Construction of a Database of Annotated Natural Human Motion and its Application to Benchmarking

Adam Chmielewski

Boston College

Computer Science Department

Senior Honors Thesis 2004

Advisor: David Martin

0. Abstract

Humans possess an uncanny ability to perceive the motion of other human beings. The performance of current algorithms on this task falls far short of human performance.

data that is anywhere near as effective or accurate as a human operator. The dataset we created in this project represents this useful ground truth to measure an algorithm's accuracy against. This gives us a means to compare the effectiveness of different algorithms. It also helps in the development of algorithms, as different versions of the same algorithm can be compared against each other to gauge effectiveness from revision to revision. For this project, we started collecting our dataset, built the necessary tools to annotate the data, and then put the data to use in comparing the effectiveness of two detection algorithms.

2. Video Data Collection

The video clip dataset is a major part of our efforts. The broad goal is to eventually create a database that contains examples of a multitude of human motions and actions, in as many different conditions as possible. Our specific goal at this time was to collect a decent number of interesting clips, with enough variety among them to do a fair benchmarking of an algorithm, and to have example clips for a lot of these variations. We looked for the following variations in both human activity and recording conditions and techniques:

Human-based variations:

-

- Height: The height of the subjects is an important variation, as tall people will perform the same actions in a different manner than shorter people.
- Shape: Similarly to height, we wanted a good variation in the size of our subjects, from thin to overweight.
- Skin tone: Especially important for evaluating face detection algorithms, we looked for a range of skin tones from very pale to very dark.
- Race: We worked to get a good variation in the race of our subjects, which was helped by the fact we did our filming in a large urban setting (Boston).
- Clothing: A person's garments make a significant difference in how successful an automatic tracking can be, and consequently we want different styles and amounts of clothing to be represented.
- Orientation: The direction a person is facing is important. For face detection, finding a profile face is similarly a different operation (depending on the

leisurely stroll. We set out to capture a number of the more common actions, and when the opportunity arose, to capture any rare or interesting actions that we happened across. Examples of the actions that we captured are listed in the descriptions of the clips later in this paper.

Condition based variations:

Filming-based variations:

- Camera Motion: While some variation in camera motion is desired, we quickly determined that rapid or jerky camera motion produced uninteresting video frames with grotesque motion blur. With that in mind, several kinds of slow camera motion were used in a number of the clips. They include:
    - Panning: horizontal or vertical movement of the camera in the scene, without change in the viewing angle.
    - Tracking: movement of the camera forward or backward in the scene, for example, following someone from behind.
    - Rotation around: A small number of the clips involved moving the camera around the subject.
    - Stationary rotation: rotating the camera on the y-axis. This is similar to panning but different in that the camera's location is not changed.
-

different ways. The camera was often held at waist level, facing a different direction than the filmer was. This naturally resulted in a sometimes less than accurate aim of the camera, but for the most part none of the subjects noticed they were being filmed. Another technique was to leave the camera on a surface, with the filmer leaving it running and not touching it. People don't seem to consider the camera to be filming when there is no one operating it, and therefore it went largely unnoticed.

Video was collected using a fairly high-end handheld digital camcorder, a Sony DCR-TRV950, recording at 720x480 resolution onto mini-DV tape. The video from the tape was imported to our workstation using Apple's iMovie. The interesting and useful parts of the video were then identified, and separated into clips ranging from five to thirty

their output. This obviously was undesired, as we did not want anything that would hurt the purity of our dataset. Our deinterlacer avoids this problem by simply removing half of the scan lines by removing every other scan line. This leaves a vertically half-size, horizontally full-size frame, but one where all of the data is from the same moment in time. We then resize the frame via bicubic sampling back to its normal size. Since we are sampling from data already in the frame, this does not introduce any new artifacts to the data set, and does a reasonably good job of removing the effects of interlacing without degrading the image quality too much. The dataset consists of both the original interlaced frames (so that reconstruction of full video clip can be accomplished) and the converted, deinterlaced frames.

The Dataset:

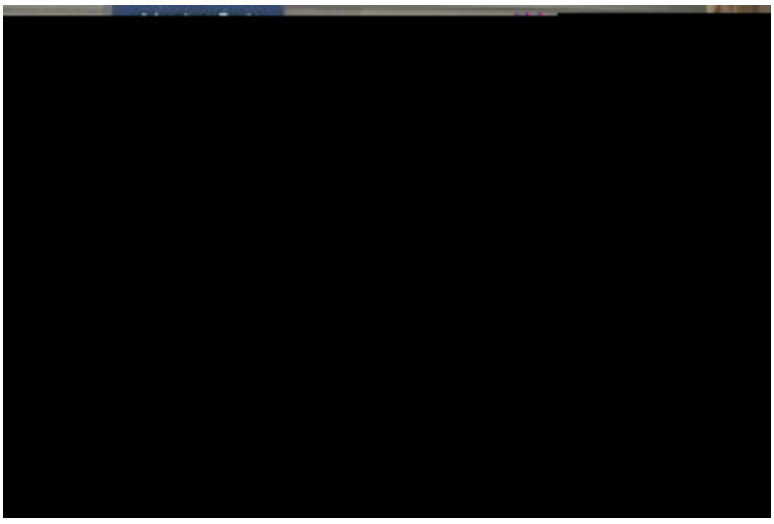
At this time, we have collected 92 different clips of varying length that are representative of the types of variations we are collecting. Because several of the clips are of similar scenes, we have grouped them together as one family of clips for the purposes of characterizing and describing them.

Description of each clip, with an example frame from each following the description:

- Attitash1-2: these two clips follow a single woman as she walks away from the camera, outdoors, on an overcast day. There is camera motion in the second clip, consisting of panning and zooming to keep track of the subject as she gets farther away. Some occlusion.

- Attitash4: two figures, one removing objects from a car trunk, the other walks by carrying a pair of skis by his side. Some occlusion occurs as cars drive by in front of them, and one figure passes in front of the other. One wears dark clothing, the other bright clothing.

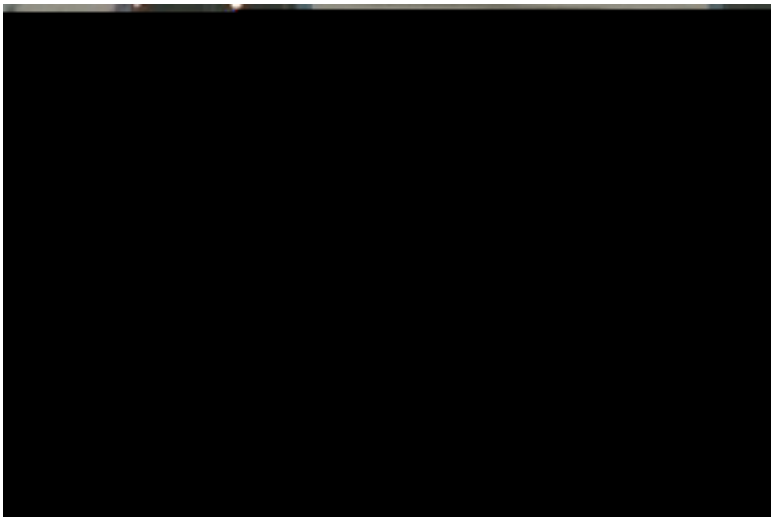
- Attitash5: five figures, all in ski clothes. Age variation, 3 children and 2 adults. Children sit on sidewalk, one throws stones. Follows one child as they walk and then sit down. Some camera movement and zoom-out halfway through the scene.



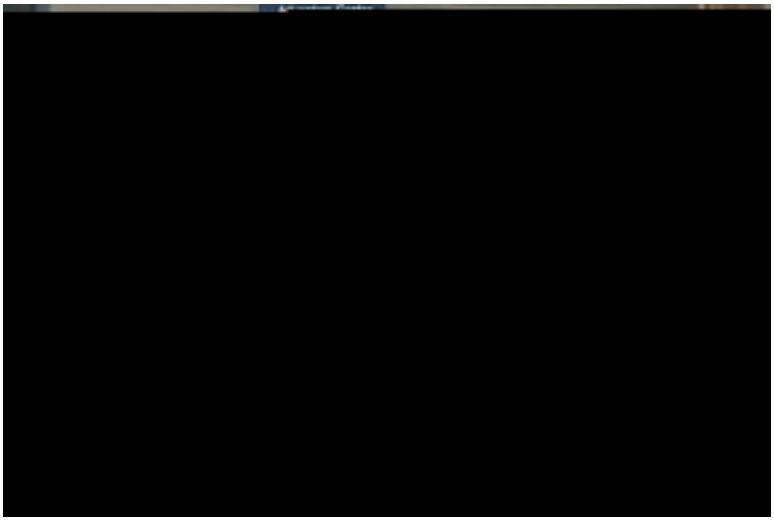
- Attitash6, Attitash 8: two similar clips, 2 figures, one reclines against a waist level fence, the other cleans and places his skis in his car. Some occlusion occurs as he

walks behind the front of the car, and he can only be seen from the waist up. Many different poses for the male figure. Slight camera motion.

- Attitash 7: Similar to 6 and 8, also includes 2 children who run up and jump up onto the fence and walk along it. Slight camera motion.

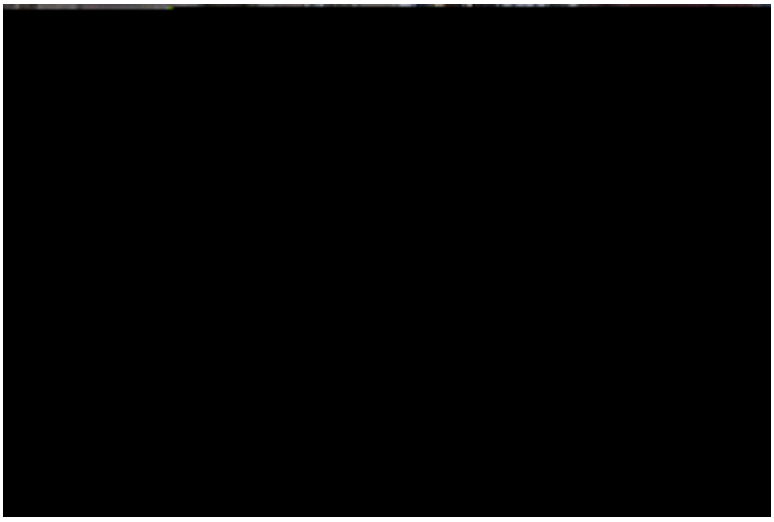


- Attitash 9: Main figure is a child who is digging in the ground. Also contains a woman who walks across the frame, and a man who is bending over to put things into his car trunk and straightening up. Slight camera motion.

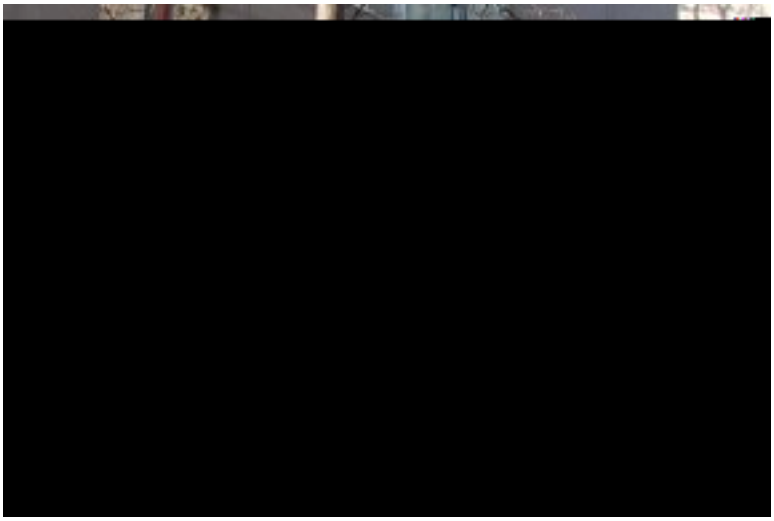


- Bicycle: Short clip of a man riding a bicycle7 Dg7mho is digging in the ground. Also contains a

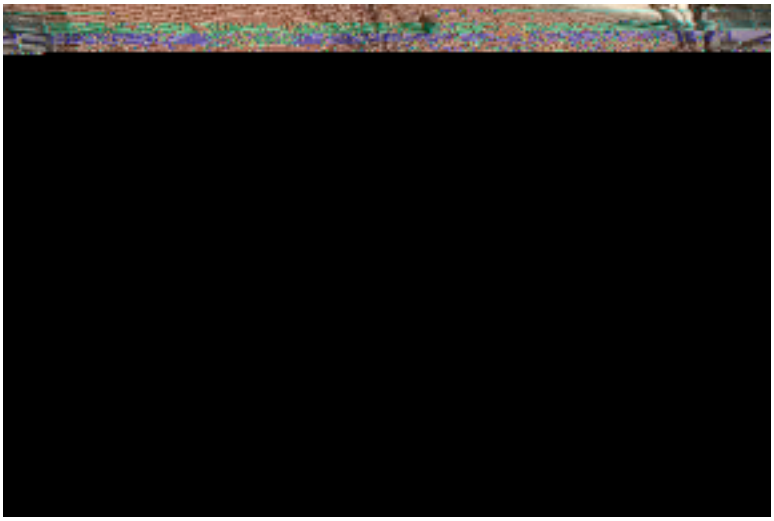
They all contain a number of people walking along the street in one of two directions. Bright sunny outdoor conditions. Winter clothes. Some waist level occlusion. Some different actions while walking; talking to each other, using a cell phone, reading, etc. Great variation in age, height, shape, race, and skin tone.



- Bu6 -7: Similar weather conditions to Bu1-5. Lighting is not as bright as there are shadows from trees in the frame. These clips contain fewer figures, but at different depths in the field, and different vectors of motion. A few interesting actions occur, such as one figure stopping to light a cigarette. Slight camera motion, no zooming.



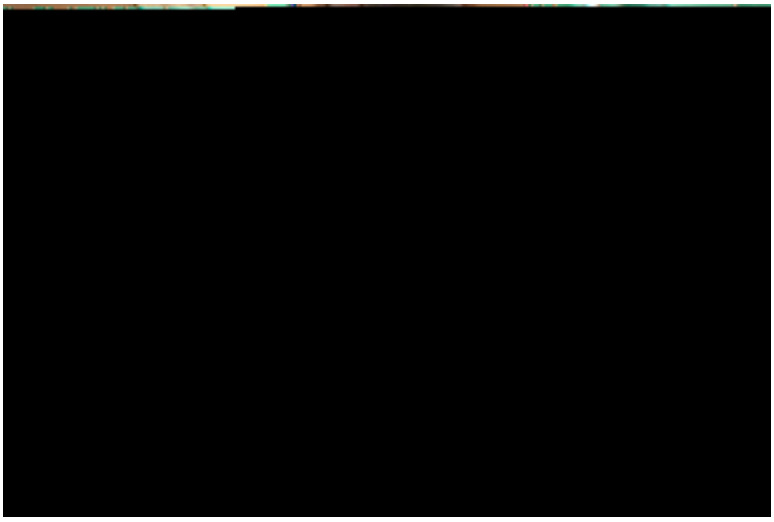
lighting, sunny, outdoors, winter conditions. Lots of occlusion, and some camera movement, including panning, rotation, and zoom.



- Cards1-4: these four clips consist of figures walking through and interacting with an aisle of a supermarket. Some variation in the height of the figures. Actions include removing items from shelves and carrying objects. Bright, indoor lighting. Several well-lit, large faces are present in some of these clips. There is no camera motion, and the only occlusion occurs when figures cross in front of each other.

- Copley3-8: These six clips are all shot at about knee level. They are scenes from a busy crossway in a mall, with a large number of people in every clip. The lighting is brighter than copley1-2, but not very bright overall. Several people walk close to the camera, so that only torsos and upper legs are visible. Many different vectors of motion. Most figures are not wearing heavy jackets, but all are in at least long sleeve shirts and pants. Variation in most human

- Escalator1-2: these clips were an opportunity to have a vertically panning camera motion. Lighting is indoors but dimly lit, mostly from lamps at eye level. Contains variations in race and skin tone as well. Three figures make an interesting path around another group of figures. Same clothing as copley3-8.



- Foodcourt1-4: these four clips are good examples of both a looking-down viewing angle and interesting interactions, such as with a salad bar and a cash register. The scene is a large self-serve foodcourt. A large number of figures in many different kinds of clothing, at varying depths in the frame. Variation also exists in the level of zoom in each clip. Slight camera motion.

- Legs1-4: For these clips, the camera was placed at about ankle level on the side of a walkway. The scene is indoors but dimly lit. The figures are limited to about waist height at the most. This would particularly useful for studying the motion of the legs while walking. Different kinds of shoes are also present: boots, sneakers, heels, etc. No camera motion.

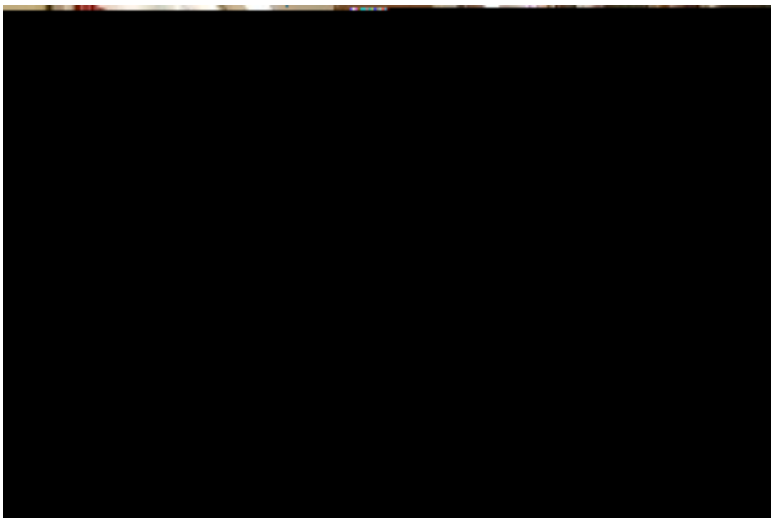
- Library1-4: these 4 clips are of a large entranceway to a library. Lighting is indoors and bright. Large variation in the types of people in the scene. Different motion vectors. Different poses. The third clip contains a woman with a seeing-eye dog. Good variation in the figures' depth in the frame. No camera motion.

- Lobby1-2: Indoor, very dimly and backlit hotel lobby. Age variation, race, and especially shape variation. Child running in second clip. Clothing level ranging from jackets to t-shirts. No camera motion.

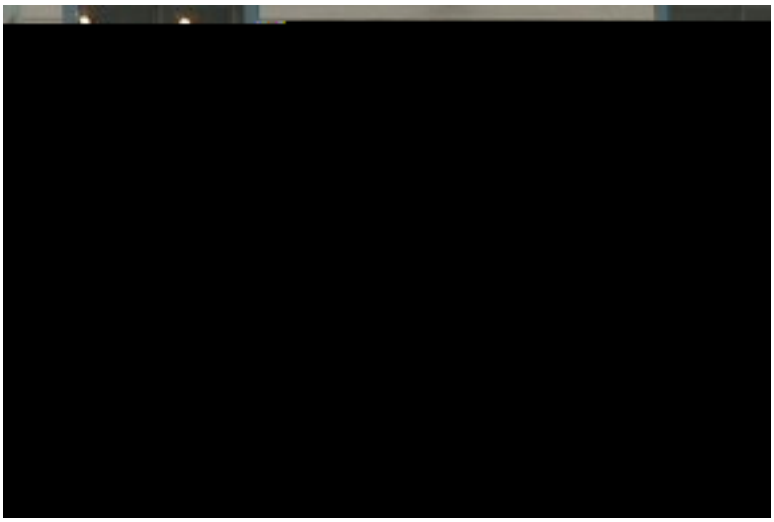
- Mallstairs1: This clip contains one of the largest numbers of figures. It is indoors, average lighting. Figures are both walking up and down a stairway, as well as walking flat along side it. With this number of people there is large variation in most human characteristics. A lot of camera motion, both panning and zooming. Viewing angle is also from above, as in foodcourt and copley9 sequences.

- Mbta1-9: this family of clips involves interactions and motion on a subway car. They contain 1 to 5 people. Because the train is moving, the shadows and lighting are always changing, creating some interesting effects as people go in and out of shadow and light. Motion of train also causes different kinds of walking than would flat ground. Age variation. Little camera motion.

- Milkaisle1-3: Scene is of a dairy aisle in a supermarket. Most figures are pushing shopping carts as they walk. Lighting is bright and indoors. Interaction with objects on the shelf and with the shopping carts. Figures walking almost directly towards and away from the camera. No camera motion, placed at waist level.



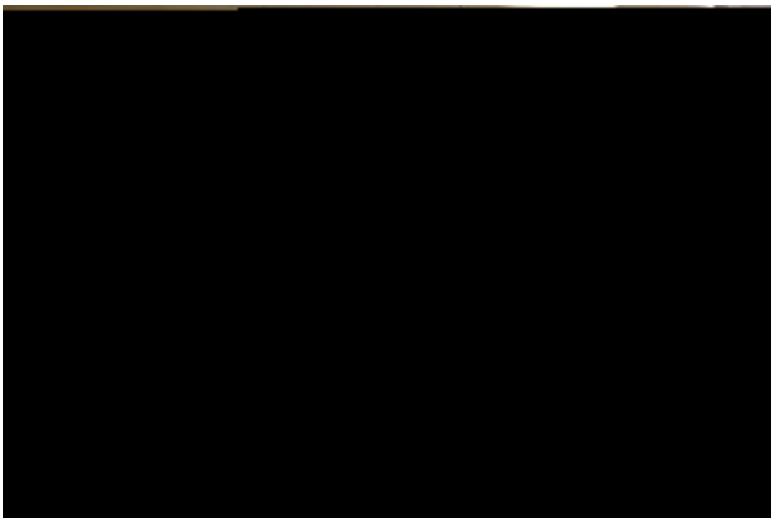
- Parkinglot1-3: These clips are the same conditions and people from the attitash4-5 clips. Parkinglot2 is an especially interesting clip, consisting of 8 figures of varying ages and sizes doing very different actions, including carrying objects, running, jumping off a fence and landing on the ground. Many different colors of clothes. Variety of poses. Some occlusion from objects. Slight camera motion.



- Runner: short (3 seconds) clip of a young woman running on the street. Outdoors,

- Shaws1: Clip consisting of four figures in a supermarket, a small corner of a bakery section. Indoors, average lighting. A couple object interactions. One figure has very bulky clothing. Some occlusion, only the head of one figure can be seen. Slight camera motion.

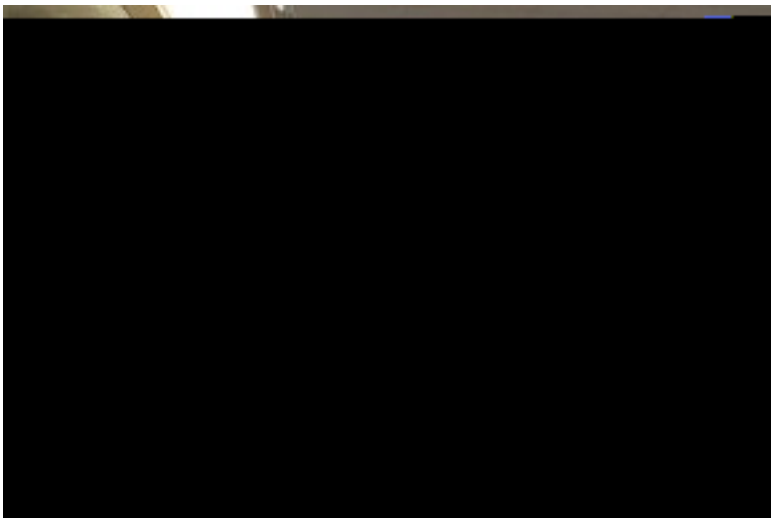
- Shaws2: Short clip of a man leaning on and pushing a shopping cart. Indoors, bright lighting. Some camera motion tracking the figure, and there is occlusion when the figure is behind the shopping cart.



- Shaws3-5: Camera was placed at eye level facing the refrigerated meat section of the supermarket. Scene is indoors, brightly light. A number of different types of

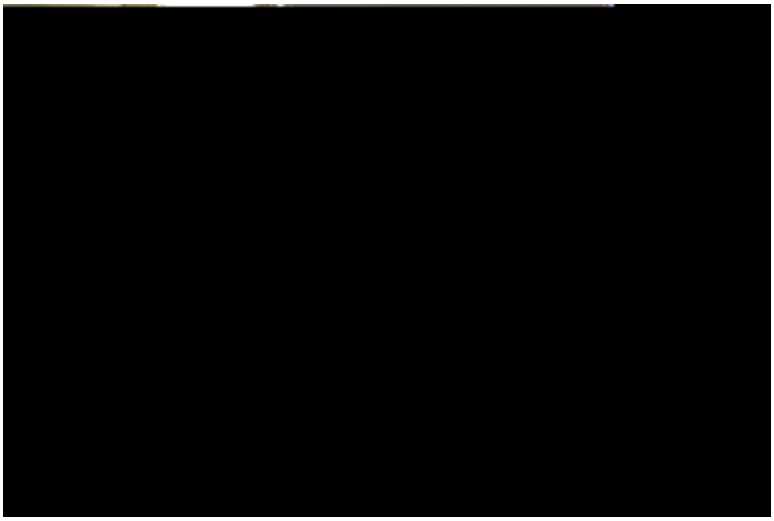
people pass through the frame. Shaws4 and 5 contain a man interacting with a box (removing objects from and placing objects in) in some detail. Some occlusion. No camera motion.

- Shaws6, 8: Shaws6 has two figures, one of whom walks toward the camera and interacts with the objects nearby. Similar conditions to shaws3-5, different aisle. No camera motion or occlusion. Shaws8 is identical to shaws6, with more figures, and figures are pushing shopping carts.



- Shaws7: Same location as shaws6, starts with a figure very close to the camera. Figure is carrying a pallet-like object on one shoulder. Grabs an object from off-

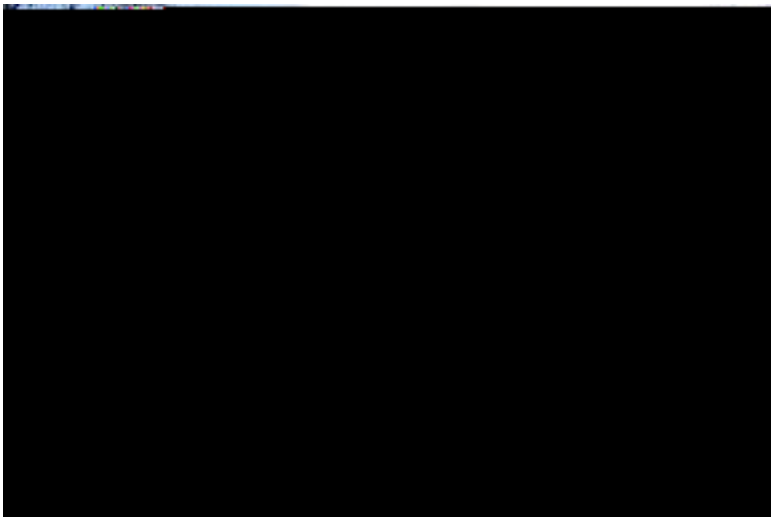
camera and slowly walks away. This is the closest and largest face in the collection so far. Same conditions as shaws3-6. No camera motion or occlusion.



- Shaws9: Camera is held in the midst of a busy section of the supermarket as

- Shawscorner1-2: Camera placed at intersection in supermarket. Many different

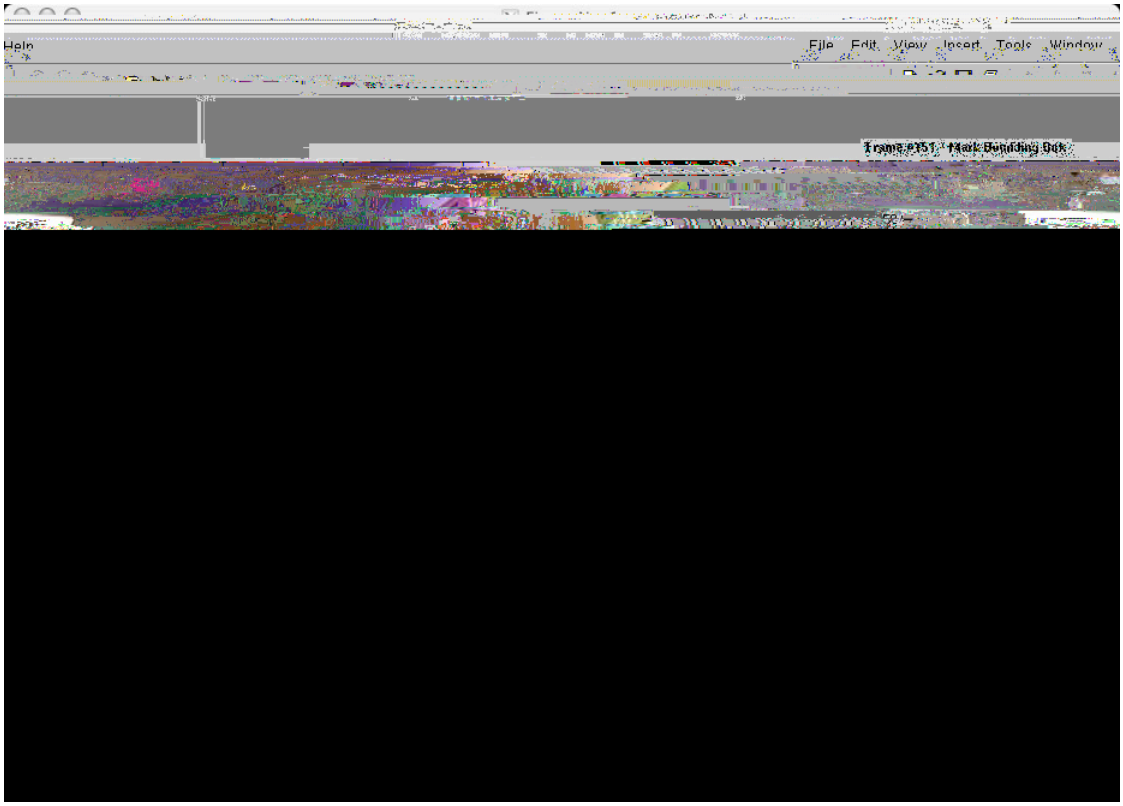
- Street1-7: Camera was placed at knee level facing a sidewalk as people walked by. Outdoors, bright day, but the scene is very backlit (the figures are almost silhouettes.) Winter clothing. Variation in the speeds of the walkers, but it is difficult to differentiate between people because of the lighting conditions. No camera motion or occlusion.



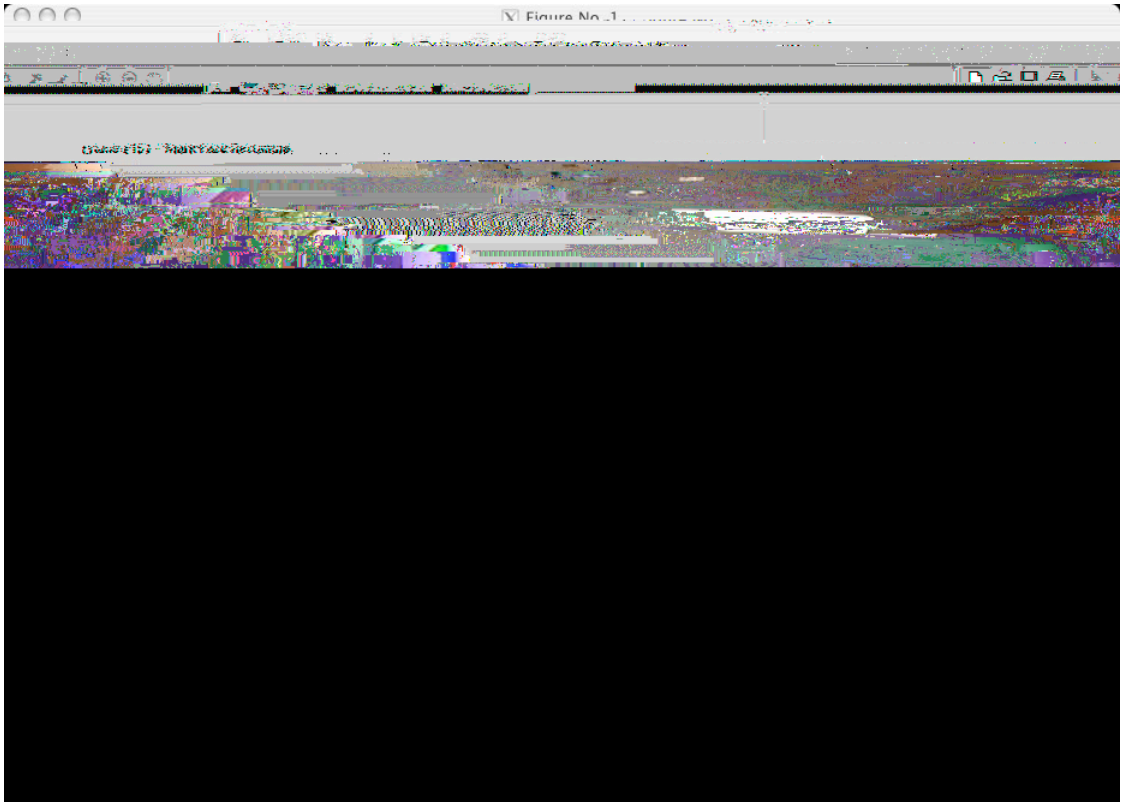
3. Video Annotation

The other aspect of our dataset is that it is human-annotated. We wanted to create data files consisting of annotations done by humans so that there would be a standard to benchmark tracking algorithms by. Ideally, these files would be annotated by many different people to average out any human differences or mistakes. There are many different kinds of annotations that we would like eventually have made, and two that we have implemented at this time. The two annotations that are currently implemented are full-figure tracking bounding boxes around each person in a frame, and face detection

figure-tracking bounding boxes, and a tool for marking faces and eye locations. In both of these, if a particular frame has already been annotated, that data is displayed for the user, and they decide whether to keep the previous annotation or to redo the annotation for that frame. For rectangle-based annotations, the user selects points at the edge of a person, and the program constructs a bounding box based on the minimum and maximum X and Y values of these points. Keyboard commands signal to the program that the user has completed a frame, and wishes to go to the next. For rectangle-based annotating, we currently have undo and redo support for selecting the points used to construct the boxes. Eye points are selected by simply clicking on the eyes. The figures on the next page illustrate the tools in use.



Above: Bounding Boxes around subjects. Below: Face and Eye locations marked:



The data is stored in XML format, making it easy to read by both outside programs, and by humans themselves. There is one xml file for each sequence. All annotations for a particular sequence go in this file. Each frame is an element in the xml tree, and each annotation is a sub-element of the frame. Below is an example of an annotation file (just 1 frame is listed):

```
<?xml version="1.0" encoding="utf-8"?>
<sequence><!-- this is an xml representation of a sequence-->

  <sequenceheader isfaceannotated="true" numboundingboxes="2"
numcomplextracks="1" numfaces="0" numframes="61" numsimpletracks="1"
sequencedescription="girl running" sequencenum="1"><!-- Header info for a sequence--
></sequenceheader>

  <frame annotated="true" face-annotated="true" framenumber="1" numfaces="1"
numpeople="2"><!-- Frames have header info and people-->

    <person bbheight="153" bbwidth="79" bbxmin="3" bbymin="153"
complextracknum="1" personnumber="1" simpletracknum="1"/>

    <person bbheight="186" bbwidth="93" bbxmin="69" bbymin="155"
complextracknum="1" personnumber="2" simpletracknum="1"/>

    <face faceheight="22" facenumber="1" facewidth="19" facexmin="117"
faceymin="170" leftx="-1" lefty="-1" rightx="129" righty="178"/>
  </frame>
```

4. Benchmarking Face Detection

As an example of the potential use of this dataset, we used our annotation data to compare the performance of two different face detection algorithms. As a method of comparison of the effectiveness, we created precision-recall curves based on their results

compared to the actual results from the same clips, human annotated. In detection, there are four possible cases to consider when evaluating an algorithms success rate.

A. True-Positive (Hit): The algorithm detected a face where there was a face annotated in the dataset.

B. False-Negative (Miss): The algorithm did not detect a face where there was one marked in the -t550 450 d

finishing. Since we needed to run it on the same clip with different thresholds, it could take several days before we had results back for a single clip. For that reason, we were

Results of the CMU Face Detector**:**

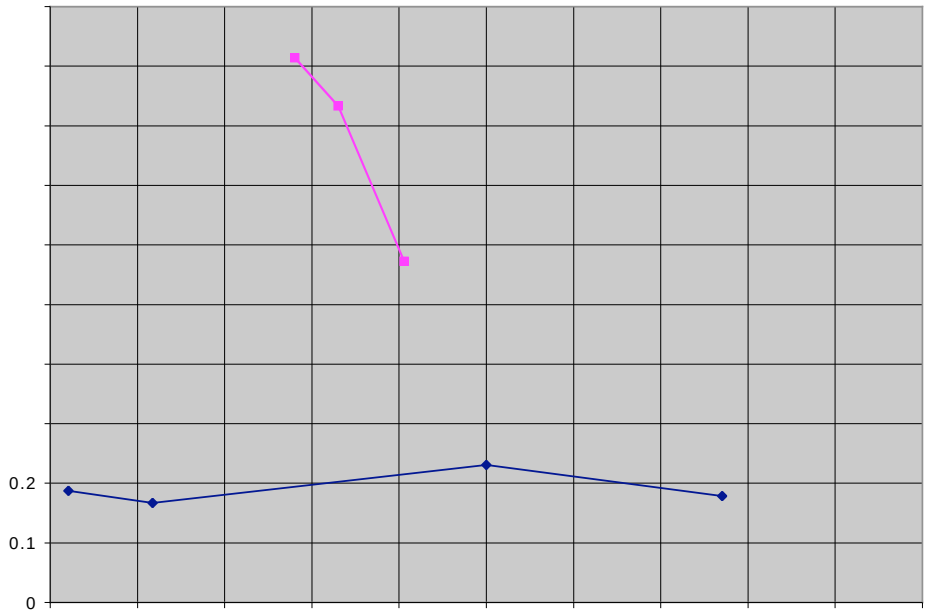
Numerical Results:

A	B	C	Precision	Recall	Clip Name	Threshold
1	5	5	0.166667	0.167	CMU Runner	1
0	6	0	~	0	CMU Runner	1.75
0	6	0	~	0		

our annotation data will be not be detected by this algorithm and regarded as misses, reducing the recall rate.

Comparing the face detectors:

Overlay of PR curves for CMU and for Color Face:



not want to miss any faces, such as detecting terrorists, then the Color Face Detector might be the better algorithm. For an application where you don't want any false positives, then clearly the more precise CMU face detector is the better algorithm. An

objectively the effectiveness of two different detection algorithms, and have shown how this data set can be useful in evaluating many different computer vision problems.