Chapter 23
VideoWeb Dataset for Multi-camera Activities and Non-verbal Communication

Giovanni Denina, Bir Bhanu, Hoang Thanh Nguyen, Chong Ding, Ahmed Kamal, Chinya Ravishankar, Amit Roy-Chowdhury, Allen Ivers, and Brenda Varda

Abstract Human-activity recognition is one of the most challenging problems in computer vision. Researchers from around the world have tried to solve this problem and have come a long way in recognizing simple motions and atomic activities. As the computer vision community heads toward fully recognizing human activities, a challenging and labeled dataset is needed. To respond to that need, we collected a dataset of realistic scenarios in a multi-camera network environment (VideoWeb) involving multiple persons performing dozens of different repetitive and non-repetitive activities. This chapter describes the details of the dataset. We believe that this VideoWeb Activities dataset is unique and it is one of the most

G. Denina (✉) · B. Bhanu · H.T. Nguyen · C. Ding · A. Kamal · C. Ravishankar · A. Roy-Chowdhury · A. Ivers · B. Varda
University of California, Riverside, Riverside, CA 92521, USA
e-mail: gdenina@vislab.ucr.edu

B. Bhanu
e-mail: bhanu@cris.ucr.edu

H.T. Nguyen
e-mail: nthoang@cs.ucr.edu

C. Ding
e-mail: cding@cs.ucr.edu

A. Kamal
e-mail: kamaa002@student.ucr.edu

C. Ravishankar
e-mail: ravi@cs.ucr.edu

A. Roy-Chowdhury
e-mail: amitrc@ee.ucr.edu

A. Ivers
e-mail: aiver001@ucr.edu

B. Varda
e-mail: bvarda@yahoo.com

B. Bhanu et al. (eds.), *Distributed Video Sensor Networks*,
DOI 10.1007/978-0-85729-127-1_23,

challenging datasets available today. The dataset is publicly available online at http://vwdata.ee.ucr.edu/ along with the data annotation.

Keywords Non-verbal communication · Human-activity dataset · VideoWeb · Multi-camera video data

1 Introduction

Research in activity recognition in video is often encumbered by the lack of labeled video datasets which depict realistic activities in practical situations. In addition, existing datasets do not focus on non-verbal communication (NVC) among multiple persons and very few datasets offer footage of the activities from multiple points of view. The *VideoWeb Activities Dataset* described in this chapter aims to fill this need by providing a diverse set of annotated multi-camera footage where the data are collected in a realistic environment and the subjects are real actors who are mimicking everyday NVC activities. The VideoWeb dataset is a collection of 2.5 hours of 51 hand-annotated scenes. Activities are performed by at least four groups of actors where each consists of four persons. The data are recorded simultaneously by four to eight cameras at full 640×480 resolution and approximately 30 frames/second. The dataset is publicly available for download at http://vwdata.ee.ucr.edu/ and is suitable for applications such as automated activity recognition, event search and retrieval, scene analysis, and multi-camera person tracking. This chapter provides the details of various considerations that were made before and during the data collection process. It provides insights in how the data can be used to facilitate the development of activity-centric computer vision algorithms.

2 Data Collection

The *VideoWeb Activities Dataset* was collected over four days using a subset of 37 outdoor wireless cameras from the VideoWeb camera network [1, 2]. In each of the 51 scenes of annotated data we utilize four-eight cameras. For each day, there are 6–18 scenes with each scene having an average length of 4 minutes.

For the first three days, each scene is composed of a collection of human activities and motions which forms a continuous storyline. There are also several object-person interactions in some scenes.

Day 4 of the VideoWeb dataset is different from the first three days. The scenes in this database involve interactions with multiple vehicles as well as multiple persons. Cameras locations for these scenes utilize a top-down view of the environment varying from 15 feet to 70 feet above the ground, mimicking zoomed-in footages from an Unmanned Aerial Vehicle (UAV).

2.1 Purpose and Significance of Data

The VideoWeb dataset is designed for evaluating the performance of human-activity recognition algorithms in a multi-camera network. The dataset is unique in that it is the only publicly available dataset that features multiple-human activities viewed from multiple cameras located asymmetrically with overlapping and non-overlapping views. The dataset also features varying degrees of illumination and lighting conditions.

The data go beyond recording simple actions and atomic activities such as walking, running, and waving. The data were designed in the context of Non-verbal communication (NVC), a process of communication without using words [3]. Non-verbal communication can fall into five categories:

- *Kinesics*—Communication using motion and body language. Examples are:
  - waving goodbye to another person
  - inviting a person to come/enter by waving your hand
  - shaking your head in agreement or disagreement
  - a tour guide pointing at objects of interest
  - drawing/writing symbols in the air
  - raising your hand for a question or answer
  - bowing to show respect
  - standing at attention
  - religious sign—"sign of the cross"
- *Proxemics*—Deriving information from the physical distances between interacting people. Examples are:
  - walking side by side
  - two people standing next to each other
  - playing a cooperative sport
  - marching
  - classroom setting
  - following someone with a distance
  - observing a person from a distance
- *Haptics*—Communication via physical contact. Examples are:
  - hand shake or hug
  - holding hands
  - slapping a person
  - punching or kicking
  - kissing another person
  - a friend giving another friend a massage
  - pinching another person
  - fixing another person's attire
  - dragging someone by hand
  - pushing a person
- *Chronemics*—Structuring time and attaching meaning to it. Examples are:
  - people coming and standing in a queue

  - someone walking to and fro anxiously
  - going back to a particular location regularly
  - security officers doing their rounds
- *Physical Appearance*—This is how a person looks or presents him/her self.
  - student
  - police officer
  - soldier
  - a person wearing a kimono
  - businessman

While this chapter will not delve deeply into the specifics of this topic, a more comprehensive study on NVC can be found in [1].

Our dataset is also geared toward multi-person interaction and interactions between or among groups of people. Many of our scenes feature at least four actors and demonstrate dozens of activities, some of which are repeated by different actors in a different manner. Additionally, actors in our dataset are not wearing highly distinguishable clothing and some are sporting additional accessories such as cane, backpack, and other portable objects.

An additional unique feature of our dataset is that we not only emphasized the person-person interaction but also person-object interaction such as picking up or handling objects. Other types of interaction included in VideoWeb dataset are vehicle–vehicle and person–vehicle interactions.

2.2 Environment for Data

VideoWeb database is collected from the VideoWeb camera network. Currently, there are 37 outdoor cameras composing the network [1, 2]. We have focused our effort on the courtyard area, where we have 17 cameras overlooking the area. Each of the cameras is capable of transmitting image of 640×480 resolution at approximately 30 frames/second.

For the first three days, the data were collected using the layout of the cameras as shown in Fig. 1. The camera locations for the data collection on fourth day are not shown in the layout, since those locations were temporary. Table 1 shows the facts about data collection for each day.

VideoWeb data location covers a space of approximately 7000 square foot and there are several plant boxes and benches. Some of these objects posed problems as they occluded some scenes. Due to the building's design, shadows greatly affect the scenes and the concrete reflects back some of the light (see Fig. 2).

2.3 Contents of Data

The VideoWeb dataset is divided into four days of data collection. For each day there are varying numbers of scenes with dozens of actions performed by multiple actors.

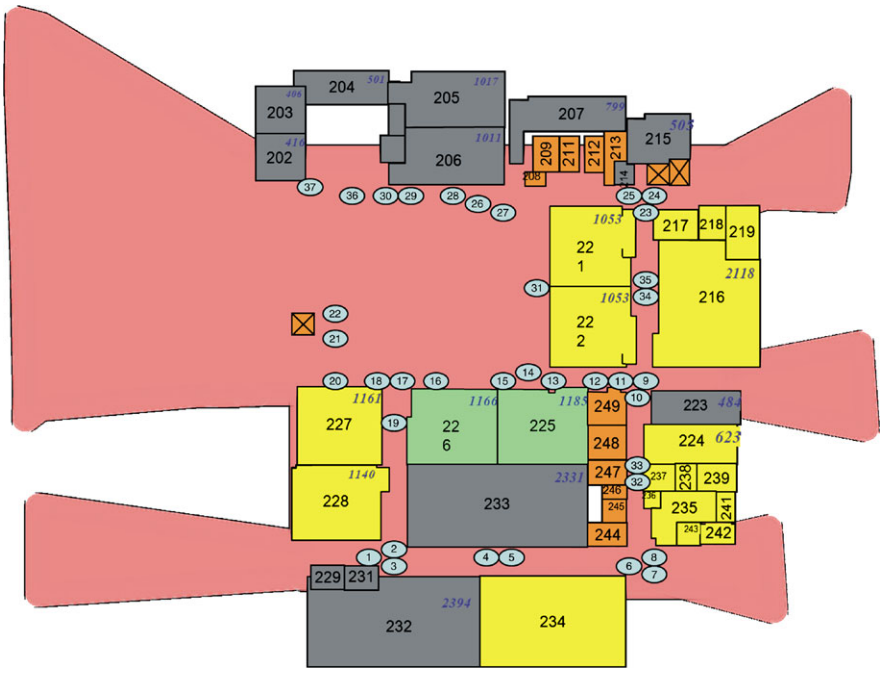

Fig. 1 Layout of the VideoWeb network. *Circles* represent the location of all 37 outdoor cameras in the network. *Pink regions* illustrate the combined field of view of all the cameras. *Blue region* indicates the region where the activities take place for the VideoWeb dataset

Fig. 2 Sample images of the same area at different times of day. (*Left*) Heavy shadows cast from the building's overhead sails. (*Right*) Shadows from the sails are gone; however, concrete is reflecting back some of the light

Each scene in the data is a continuous flow of a story line and there is a corresponding script for each of the scenes. Some scenes will repeat the same script using different actors in different attires, so we have multiple instances of the same activity. Table 2 provides a sample script.

Table 1 Quick facts about each of the data days

	# of Scenes	# of Cameras	Vehicles in video?
Day 1	8	4	No
Day 2	19	8	No
Day 3	18	8	No
Day 4	6	7	Yes

Table 2 Sample script of characters and their corresponding actions

"Selling and Hawking Scene"	
Characters	Actions
Boss: big, expressive guy	• Hand on waist when waiting
	• Shakes two handedly
	• Bumps shoulders with friends
	• Punches arms for fun
	• Points at people
	• Hits smaller guy on head
	• Pushes smaller guy when having a private conversation
Employee: meeker, smaller guy	• Stands with hands behind back
	• Covers ears when it is loud
	• Raises hand in a gesture to control things
	• Wipes his brows
	• Bows when he meets someone new
	• Hand in "prayer" when things get bad

There are dozens of activities and actions featured in this dataset. We have identified 51 significant activities related to NVC. The complete list is given in Table 3.

Table 3 List of 51 common activities in VideoWeb dataset. Notice one action can have multiple meaning in the context of NVC, e.g. raising hands could mean frustration, to interrupt, or a question

Common Activities in VideoWeb Dataset		
Hand off Object	Lie Down on Ground	Waiting Impatiently
Guide Person Away	Stand Up	Raised Hand (question)
Toss/Throw Object	Talk on Phone	Sit on Bench
Explain/Story-telling/Teaching	Argue Within Two Feet (aggressive)	Group Corners Single Person
Walk Backward	Walk Close within 2 Feet	Listening to music/dance
Point (indication)	Raised hands (passive)	Hug
Direct Friend with Wave	Text on Phone	Lean close, obfuscation
Crossed Arms	Wave Off (ignore)	Look About/Scanning Area
Touch to Get Attention	Reading Book	Courteous Nod
Running	Spin while Talking	Slap Self
Raised Hand (interrupt)	Slow Tired Walk	Wave 1-Hand
Shake Hands	Raised Arms (frustration)	Signing Paper
Shoulder Bump Into Someone	Raised Arms (gathering)	Show Object to Someone
Search for Object	Flirting	Pull Someone Away
Find Object	Walking Close	Observe from Afar
Pick Up Object	Sneak Away from Group	Wave Off (ignore)
Sit Cross Legged	Push Button	Shove

Examples The following images (Figs. 3, 4, 5, 6) are sample scenes from the four different days of data collection.

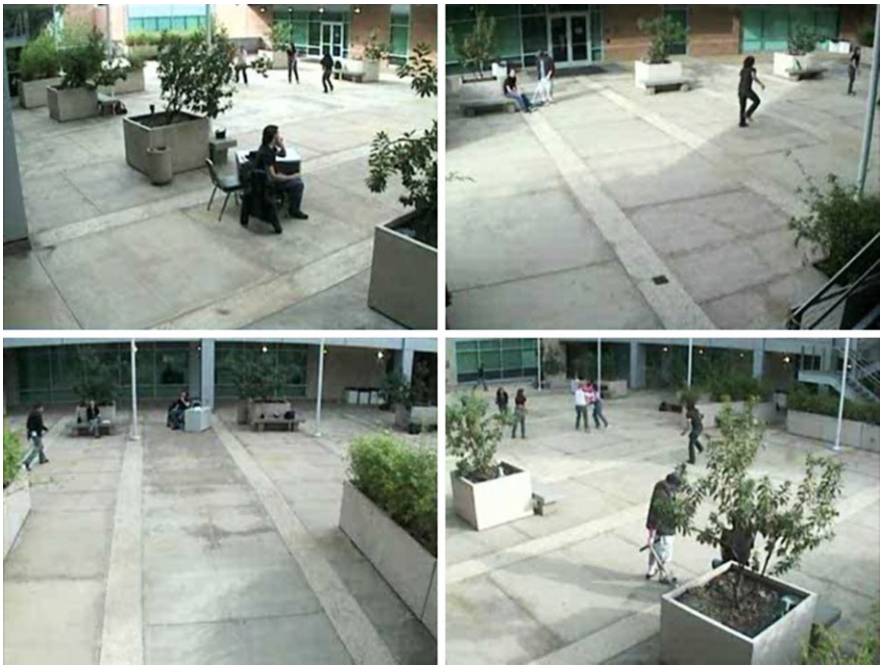

Fig. 3 Day 1. Courtyard scene with four camera views. Some of the actions visible include Dancing, Standing up, Running, Sitting on the bench, Pointing (to indicate), Tossing/Throwing an object, Catching/Picking up object, Talking on phone, Observing from Afar (more than 10 feet), and Sitting cross legged

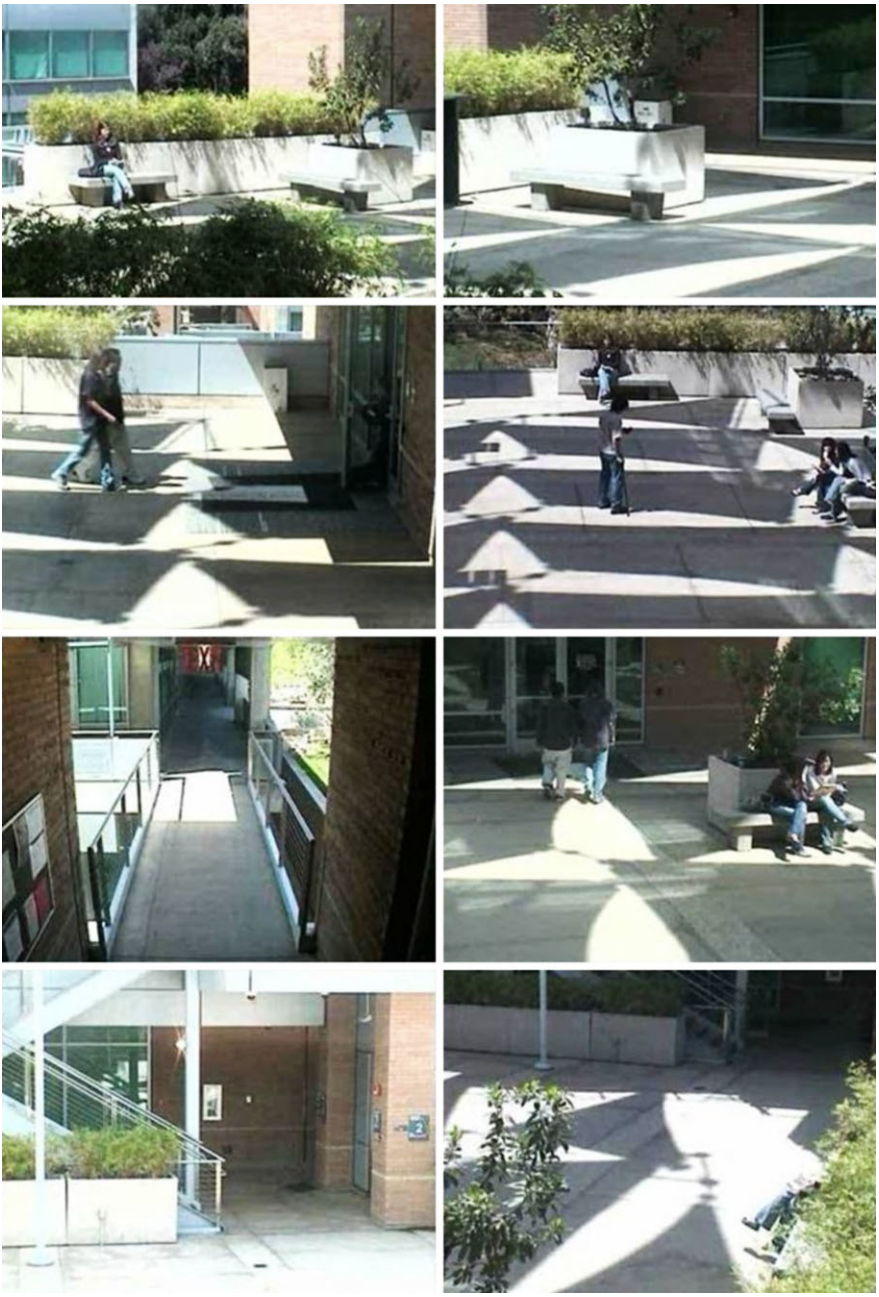

Fig. 4 Day 2. Courtyard scene with eight camera views. Some of the actions visible include Running, Sitting on bench, Waving off (to ignore), Observing from afar (more than 10 feet), Sitting cross legged, Walking with crutches, and Walking side-by-side (within one foot)

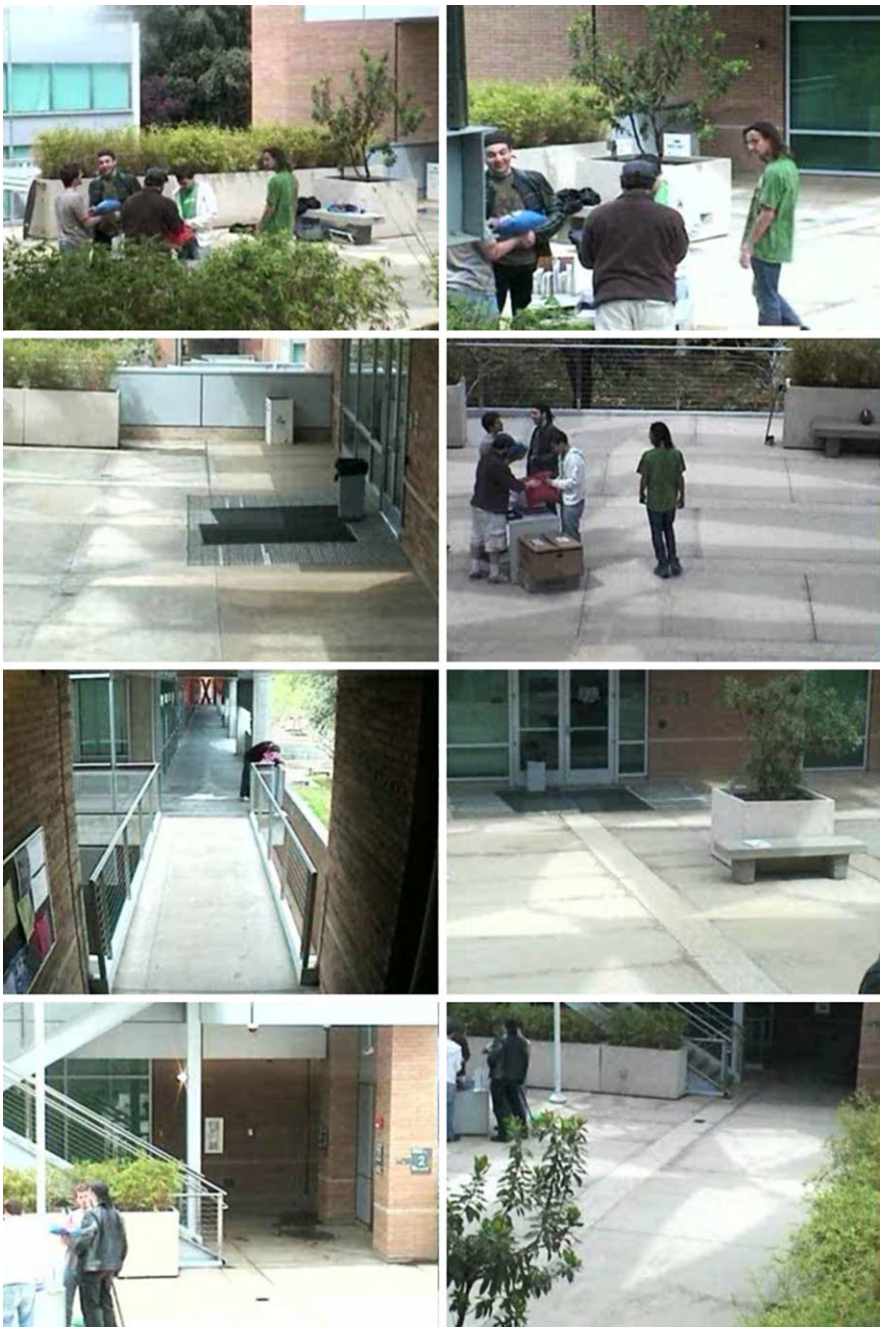

Fig. 5 Day 3. Courtyard scene with eight camera views. Some of the actions visible include Walking, Looking around/Scanning area, Showing an object to someone, Arguing (aggressive), Leaning over rails, Walking slowly, and Holding an object

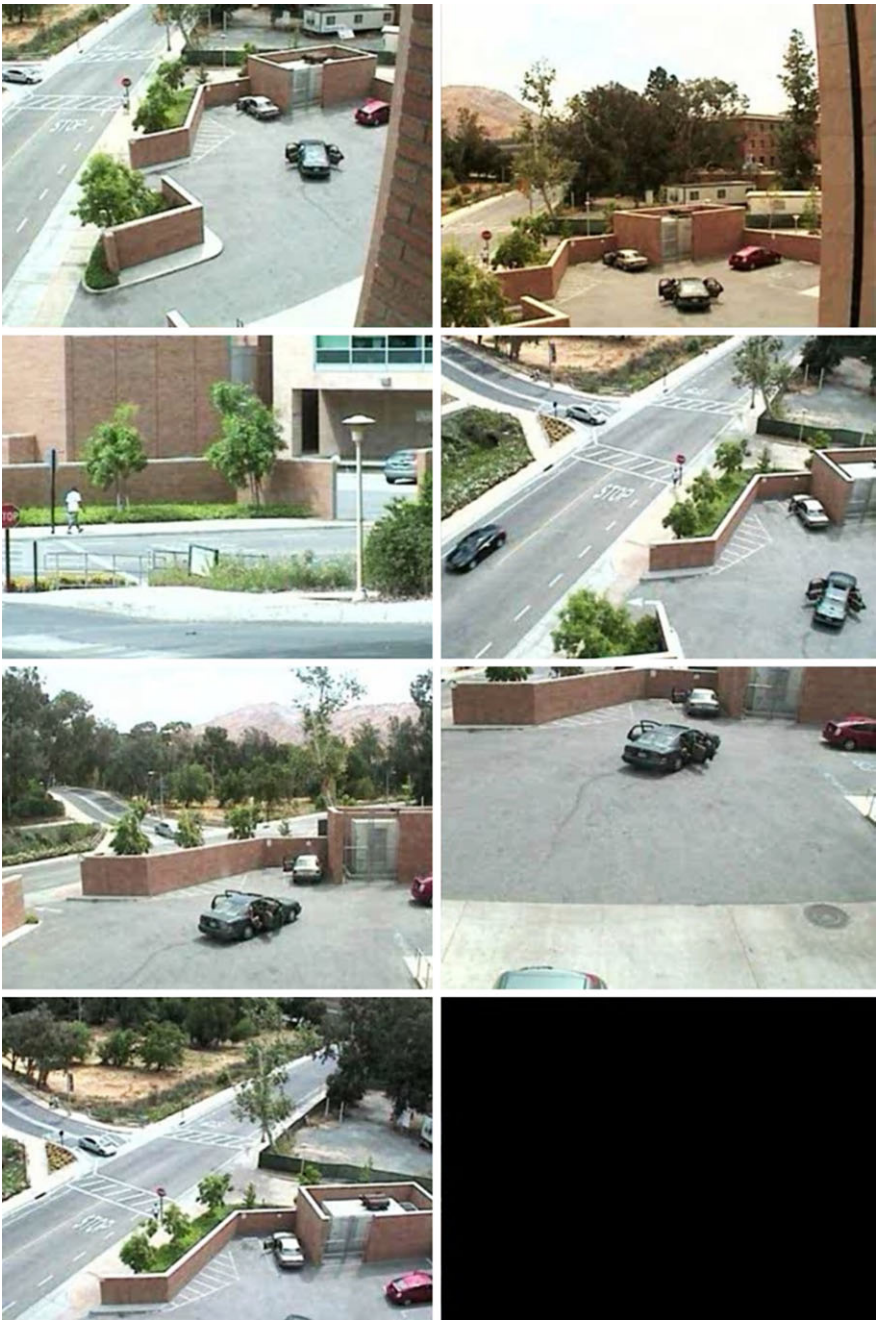

Fig. 6 Day 4. Intersection scene with seven camera views. Some of the human activities are: Getting out of a vehicle, Walking, Closing door, Standing by the vehicle, and Carrying an object. Vehicle activities include Right turn and Stop at intersection

2.4 Ground-Truth Annotations

The VideoWeb dataset includes ground-truth hand-annotation for all of the 51 scenes. To generate the ground-truth each scene was scanned frame by frame; once an activity is identified, the beginning and ending of the action is marked and recorded.

Ground-truth data are stored in XLS format and are organized as follows:

	Instance A	Instance B	...
Action 1	(camera #, start frame, end frame)	...	...

To clarify, we will use an actual data entry from Day 1 Scene 2.

Running	(14, 2136, 2195), (27, 3122, 3208), (37, 2127, 2167)	(14, 3094, 3134)	(14, 3197, 3183)

We look at one of the actions in our list-running. Cameras 14, 27, and 37 have seen the action. For camera 14, the start time is at frame 2136 and end time is at frame 2195. The similar things go for cameras 27 and 37. Another instance of running is seen by camera 14 from frame 3094 until frame 3134, and so on.

Note that ground-truth data do not specify the identity of the person who performed the action. As long as the action takes place, it is identified and marked.

2.5 Availability of the Data

The VideoWeb Dataset is available at http://vwdata.ee.ucr.edu/. After submitting the release form, users will be issued an account and password to access and download the data. Footage is available as both MPEG1-encoded videos as well as raw Motion JPEG frames and ground-truth annotations are provided as XLS files. The dataset consists of 51 scenes recorded across 368 clips with a total size of 85 GB and 123 GB for the MPEG videos and Motion JPEG (MJPEG) data, respectively.

There are four days of data for the VideoWeb Activities Dataset and the days are labeled *Day1*, *Day2*, *Day3*, and *Day4*. Each day contains a folder for each scene for that particular day. Under each *scene* directory, there are videos from each camera, these videos are for visualization purposes. The videos are in .mpeg format. Also under each scene there is an excel file which contains the video annotation. The video annotation is in .xls format. Within the *scene* folder there is also a folder which contains the zip files for all the video images.

The naming convention for the videos is as follows:

- "Day#_Scene#_Camera#.mpeg"

The naming convention for the excel files is as follows:

- "Day#_Scene#.xls"

The video annotation file contains a given list of actions that are observed by a viewer. The annotation has each action separated by the frame number. Within the annotation file, when an action occurs in the video, it is identified by the start frame and end frame.

The naming convention for identifying the action is:

- "(camera#, start frame, end frame)"

  Each column indicates an occurrence of when the action occurred.
  The Utilities folder contains a number of utilities that you might find useful.

- A MATLAB renaming script that renames jpegs so windows correctly lists all the jpegs in order.
- A program that converts MJPEGs in a folder into a mpeg video with the same resolution as the MJPEGs.

It should be noted that users should re-generate videos using the raw Motion JPEGS and the utilities that are provided. We are in the process of replacing the provided videos to deal with a frame skip issue due to slight network lag. Timestamps across videos may not correspond. Solving for the time offsets between cameras is possible by comparing two frames of the same activity from different cameras.

3 Conclusions

VideoWeb dataset is one of the most challenging dataset for human action recognition as of the writing of this chapter. It is also a unique dataset as it is the only dataset on multiple-human interactions involving multiple actions in a multi-camera network. The data are collected in a realistic environment in the context of Nonverbal communication. This is an important feature of our dataset since one human action can have more than one meaning in the context of NVC. For example, raising hand could mean to interrupt, show frustrations, and asking a question. All data are publicly available online at http://vwdata.ee.ucr.edu/ as MPEG videos or raw Motion JPEGs including the hand-annotated data. We hope that the computer vision and pattern recognition community will be excited to use these new data. This will lead to the advancement of the field as it will allow the community to compare different technical approaches on the same data.

Acknowledgements This work was supported in part by ONR grant N00014-07-C-0311, N00014-07-1-0931 and NSF grants IIS 0551741 and ENGR 0622176.

References

1. Nguyen, H., Bhanu, B., Patel, A., Diaz, R.: VideoWeb: Design of a wireless camera network for real-time monitoring of activities. In: Third ACM/IEEE International Conference on Distributed Smart Cameras, Como, Italy, 30 August–2 September 2009
2. Nguyen, H., Bhanu, B.: Videoweb-optimizing a wireless camera network for surveillance. In: Bhanu, B., Ravishankar, C., Roy Chowdhury, A., Terzopoulos, D., Aghajan, H. (eds.) Distributed Video Sensor Networks. Springer, Berlin (2010), Chapter 22
3. Andersen, P.: Nonverbal Communication. Waveland Press, Long Grove (2008)