

2009 AVSS Multiple Camera Person Tracking (MCPT) Evaluation Plan

1 Introduction

Technologies that extract information from video sensors are being used for a variety of security applications from a variety of domains such as mass transit monitoring, large entertainment venue security, building security, etc. A key component technology for these domains is the ability to track people as they move through a network of video cameras. Several previous evaluations have focused on different aspects of person tracking challenges including: Classification of Events, Activities, and Relations (CLEAR) [3], Performance Evaluation of Tracking and Surveillance (PETS) [4], TRECVID's Event Detection Track [6].

To continue the effort to develop person tracking technologies, the Advanced Video and Signal Based Surveillance (AVSS) IEEE Conference is sponsoring the Multiple Camera Person Tracking challenge evaluation in conjunction with the Home Office Scientific Development Branch (HOSDB), Centre for the Protection of National Infrastructure (CPNI), and the National Institute of Standards and Technology (NIST).

The goal of the effort is to facilitate research via a common evaluation task that focuses on one aspect of person tracking technologies: the ability to track a specified person within a video sensor field using a small set of *in situ* exemplar video images to specify the person. We refer to these technologies as **Single Person Tracking (SPT)** technologies.

The rest of this document describes the evaluation tasks supported by the evaluation, the data set provided to the participants at no charge, the metrics used to evaluate system performance, data formats, and system submission instructions.

2 The Evaluation Tasks

The evaluation supports three evaluation tasks: Multi-Camera Single Person Tracking (MCSPT), Single Camera Single Person Tracking, and Camera Pair Single Person Tracking. The first task, MCSPT, is the compulsory task that all participants must build systems address. The latter two are voluntary, contrastive evaluation tasks designed assess factors of SPT system performance.

2.1 Multi-Camera Single Person Tracking (MCSPT) Task

The Multi-Camera Single Person Tracking task is to spatio-temporally track a single person as they traverse a multi-

camera field, after the person to track has been specified by five *in situ* video frames. The *in situ* video frames, called **Target Tracking Frame(s) (TTF)**, will be selected from a single camera view and the frames will be the first five annotated frames where the subject is 100% within the frame boundaries (75% for elevator close-up camera) and with no greater than 50% occlusion. The system must then track the person in all video streams from the next frame on.

Implicit in the MCSPT task, systems are expected to be able to re-acquire the subject during camera transitions that may or may not overlap and regardless of whether or not the system loses a track.

2.2 Single-Camera Single Person Tracking (SCSPT) Task

The Single-Camera Single Person Tracking task is to spatio-temporally track a single person as they move through a single camera view after the person to track has been specified by five TTFs. The system must then track the person after the last TTF for the remainder of the video.

This evaluation task is a contrastive condition to determine the system's ability to track the person within camera by factoring out camera-to-camera target re-acquisition of the MCSPT task.

2.3 Camera-Pair Single Person Tracking (CPSPT) Task

The Camera Pair Single Person Tracking task is to spatio-temporally track a single person as they traverse two camera fields of view after the person to track has been specified by five TTFs. The system must then track the person in the pair of video streams.

This evaluation task is a contrastive task focusing on the system's ability to successfully re-acquire the target between a pair of cameras. The "second" camera in the pair may or may not contain images of the person and the camera's field of view may or may not overlap.

3 The Data

The training and test data for the evaluation comes from HOSDB's i-LIDS Multiple Camera Tracking Training Corpus (MCTTR) [1]. The data set was collected at the London Gatwick Airport from a 5-camera airport surveillance field and it has a total of ~44 camera hours of data. The collection consists 107 5-camera excerpt sets from 12 collection epochs. Each excerpt

contains a track for a single person annotated according to the guidelines in the i-LIDS User Guide [2].

The data set will be divided into two subsets for the AVSS MCPT evaluation: a ~29-hour training corpus and a ~15-hour testing corpus. The data will be divided by using the following prioritized factors:

- Excerpts from a collection epoch will not be split across test/train data sets,
- Subject trajectories through the camera network will be represented in both test/train data sets
- Crowd density will be balanced

The MCTTR data set has already been exposed in that some participants may have worked with the data already. As such, participants will be expected to state the extent of their use of the video evaluation data prior to developing a system for AVSS and any system output submitted for AVSS may not make use of the test data for system training/tuning.

4 The Metrics

Systems will be evaluated with several different metrics. The primary evaluation metric is TBD but expected to be one of following:

- F1 (as described in HOSDB Pub 28-08) [2]
- Multiple Object Tracking Accuracy [3, 5]
- Multiple Object Tracking Precision [3, 5]
- Normalized Detection Cost Rate [6]

This section will further describe the protocols and metrics at a later date.

5 System Output Submission Instructions

The packaging and file naming conventions for the AVSS evaluation relies on Experiment Identifiers (EXP-ID) to organize and identify the files for each evaluation condition and link the system inputs to system outputs. Since EXP-IDs may be used in multiple contexts, some fields contain default values. The following section describes the EXP-IDs to be used for the i-LIDS MCTTR dataset.

The following BNF describes the EXP-ID structure:

```
EXP-ID ::= <SITE>_<YEAR>_<TASK>_<DATA>_<LANG>_  
<SYSID>_<VERSION>
```

where,

<SITE> ::= expt | short name of participant's site

The special SITE code "expt" is used in the EXP-ID to indicate an Experiment Control File (see Appendix B).

<YEAR> ::= 2009

<TASK> ::= MCSPT | SCSPT | CPSPT

<DATA> ::= DRYRUN09 | DEV09 | EVAL09

DRYRUN09 refers to the "dry run" for the evaluation, in which systems run on a portion of the training data. DEV09 is used to indicate the training subset, and EVAL09 is used to indicate the testing subset of the i-LIDS MCTTR data.

<LANG> ::= ENG

<SYSID> ::= a site-specified string (that does not contain underscores or spaces) designating the system used

The SYSID string must be present. It is to begin with p- for a primary system or with c- for any contrastive systems. For example, this string could be p-baseline or c-contrast. This field is intended to differentiate between runs for the same evaluation condition. Therefore, a different SYSID should be created for runs where any changes were made to a system.

<VERSION> ::= 1..n (with values greater than 1 indicating multiple runs of the same experiment/system)

In order to facilitate transmission to NIST and subsequent scoring, submissions must be made using the following protocol, consisting of three steps: (1) preparing a system description, (2) packaging system outputs and system descriptions, and (3) transmitting the data to NIST.

System Descriptions

Section 1. Experiment Identifier(s)

List all the experiment IDs for which system outputs were submitted.

Section 2. System Description

A brief technical description of your system; if a contrastive test, contrast with the primary system description.

Section 3. Training:

A list of resources used for training and development.

Section 4. References:

A list of all pertinent references

Packaging Submissions

All system output submissions must be formatted according to the following directory structure:

output/<EXP-ID>/<EXP-ID>.txt

output/<EXP-ID>/<TASK>/tracking_trial_id*/*.xml

where, EXP-ID is the experiment identifier, <EXP-ID>.txt is the system description file as described above, <TASK> is the type of tracking task, tracking_trial_id* is the set of directories corresponding to the "id" attribute of the "tracking_trial" elements in the ECF, and *.xml is the set of ViPER XML files containing the system-generated tracking data.

Transmitting Submissions

To prepare your submission, first create the previously described file/directory structure. This structure may contain the output of multiple experiments, although you are free to submit one experiment at a time if you prefer. The following instructions assume that you are using the UNIX operating system. If you do not have access to UNIX utilities or ftp, please contact NIST to make alternate arrangements.

First, change directory to the parent directory of your "output/" directory. Next, type the following command:

```
tar -cvf - ./output | gzip > <SITE>_<SUB-NUM>.tgz
```

where,

<SITE> is the ID for your site

<SUB-NUM> is an integer 1 to n, where 1 identifies your first submission, 2 your second, etc.

This command creates a single tar/gzip file containing all of your results. Next, ftp to jaguar.ncsl.nist.gov giving the username 'anonymous' and (if requested) your e-mail address as the password. After you are logged in, issue the following set of commands, (the prompt will be 'ftp>'):

```
ftp> cd incoming
```

```
ftp> binary
```

```
ftp> put <SITE>_<SUB-NUM>.tgz
```

```
ftp> quit
```

Note that because the "incoming" ftp directory (where you just ftp'd your submission) is write protected, you will not be able to overwrite any existing file by the same name (you will get an error message if you try), and you will not be able to list the incoming directory (i.e., with the "ls" or "dir" commands). Please note whether you get any error messages from the ftp process when you execute the ftp commands stated above and report them to NIST.

The last thing you need to do is send an e-mail message to jonathan.fiscus@nist.gov, travis.rose@nist.gov, and martial@nist.gov to notify NIST of your submission. The following information should be included in your email: the name of your submission file, the file size, and a listing of each of your submitted experiment IDs.

Please submit your files in time for us to deal with any transmission errors that might occur well before the due date if possible. Submissions must validate using the AVSS scoring tools (part of the NIST F4DE software [7]), or they will be rejected. Note that submissions received after the stated due dates for any reason will be marked late.

6 Data Formats

The reference annotations and system outputs will be ViPER-formatted XML files following the conventions used for the CLEAR person tracking evaluation. Appendix A documents the ViPER XML format. Appendix B documents the Experiment Control File (ECF) format.

7 References

1. Home Office Multiple Camera Tracking Scenario data, <http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/video-based-detection-systems/i-lids>
2. i-LIDS User Guide: <http://scienceandresearch.homeoffice.gov.uk/hosdb/publications/cctv-publications/28-08 - i-LIDS User Guide.pdf>
3. 2007 CLEAR Evaluation Protocol, http://isl.ira.uka.de/clear07/downloads/ClearEval_Protocol_v5.pdf
4. PETS: Performance Evaluation of Tracking and Surveillance, <http://www.cvg.rdg.ac.uk/slides/pets.html>
5. Kasturi, Goildgof, Soundararajan, Manohar, Garofolo, Boonstra, Korzhova, Zhang, "Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 2, Feb. 2009
6. TRECvid Event Detection Evaluation Plan: <http://www.nist.gov/speech/tests/trecvid/2008/doc/EventDet08-EvalPlan-v07.htm>
7. Framework For Detection Evaluations software tools: <http://www.itl.nist.gov/iad/mig/tools/>

Appendix A: ViPER XML Format

This evaluation uses the Video Performance Evaluation Resource (ViPER) XML format, following the conventions developed for the Classification of Events, Activities, and Relationships (CLEAR) evaluation (<http://www.clear-evaluation.org/clear06> and http://www.clear-evaluation.org/clear06/?Evaluation_Tasks). ViPER has previously been used for spatio-temporal annotation of objects in video, and has online documentation (<http://viper-toolkit.sourceforge.net/docs/>). This appendix describes a ViPER XML file format specialized for the AVSS evaluation. The ViPER structure below is used to represent both reference data files and system output data files by restricting the use of some fields as detailed below.

The ViPER hierarchy is divided into “config” and “data” sections. Thus, all metadata files will contain a “config” and a “data” section. For this evaluation task, following the CLEAR conventions, the config section consists of several descriptors: one to represent the relevant source file, one to define a “person” as a ViPER object, one to define a list of frames, and one to define a list of I-frames. An XML fragment showing an example of the config section is as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<viper xmlns="http://lamp.cfar.umd.edu/viper#" xmlns:data="http://lamp.cfar.umd.edu/viperdata#">
  <config>
    <descriptor name="Information" type="FILE">
      <attribute dynamic="false" name="SOURCETYPE" type="http://lamp.cfar.umd.edu/viperdata#lvalue">
        <data:lvalue-possibles>
          <data:lvalue-enum value="SEQUENCE"/>
          <data:lvalue-enum value="FRAMES"/>
        </data:lvalue-possibles>
      </attribute>
      <attribute dynamic="false" name="NUMFRAMES" type="http://lamp.cfar.umd.edu/viperdata#dvalue"/>
      <attribute dynamic="false" name="FRAMERATE" type="http://lamp.cfar.umd.edu/viperdata#fvalue"/>
      <attribute dynamic="false" name="H-FRAME-SIZE" type="http://lamp.cfar.umd.edu/viperdata#dvalue"/>
      <attribute dynamic="false" name="V-FRAME-SIZE" type="http://lamp.cfar.umd.edu/viperdata#dvalue"/>
    </descriptor>
    <descriptor name="PERSON" type="OBJECT">
      <attribute dynamic="true" name="LOCATION" type="http://lamp.cfar.umd.edu/viperdata#bbox"/>
      <attribute dynamic="true" name="AMBIGUOUS" type="http://lamp.cfar.umd.edu/viperdata#bvalue"/>
      <attribute dynamic="true" name="OCCLUSION" type="http://lamp.cfar.umd.edu/viperdata#bvalue"/>
      <attribute dynamic="true" name="PRESENT" type="http://lamp.cfar.umd.edu/viperdata#bvalue"/>
      <attribute dynamic="true" name="SYNTHETIC" type="http://lamp.cfar.umd.edu/viperdata#bvalue"/>
    </descriptor>
    <descriptor name="FRAME" type="OBJECT">
      <attribute dynamic="true" name="EVALUATE" type="http://lamp.cfar.umd.edu/viperdata#bvalue"/>
    </descriptor>
    <descriptor name="I-FRAMES" type="OBJECT"/>
  </config>
  ...
</viper>
```

In the data section, specific details about the source file, person, frames, and I-frames are provided. The source file descriptor includes information about its file location, number of frames, frame rate, etc. The person object contains information about the person’s unique identifier (id), the temporal extent of the person’s appearance in the video (framespan), and location information expressed as a set of ViPER bounding boxes. This is the minimal set of information that a tracking system would need to output for this evaluation. Other person-related features include “ambiguous”, “occlusion”, “present”, and “synthetic”, however these are only used for the purpose of detailed (i.e., reference) annotation and are not evaluated. The frames object contains a list of frames that can potentially be evaluated, and the I-frames object lists all the I-frames in the video. For this evaluation, the frames represented in the frames and I-frames objects are the same. Generally, the “I-frames” object can represent a superset of the frames contained in the “frames” object. Note: in this case, the I-frames object does not actually list the I-frames that occur in the video’s Group of Pictures (GOP); rather, this object enumerates the set of frames that can be annotated. The person, frames, and I-frames objects are all children of the source file object. Finally, each ViPER XML file is limited to a single source file (i.e., one camera view).

An XML fragment showing an example of the data section is as follows:

```

<?xml version="1.0" encoding="UTF-8"?>
<viper xmlns="http://lamp.cfar.umd.edu/viper#" xmlns:data="http://lamp.cfar.umd.edu/viperdata#">
...
<data>
  <sourcefile filename=" MCTTR0101a.mov.deint.mpeg">
    <file id="0" name="Information">
      <attribute name="FRAMERATE">
        <data:fvalue value="1.0"/>
      </attribute>
      <attribute name="H-FRAME-SIZE"/>
      <attribute name="NUMFRAMES">
        <data:dvalue value="16226"/>
      </attribute>
      <attribute name="SOURCETYPE"/>
      <attribute name="V-FRAME-SIZE"/>
    </file>
    <object name="PERSON" id="6" framespan="12385:12385 12390:12390 12395:12395 12400:12400 12405:12405 ...">
      <attribute name="LOCATION">
        <data:bbox framespan="12385:12385" x="39" y="183" width="119" height="379"/>
        <data:bbox framespan="12390:12390" x="52" y="177" width="123" height="377"/>
        <data:bbox framespan="12395:12395" x="68" y="174" width="118" height="368"/>
        <data:bbox framespan="12400:12400" x="82" y="161" width="116" height="356"/>
        <data:bbox framespan="12405:12405" x="93" y="163" width="116" height="354"/>
        ...
      </attribute>
    </attribute>
  </object>
  <object name="I-FRAMES" id="0" framespan="1:1 5:5 10:10 15:15 20:20 25:25 ... 16220:16220 16225:16226"/>
  <object name="FRAME" id="0" framespan="1:1 5:5 10:10 15:15 20:20 25:25 ... 16220:16220 16225:16226">
    <attribute name="EVALUATE">
      <data:bvalue framespan="1:16226" value="true"/>
    </attribute>
  </object>
</sourcefile>
</data>
</viper>

```

Appendix B: Experiment Control Files

Experiment Control Files (ECFs) provide a means for the evaluation infrastructure to specify a set of tracking trials as an experimental condition. A system input ECF will be provided for the evaluation task to indicate which videos should be processed and the target tracking frames. The evaluation code also uses an ECF to determine the range of data to evaluate systems on. In the event a problem is discovered with the data, a special scoring ECF will be used to specify the time intervals to be scored. The ECF is a plain text file in XML format.

For this evaluation, we have divided the MCTTR corpus into training and testing subsets. For the training subset, we provide annotations in several formats: (a) the original XML files from UK Home Office, (b) reformatted ViPER annotation files, compatible with NIST scoring software, (c) starter system files that contain only the target tracking frames, and (d) starter system files that contain no tracking data. For the testing subset, we provide: (a) starter system files that contain only the target tracking frames, and (b) starter system files that contain no tracking data. These files are intended to be used for specific evaluation tasks (multi-camera single person tracking, single-camera single person tracking, or camera-pair single person tracking). Each task will consist of a set of tracking trials that indicate which person to track (specified by five *in situ* video frames). The *in situ* video frames, called *Target Tracking Frame(s) (TTF)*, will be selected from a single camera view, and are represented in one starter system file for each tracking trial (file extension: .ss.xml). Systems will need to extend this starter system file with tracking data for the indicated person. For the other camera views indicated in the tracking trial, systems will need to extend starter system files that contain no tracking data (file extension: .empty.xml). All of the tracking trials and their associated starter system files are encoded in an Experiment Control File (ECF), described below.

ECF Format Description

The ECF consists of the following hierarchically organized XML nodes: "ecf", "version", and one or more "tracking_trial" nodes. In particular, a tracking trial consists of a type (derived from the type of evaluation task), a unique identifier, an interval (expressed in frames), and information about the cameras. Each camera will specify a filename, its camera identifier, a template XML file (which will either be a "starter system" or "empty" ViPER XML file), whether target training frames are provided, any intervals for "don't care frames", and any "don't care regions"; these can have their own intervals (expressed in frames), with the condition that the interval occurs within the tracking trial. Each "don't care region" specifies one or more bounding boxes.

An example ECF for a multiple camera single person tracking task (consisting of one tracking trial) is as follows:

```
<?xml version="1.0" ?>
<ecf xmlns="http://www.itl.nist.gov/iad/mig/avss09ecf#">
  <version>
    Created 20090616
  </version>
  <!-- Multiple Camera Single Person Tracking Task -->
  <tracking_trial type="mcspt" id="MCSPT_0101a" framespan="1300:21000">
    <camera file="MCTTR0101a.mov.deint.mpeg" camid="1" template_xml="MCTTR0101a.ss.xml" target_training="true">
      <dont_care_frames framespan="1300:1500 20000:21000"/>
      <dont_care_region id="0" framespan="1501:15000">
        <bbox framespan="1501:5000" x="39" y="183" width="119" height="379"/>
        <bbox framespan="5001:12500" x="228" y="104" width="78" height="245"/>
        <bbox framespan="12501:15000" x="228" y="400" width="478" height="145"/>
      </dont_care_region>
    </camera>
    <camera file="MCTTR0102a.mov.deint.mpeg" camid="2" template_xml="MCTTR0102a.empty.xml" target_training="false">
      <dont_care_frames framespan="1300:1600 10000:12000"/>
    </camera>
    <camera file="MCTTR0103a.mov.deint.mpeg" camid="3" template_xml="MCTTR0103a.empty.xml" target_training="false">
      <dont_care_region id="0" framespan="1501:2000">
        <bbox framespan="1501:2000" x="39" y="183" width="119" height="379"/>
      </dont_care_region>
      <dont_care_region id="1" framespan="5000:12000">
        <bbox framespan="5000:7000" x="39" y="183" width="119" height="379"/>
        <bbox framespan="7001:12000" x="329" y="13" width="19" height="37"/>
      </dont_care_region>
    </camera>
    <camera file="MCTTR0104a.mov.deint.mpeg" camid="4" template_xml="MCTTR0104a.empty.xml" target_training="false">
    </camera>
    <camera file="MCTTR0105a.mov.deint.mpeg" camid="5" template_xml="MCTTR0105a.empty.xml" target_training="false">
    </camera>
  </tracking_trial>
</ecf>
```

Multiple-camera Single Person Tracking Task

For this task, the ECF will provide a set of tracking trials that specify one camera and its associated starter system file containing the target tracking frames (TTF); the tracking trial will include other camera views that will be associated with starter system files containing no TTF. Systems will need to extend the starter system files by populating tracking data for the indicated person over the tracking trial's interval.

Single-camera Single Person Tracking Task

For this task, the ECF will provide a set of tracking trials that specify one camera and its associated starter system file containing the target tracking frames (TTF). Only one camera view will be included in each trial. Systems will need to extend the starter system files by populating tracking data for the indicated person over the tracking trial's interval.

Camera-pair Single Person Tracking Task

For this task, the ECF will provide a set of tracking trials that specify one camera and its associated starter system file containing the target tracking frames (TTF); the tracking trial will include one other camera view that will be associated with starter system files

containing no TTF. Only two camera views will be included in each trial. Systems will need to extend the starter system files by populating tracking data for the indicated person over the tracking trial's interval.