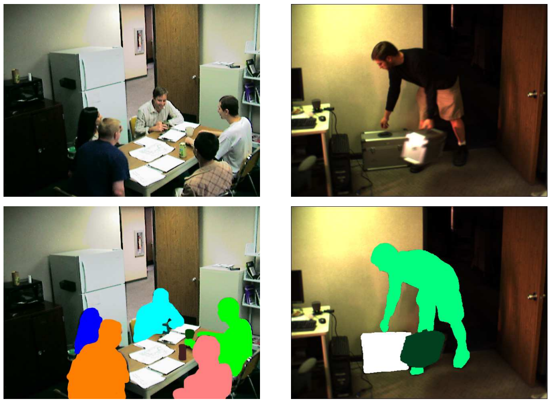


Figure 3: Frames from two different scenarios with color coding of identities

3.3. Individual data

Recognizing people is an important high-level task in video analysis. Oftentimes, additional information about the typical subjects who are likely to be present within a given area is available as input to video surveillance algorithms, e.g. in the form of mugshot images and profile gait sequences. We provide this information, which enables performance evaluation of face and gait-based recognition methods.

Face

For each subject a frontal face image was collected from a stationary camera placed approximately 1 meter in front of the subject. Face images are stored as JPEG encoded frames in the directory `subject-database/faces`. Filenames are labeled *nnn*.jpg, where *nnn* is the corresponding subject ID label. Example mugshots are shown in Figure 4.

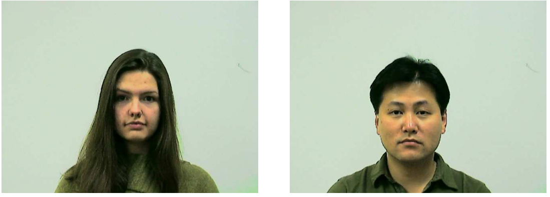


Figure 4: Example mugshots for subjects captured in the surveillance network. Subject files 004.jpg and 006 .jpg are shown here.

Gait

Gait collection took place in the same space as the rest of the data collection under similar illumination conditions. Each subject was asked to walk through the main hallway passing in front of a digital video camera, approximately orthogonally to the camera's optic axis. Subjects passed in front of the camera in two directions to capture gait from both sides of the subject. Approximately fifty frames are captured before the person enters the field of view of the camera to estimate the background model to acquire the binarized silhouettes used in gait recognition. The corresponding JPEG encoded frames from these collection experiments can be found in the directory: `subject-database/gait/`*nn*/, where *nn* is the subject identification label. Figure 5 shows a sample frame for the same two subjects shown in Figure 4. Both face and gait databases were captured using a SONY DFW-VL500.

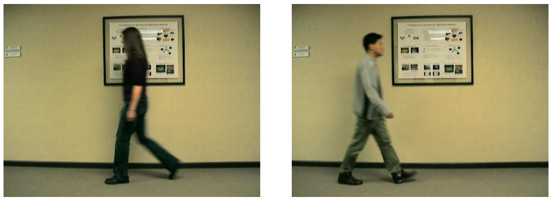


Figure 5: Example frames from the gait sequences corresponding to two different subjects (004 and 006).

4. Obtaining the Dataset

The Terrascope data is freely available for research and educational purposes. Information about obtaining the dataset can be found at *webpage removed for review*. Because the entire dataset is over 100 Gigabytes, direct download of the data is currently infeasible. Instead, those who wish to obtain the entire set of data should mail a hard drive to the maintainers of the Terrascope data. Once received, the data will be copied onto the hard drive and shipped to the user. More information about this process can be found on the Terrascope website.

Partial video sequences from the nine cameras can be downloaded directly from the website as well. In addition to the main website, the Terrascope data can also be found at the VIVID evaluation website *webpage removed for review*. Data on the VIVID site is in JPEG compressed format and broken into a total of 36 different files representing the data captured from the nine different cameras for each of the four scenarios. This data is far smaller that the uncompressed repository found on the main website and can be downloaded in far less time. When using Terrascope data in support of research, we simply request that this paper is cited where appropriate.

5. Extensions and future work

The dataset was intended to support research and development of indoor video surveillance algorithms in multi-camera networks. Although it is already being used in our laboratory and seems to be a valuable resource, we expect that it will continue to evolve over time. The network of cameras from which the data was collected is growing from a nine-camera system to at least 24 cameras over the next year. We expect to re-release a scripted dataset similar to the one described here that utilizes the additional cameras then.

There is an increasing interest in the use of radio-frequency identity tags in the video surveillance domain. We hope to use the same methods described here to provide a controlled multi-camera surveillance dataset that includes some subjects who are wearing RFID.

Although the groundtruth data is fairly sparse, the effort required over 100 human-hours of effort. This number includes only the hand-segmentation and labeling of subjects and objects in the video sequences. Clearly a more efficent approach is needed. We are developing a learning algorithm that utilizes the hand-segmented frames to create new reliable ground-truth frames for the dataset. As soon as this technique is complete, new groundtruth will become available. Finally, we encourage users of the dataset to contribute to groundtruth collection efforts by submitting any groundtruth data to us via the dataset website. In this way, the dataset can increase in utility and continue to serve as a valuable testbed for the video surveillance community.

6 Acknowlegements

Many thanks to Dr. Robert Collins, Xuhui Zhou, and Seng Keat Teh for allowing the use of the VIVID evaluation website as a mirror for the Terrascope data.

References

[1] P. J. Philips, H. Moon, and S. A. Rizvi, "The feret evaluation methodology for face-recognition algorithms," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 22, no. 10, pp. 1090–1100, October 2000.

[2] S. Sarkar, P. Jonathon Phillips, Z. Liu, I. Robledo, P. Grother, and K. W. Bowyer, "The human id gait challenge problem: Data sets, performance, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003(submitted).

[3] Ruo Zhang, P.S. Tsai, J. Cryer, and M. Shah, "Shape from shading: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, Aug 1999.

[4] "*Removed for Review*," .

[5] V. Kettnaker and R. Zabih, "Bayesian multi-camera surveillance," in *IEEE Computer Vision and Pattern Recognition*, 1999.

[6] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, "Tracking across multiple cameras with disjoint views," in *The Ninth IEEE International Conference on Computer Vision*, Nice, France, 2003.

[7] G .Unal and A. Yezzi, "A variational approach to problems in calibration of multiple cameras," *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. I–172– I–178, 2004.

[8] I.Haritaoglu, D.Harwood, and L.Davis., "W4: Who, when, where, what: A real time system for detecting and tracking people," *Proceedings of the Third IEEE Conference on face and gesture recognition*, pp. 222–227, 1998.

[9] L. Sigal and S. Sclaroff, "Estimation and prediction of evolving color distributions for skin segmentation under varying illumination," *Proceedings of IEEE CVPR*, 2000.

[10] R. B. Fisher, "Pets04 surveillance ground truth data set," *Proc. Sixth IEEE Int. Work. on Performance Evaluation of Tracking and Surveillance (PETS04)*, pp. 1–5, May 2004.

A. Data organization

This section describes the file organization of the dataset. The directory tree of the hard disk of the dataset has the following structure:

```
terrascope
|-- annotation
|    |-- scenario1
|    |    |-- camera1    Annotation files
|    |    |      ...     for scenario1
|    |    ‘-- camera9
|    |-- scenario2
|    |    |-- camera1
|    |    |      ...
|    |    ‘-- camera9
|    ‘-- scenario3
|         |-- camera1
|         |      ...
|         ‘-- camera9
|-- scenario1
|    |-- camera1         Images for scenario1
|    |     ...
|    ‘-- camera9
|-- scenario2
|    |-- camera1
|    |     ...
|    ‘-- camera9
|-- scenario3
|    |-- camera1
|    |     ...
|    ‘-- camera9
|-- natural
|    |-- camera1         Natural image
|    |     ...           footage
|    ‘-- camera9
‘-- subjects
     |-- face            Mugshots
     ‘-- gait
          |-- 001        Individual gait
          |    ...       sequences
          ‘-- 012
```