The author(s) shown below used Federal funds provided by the U.S. Department of Justice and prepared the following final report:

Document Title: Practitioner Centric Video Analytics, Final Summary Overview

Author(s): Ming-Ching Chang, Jixu Chen, Siwei Lyu, Peter Tu

Document No.: 250272

Date Received: October 2016

Award Number: 2013-IJ-CX-K010

This report has not been published by the U.S. Department of Justice. To provide better customer service, NCJRS has made this federally funded grant report available electronically.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Practitioner Centric Video Analytics

Final Summary Overview

Ming-Ching Chang, Jixu Chen, Siwei Lyu, Peter Tu (PI)

GE Global Research

Applied Research and Development Program
Grant Number 2013-IJ-CX-K010
DUNS 0861884010000, EIN 14-0680340
Project/Grant Period: Oct. 1, 2013 to Sep. 30. 2015

Submitted to
U.S. Department of Justice
Office of Justice Programs
National Institute of Justice

Submitted by
GE Global Research
One Research Circle
Niskayuna NY 12309

Technical Point of Contact
Peter Tu
Principal Investigator
Phone: (518)387-5838
tu@ge.com

Administrative Point of Contact
Cheryl L. Sabourin
Business Development Manager
Phone: (518) 387-5378
sabourin@ge.com

Oct.26 2015
Submitted Official: John Burczak
Signature:
John D.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

1. Program Goal

Using practitioner feedback as input, this program focuses on the development of three new video analytics technologies: (1) **3D Video Representation and Event Summarization Front-end**, which allows users to observe view-independent 3D event summarizations with highlights and greater contextual support; (2) **One-shot Learning for Action Recognition**, which allows for the recognition of novel events using a single user-provided example; and (3) **Person Specific Face Recognition**, which enables search against a face catalog by specifying certain distinguishing facial features such as visible scars and hair styles. Practitioner feedback is incorporated into the development cycle. Program results have been disseminated through progress reports, demonstrations, conferences, and peer-reviewed scientific publications.

2. Design, Methods, and Data Analysis

2.1 3D Video Representation and Event Summarization Front-end

The goal of this task is to construct a 3D model of a surveillance site, such that live-feed video analytics meta data (such as tracking and face catalog information) can be visualized from an interactive, user-specific view angle. The representation can also be enriched by adding site-specific ontology information.

The SUNY-GE team has completed a prototype demo system. Starting with a large number of survey photos taken from different view angles and covering parts of a site, we first apply state-of-the-art 3D photometric reconstruction algorithms to calculate a 3D model of the site represented as a dense point cloud. We then apply a coordinate system transform between the virtual coordinate system of the 3D site model to the world coordinate system of the physical site. With such coordinate correspondences, meta data feeds such as results from the tracking modules can be transmitted via a message-passing mechanism and mapped on to the 3D coordinate system of the virtual environment. In this way objects under tracking can be rendered directly into a 3D display environment. Additional meta-data distilled from other video analytics modules (such as face detection and recognition, gaze and affective body pose) can be attached to the synthesized camera view to provide enriched visualization of a given scenario. Figure 1 provides an overview of the framework.

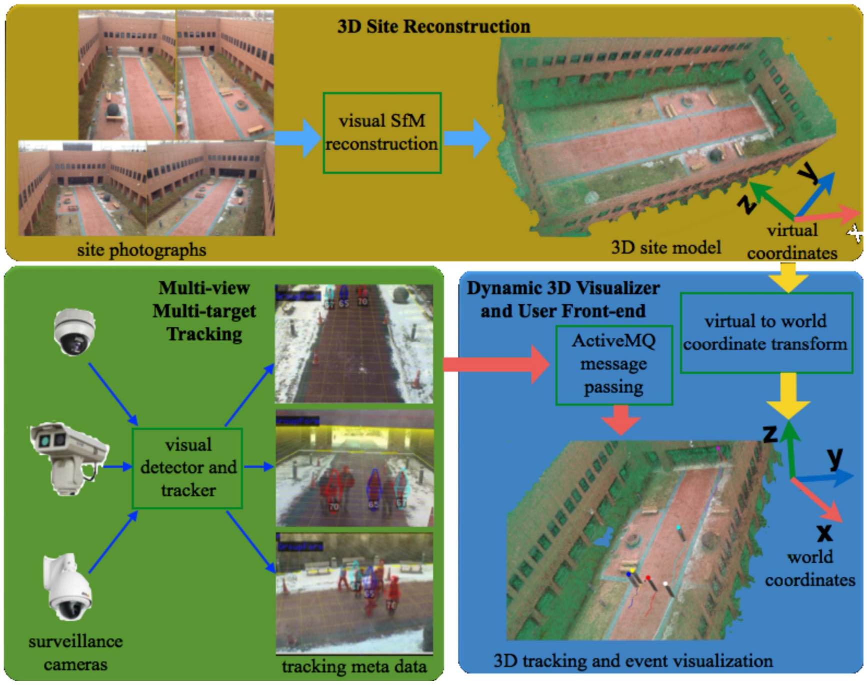
3D Site Reconstruction
visual SfM reconstruction
site photographs
3D site model
virtual coordinates
z
y
Multi-view Multi-target Tracking
visual detector and tracker
surveillance cameras
tracking meta data
Dynamic 3D Visualizer and User Front-end
ActiveMQ message passing
virtual to world coordinate transform
z
y
x
world coordinates
3D tracking and event visualization

Figure 1. Overview of the 3D Video Event Visualization Framework.

To address issues associated with estimating the coordinate transform, we investigated the use of a robust RANSAC scheme in addition to the estimated Singular Value Decomposition (SVD) approach in order to obtain a more reliable 3D model fusion method. As commercial 3D sensors

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

have become pervasive and 3D models more easily obtainable, we investigate the means of integrating external 3D models into our framework, where the coordinate transformation can be estimated using the developed RANSAC-based module. In cases where depth sensors (such as the Microsoft Kinect sensors) are available, we can also visualize live 3D video feeds in the form of dynamic 3D point clouds embedded in the static 3D site model. We have also developed a scalable fusion method, which can repeatedly fuse multiple 3D point clouds using a hierarchical framework to produce large-scale 3D site models for site-wide event visualization. See Appendix A6 for a discussion of the accuracy associated with these experiments.

2.2 One-shot Learning for Action Recognition

The goal of this task was the development of capabilities associated with the recognition of behaviors given as few as a single example (Note that from a social science perspective the term behavior has a more specific interpretation). For review purposes, the various concepts associated with the one-shot-learning paradigm are now defined. **Behaviors**: are sequences of events that are performed by people. **Video Analytic Streams**: data that is generated by base video analytics such as the locations, gaze directions, expressions and motion fields of each person (these are under constant development by the community). **Signal Generators**: various interpretations of the data in terms of a single variant time series ranging between 0 and 1 (See Appendix A4 for more details). A single value known as an *affect score*, ranging between 0 and 1, is used to describe the overall signal. It can be argued that the affect scores can represent a semantically meaningful description. **Signatures and Recognition:** signatures are used to characterize a given behavior. They encode the observed affects. Two signatures can be

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

compared with respect to similarity in a straightforward manner by considering a weighted sum of their affect differences. Recognition is achieved by measuring the similarity scores between a query behavior and all behaviors resident in a database.

Behaviors: For testing and evaluation purposes, thirteen types of behaviors where enacted and recorded. Each behavior involved three individuals and was performed twice resulting in a query and true mate/database instance. The true mates can be viewed as a set of database behaviors. The overall goal of this task is to be able to compare each query behavior with all members of the behavior database with the hope that the highest similarity score occurs when comparing the query against its true mate.

Video Analytics Streams: Each of the thirteen behavior pairs were processed using GE's social interaction analysis system. The site was instrumented with a set of range and PTZ cameras. Each person was tracked and the PTZ cameras were automatically focused onto targeted faces. The output of this module was the location, articulated motion patterns, gaze direction and facial expressions for each individual (see Appendix A3 for a description of these capabilities).

Signal Generators: Under this program, 6 prototype signal generators were instantiated. They focused on: emotion valence, gaze standard deviation, gaze engagement, location proximity, speed and motion magnitude. Each signal generator is associated with a set of parameters that define its performance. Users are allowed to define multiple versions of each signal generator by either selecting parameter values manually or setting them via a random number generator.

Signatures: For the purposes of experimentation, 3 versions of each of the prototype signal generators were instantiated resulting in 18 affect scores for each of the 13 behavior types (See

Appendix A1 for a list of the test behaviors). Initially a uniform affect weighting function was used to compute the similarity scores. By considering all pair-wise similarity scores between the query and database behaviors, a cumulative match characteristic curve was generated to represent overall recognition performance. It was found that 4 of the queries resulted in a top ranking for their true mates and 9 of the queries resulted in a rank of 4 or better for their true mates. By comparison, chance alone would be expected to generate only 1 top ranking. An optimization algorithm was then developed for the purposes of defining more optimal affect weights for the similarity scores. This process was based on a random search through possible weighting values. After 1000 iterations it was found that the average rank of the true mates went from 3.7 down to 2.1. This resulted in 7 out of the 13 queries receiving a top rank for their true mates. In addition 12 of the 13 queries resulted in a ranking of 4 or better for their true mates – (see Appendix A5 for the Cumulative Match Characteristic Curve which provides a more comprehensive measure of system performance).

In summary, a one-shot learning system for action recognition was developed. A comprehensive set of state of the art video analytics was used to characterize a representative collection of human behaviors. A set of prototype signal generators, capable of producing a wide variety of *parametric off spring,* were instantiated. Testing using uniform weights resulted in recognition performance that was markedly better than chance. After a single round of optimization, recognition rates were observed to improve significantly.

2.3 Person Specific Face Recognition

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Traditional face recognition functions by searching a catalog of face images and determining the best match to a query instance. While this method is feasible subject to certain assumptions and limitations, we hypothesis that the recognition accuracy can be further improved by using distinguishing features that are specific to a person of interest (See Appendix A2 for a clarification of the term person specific face recognition).

Person-specific features includes distinguishing features contained within the face region, such as moles, blemishes and scars, and hair features that are located outside of face region. The GE team has developed algorithms to extract these two types of person-specific features:

- **Patch-based features** are extracted by asking a user of the system to select a small rectangle region of interest (ROI) from the probe face image. This ROI should include some unique patterns, such as moles and scars that can distinguish the person from others. The GE team has developed a system with user interface which allows users to select ROIs from a video, and extract texture feature from the ROIs.
- **Hair features are** extracted automatically from a face image using hair segmentation. The GE team has developed an efficient hair-segmentation algorithm that detects hair regions in real-time with high accuracy on both standard and low-resolution images. Given the hair segmentation result, color and texture features are extracted from the hair region resulting in person-specific features.

The GE team has developed a prototype system (Figure 2) to extract the above person-specific features and perform face recognition. In order to compensate for variations in

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

pose, an automatic face alignment/warping pipeline is integrated into the system as a pre-processing step.

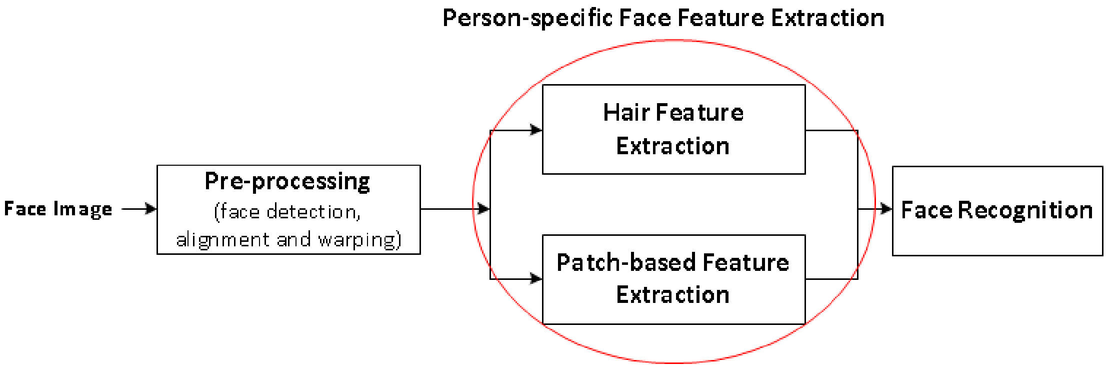
Person-specific Face Feature Extraction
Face Image
Pre-processing
(face detection, alignment and warping)
Hair Feature Extraction
Patch-based Feature Extraction
Face Recognition

Figure 2. System Diagram of the person-specific face recognition pipeline.

The person-specific face recognition system was evaluated on a subset of the Multiple Biometric Grand Challenge (MBGC) dataset. Two types of person-specific features, i.e., patch-based features and hair-based features, and one conventional feature (Local Binary Pattern), were used together for evaluation purposes. The recognition performance is presented as a Cumulative Match Characteristic (CMC) curve (Figure 3). As shown in Figure 3, compared to the baseline with only LBP feature (Face_LBP), the performance is improved by using person-specific features (Patch_features and Hair_features). The recognition rate (rank 1th in the CMC curve) is improved from 49.3% to 55.2% and the rank5 accuracy (i.e., percentage of current identification included in the top 5 candidates) is improved from 74.6% to 83.6%.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

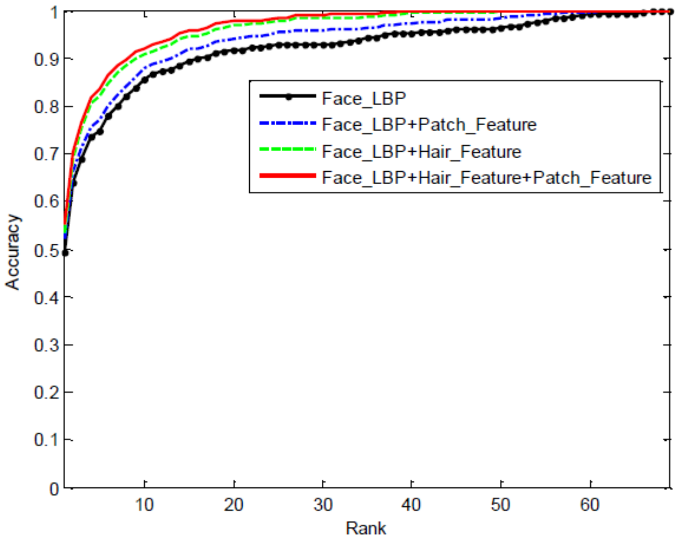

Figure 3. Face recognition CMC curves using different features : LBP features (Face_LBP), patch_based person-specific features (Patch_feature) and hair features (Hair_feature).

3. Scholarly Products and in Process Documents

The following is a list of all publications, manuscripts in submission, and patent disclosures that have been produced based on the output of this program.

1. Yueming Yang, Ming-Ching Chang, Siwei Lyu, and Peter Tu, "3D Video Visualization for Event Summarization", the Forth Greater New York Area Multimedia and Vision Meeting (GNYMVM), New York NY USA, Oct. 2014.
2. Yueming Yang, Ming-Ching Chang, Peter Tu, and Siwei Lyu, "Seeing as It Happens: Real Time 3D Video Event Visualization", IEEE International Conference on Image Processing (ICIP), 2015.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

3. Andrew Pulver, Ming-Ching Chang, and Siwei Lyu, ``Shot Segmentation and Grouping for PTZ Camera Videos", Annual Symposium on Information Assurance (ASIA), Albany NY USA, June 2015.
4. "Real-time Hair Segmentation Using Tiered Structure", submitted to the Workshop on Low-Resolution Face Analysis (LRFA), IEEE international conference on Advanced Video and Signal based Surveillance (AVSS), 2015.
5. Patent Disclosure: "Hair Segmentation Using Tiered Structure". **GE Docket No.:** 277201.
6. Patent Disclosure: "One-Shot Learning for Scenario Action Recognition" in preparation for submission.

4. Implications for Criminal Justice Policy and Practice in the United States

Implications with respect to both policy and practice of this work can be viewed from the perspective of new capabilities that will be enabled due to our successful research agenda as well as insights gained due to direct interaction with the practitioner community. Research results include:

(i) **Video Representation and Event Summarization**: methods associated with 3D site reconstruction and the embedding of video analytic derived meta data such as tracks and face information will provide for a new form of "recognition at a glance" allowing for greater understanding of complex multi-camera imagery.

(ii) **One-Shot Learning for Action Recognition**: armed with this new form of event recognition, analytics engines will be able to detect new types of behaviors with as

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

few as a single example. In this way the surveillance systems of the future will be able to more readily keep pace with the ever-evolving demands on low enforcement.

(iii) **Person Specific Face Recognition**: By considering person specific cues such as hair styles and facial markings, it will become possible to more accurately detect specific persons of interest.

In addition to the research implications of this work, our direct interactions with the police community have resulted in a number of insights regarding the use of video analytics with respect to the day-to-day efforts associated with law enforcement. Methods developed by GE for decomposing tour based PTZ imagery into multiple static videos has resulted in speedier analysis of forensic data. GE's video synopsis methods have allowed for the identification of key events without the need for extensive manual review. GE's facial cataloging and analysis methods constitute a foundation for the systematic analysis of individuals associated with specific sites of interest.

By combining cutting edge research with real-world industrial grade analytics that have been deployed directly with the practitioner community, GE along with its partners SUNY Albany and the Schenectady Municipal Police Department have significantly advanced the field of video surveillance as it pertains to a variety of law enforcement needs.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Appendix

A1: Three Person Behaviors used for Experiments in One-Shot-Learning

Three Person Test Behaviors

Three people (who know each other) just passing through
Strangers passing by
A chance meeting between friends (happy)
Two people are waiting for a third, who is late for the meeting
A group forms, an argument starts and ends in a fight
Two people approach a drug dealer and purchase drugs
Two people are lost and ask a bystander for directions
A game of tag (one person is “it” and tries to tag the others)
Three strangers standing around, one faints and the others try to assist
A pan-handler asks for change
A busker (juggler or musician) gets a tip
A stalker starts to follow a pair of people
Two punks start “tagging” while one is a lookout (graffiti).
A group of people break a bike lock and ride away
A child throws a temper tantrum and the parents try to calm him down
Collision avoidance, one person tries not to collide with a pair of on comers
Follow the leader, a leader emerges and everyone follows him
A loud bang and everyone tries to run away
A gun shot occurs and everyone takes cover
Two bullies try to take lunch money from a nerd

A2: General vs. Person Specific Face Recognition

A general purpose face recognition engine (GPFRE) can be viewed as a discriminative process. Given any pair of images a GPFRE must determine whether or not the two images are from the same individual or

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

from different individuals. In principle a GPRFRE can take any query image and compare it against a database of known individuals for the purposes of determining the identity of query image. In this work we argue that in order to allow for such generic use, the features used to describe a given image must be “universal” in nature. In contrast a person specific face recognition engine (PSFRE) is only concerned with determining whether or not a query image matches the image of a specific person. Given this more limited scope the PSFRE is free to focus on features that while not universally applicable can be used to greater advantage with respect to the task of determining whether or not a query image matches a specific person of interest. For this reason we believe that a PSFRE is different from a GPFRE.

A3: Base Video Analytic Capabilities

With respect to the behavior recognition system, we are leveraging prior work where fixed RGB(D) cameras are responsible for site-wide tracking plus body motion analysis. Pan Tilt Zoom cameras are then responsible for capturing high resolution facial images resulting in facial images in the range of 75 pixels ear to ear. Gaze and expression recognition is performed on such imagery. For more details See “Bridging computer vision and social science: a multi-camera vision system for social interaction training analysis” Jixu Chen, Ming-Ching Chang, Tai-Peng Tian, Ting Yu, Peter Tu , ICIP 2015.

A4: Signal Generator Description

A signal generator is a module that a) will consume a video analytic stream, b) must analyze this stream so as to produce a time series with values ranging from 0 to 1, c) at the completion of each behavior must be able to produce an affect score between 0 and 1) and d) must have a set of parameters that define the behavior of this module. By allowing for a parametric representation for each signal generator, a user can instantiate a particular variant of a given signal generator. Conversely, multiple variants of a signal generator can be produced by considering various permutations of the signal

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

generator parameters. A signal generator bank allows maintains the set of signal generators that will be used to characterize a given behavior.

A5 Cumulative Match Characteristic Curves for One Shot Learning

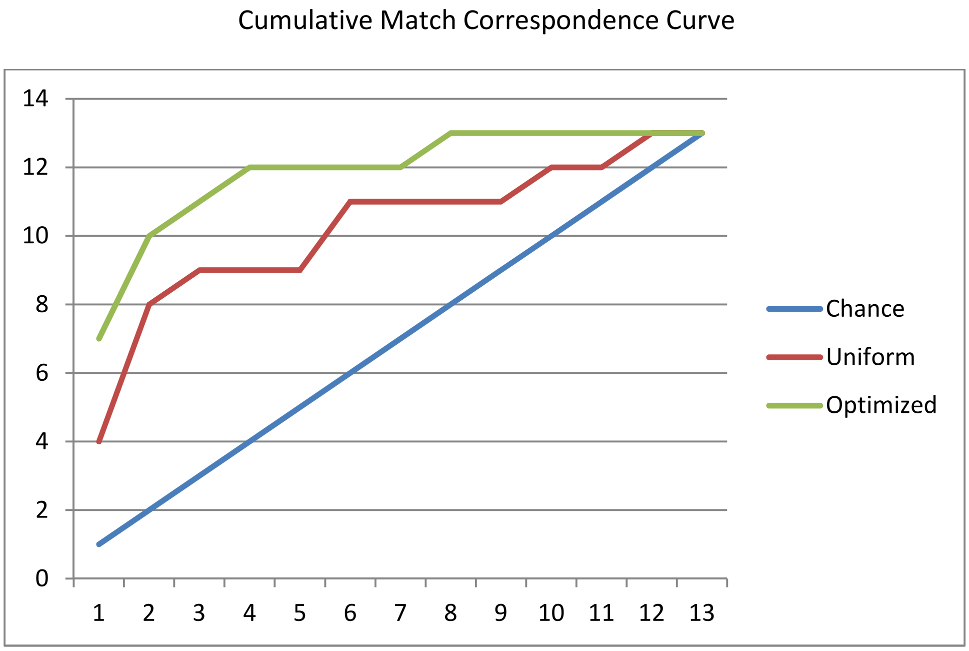

Figure 1: A Cumulative Match Characteristic (CMC) curve for the one shot learning experiments. The y axis defines the number of true matches that receives a rank of x or better. Blue curve: performance that would be expected using chance alone. Red Curve: performance observed using a uniform weighting function w (all signal generators contributing equally). Green: performance observed using an optimized weighting function w.

A6 Accuracy Measurements for 3D Reconstruction

For the evaluation of the effectiveness of the 3D video representation task, we provide an evaluation of the accuracy of one of the 3D models reconstructed from an experiment using sets of 2D images. Specifically, we measure the 3D distances between a set of actual landmark points in world coordinates (meters) and compare the corresponding distances to the synthesized 3D model after the proposed world-coordinate transformation. The ratios between these distances are calculated, where the ideal value of such ratio should be 1 (an exact match). All selected landmarks are on the ground plane (i.e. z=0), as shown in Figure 4. Note that the evaluation is performed on a single selected PTZ view (out of a total of 4), thus only 4 landmarks are used in this evaluation.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

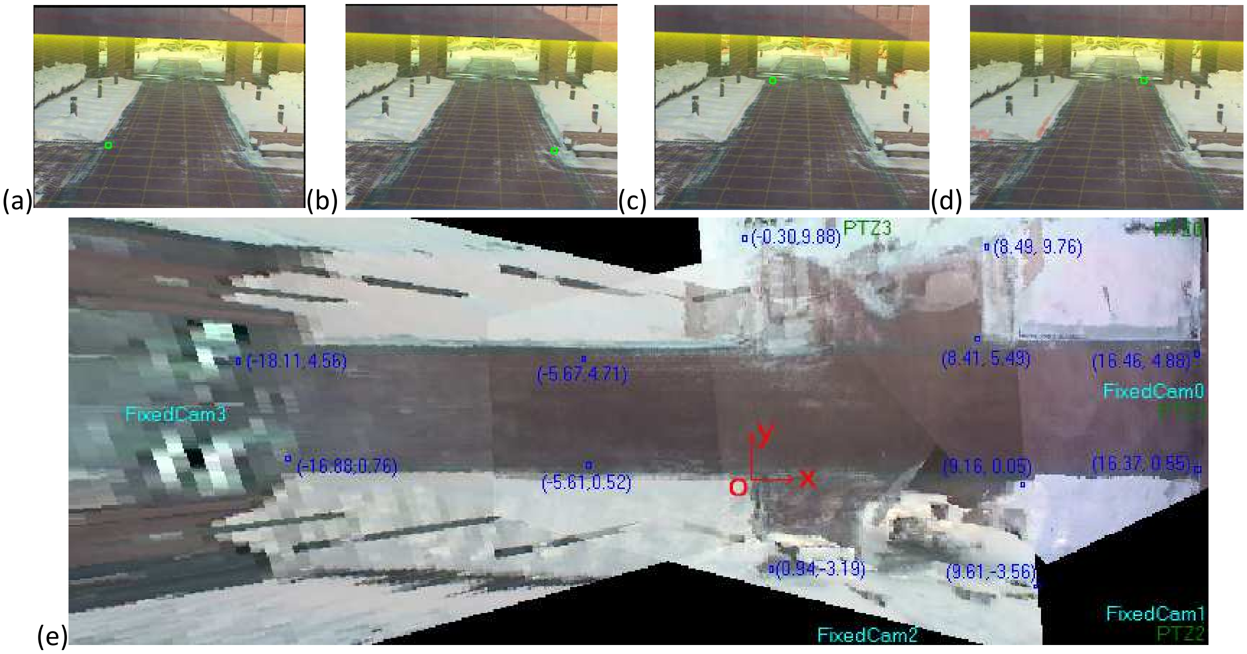

Figure 4. The ground point coordinates of landmark A in (a) is (0.773808, 0.557868), for landmark B in (b) is (0.780372, 4.718262), for landmark C in (c) is (-16.879016, 0.835553), for landmark D in (d) is (-18.110363, 4.562231). Figure (e) shows a synthetic top-down view generated by the fusion of the 4 fixed camera views with selected landmarks.

Table 1 summarizes the resulting evaluation. Given the 4 landmarks points (A, B, C, D), the distances are measured / computed for the 6 pair-wise distances (AB, AC, AD, BC, BD, CD) . The overall average difference in distances is 0.378 (meter) or 37 (cm). The standard deviation of these values is 0.419 (meters) or 41 (cm). Note that landmark point D is in the far field resulting in a relatively large measurement error. The additional error may have been introduced during the manual labeling process. If landmark D is omitted, the average measured error reduces to 0.1 (meter) or 10 (cm).

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Table 1. Comparison of the distances of pairwise landmark points between the world and transformed virtual coordinate systems.

Landmark pairs	Distance measures in world coordinate system	Calculated distance in transformed virtual coordinate system	Absolute difference	Ratio
AB	4.16	4.35	0.19	0.956
AC	17.65	17.71	0.06	0.997
AD	19.30	20.19	0.89	0.955
BC	18.08	18.15	0.07	0.996
BD	18.89	19.83	0.94	0.953
CD	3.92	4.0375	0.12	0.970

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.