Enhancing Re-identification Through Contextual Trajectory Forecasting

Pranav Manttini
Department of Computer Science
University of Houston
Houston, Texas 77004
Email: pmantini@cs.uh.edu

Shishir Shah
Department of Computer Science
University of Houston
Houston, Texas 77004
Email: sshah@central.uh.edu

***Abstract*—Person re-identification (re-ID) is the ability to associate the identity of a person observed at one time and location with the same subject when acquired at a different time and location. Trajectory forecasting is the task of predicting the likely path that a person might take to reach a destination. Contextual trajectory forecasting (CTF) leverages the 3D geometric information of the environment along with observed behavioral norms for human path prediction. Re-ID involves feature matching to find an identity in the database with similar features. The features encompass information regarding appearance of the person like color and texture, or context of the scenario like the time and location of the human subject. CTF provides a future estimate of the likely time and spatial location of previously observed subjects. Embedding this information into traditional re-ID algorithm significantly boost their performance. In this paper, re-ID is performed across non-overlapping cameras with real world human subjects. CTF is embedded into a re-identification algorithm that uses symmetry driven accumulation of local features (SDALF) [1] to evaluate the performance. Experiments suggest a significant improvement in re-identification by embedding CTF.**

***Keywords*—Human trajectory, motion forecasting, re-identification.**

I. Introduction

Given the observation of an individual from different cameras, over disparate time and location, automated re-identification algorithms deal with the task of associating the identity correctly of the individual across all observations. Re-ID is an everyday trivial task for human beings. Replicating such system is a confounding task because of various difficulties originating from low quality images, occlusion, changes in illumination, view, and pose across cameras [2]. Furthermore the lack of robust algorithms to infer the topology of the network and calibrate camera locations for leveraging contextual information complicates this process.

Networked cameras are widely used for monitoring human activity in public areas. Camera networks spanning from hundreds to thousands of cameras per network is a common occurrence in busy public locations like airports. These camera networks generate massive quantities of video data. At the time of need, manually fishing for a single human subject from the sea of data is a tedious and time consuming task. Re-ID algorithms find a natural place in these scenarios. Given the videos from these networks, the task of re-ID performed by security personnels though tedious is still extremely reliable. To design a re-ID algorithm we take motivation from how these security personnels might use a combination of appearance features along with contextual information to identify subjects in the videos. For example, if a person is observed in a particular video, the human performing re-ID will notice information such as the color of clothing, the direction of motion and their velocity (run or walk). The human performing re-ID has knowledge regarding the geometry of the environment, the topology of the camera network. This knowledge can be used to process the observed information to arrive at an estimate of the likely future time and location of the observed person. This estimate will allow the human to only search for a small window of time in specific locations for a match. Hence using information regarding the geometry of the environment can be of vital assistance in re-ID.

This paper showcases a method to perform human trajectory forecasting based on the geometry of the environment and observed human behavioral norms and then leverages these predictions to assist in re-ID. We employ data collected from real world surveillance cameras with non overlapping field of view to evaluate the performance of this algorithm.

II. Related Work

Considerable amount of research has been performed in the area of re-ID. Re-ID problems are widely viewed as recognition problems. A database of known identities is called a gallery set. Given an observation whose identity is unknown called a probe, the goal of re-ID algorithm is to rank the identities in the gallery set based on a similarity score to the probe. Re-ID approaches can be broadly categorized as appearance based methods or context based methods. The former only uses information regarding the appearance like color and texture to construct features that describe an identity for matching, while the latter often augments this information with context like spatial and temporal data to match with the gallery set. This paper can be categorized under the latter. A complete survey of re-ID was conducted by Gala and Shah [2].

Appearance based methods are more commonly considered than context based methods. This can be attributed to the unavailability of the environmental geometry and camera topography for commonly used public datasets. The proposed method suggests a technique for embedding this information to build context based re-ID algorithms. A complete survey

of appearance based methods was conducted in [3] by Satta. A huge body of work exists that employ different appearance based features like color, texture, gradient and shape for re-ID, few of which are [4], [5], [6], [7], [8], [9], [10]. Since this work employs a contextual based method, they are discussed further in detail. However, the proposed algorithm is used in conjunction with an appearance based method suggested by Bazzani *et al.* in [1], which constructs a symmetry based description to characterize the human body for re-ID.

The need to associate trajectories across multiple-camera network for tracking contributed to the genesis of contextual re-ID algorithms. These methods try to understand the relationship between the cameras in a surveillance network to estimate the space or time dependency among the observations for re-ID. Makris *et al.* in [11] suggested the need for "network calibration", that describes the association among cameras for tracking across non-overlapping cameras. The network topology is represented as a graph with nodes representing the entry/exit zones of the cameras present on the network, and the edges represented the transition time and probability between the nodes. Observed trajectories are later used as training data to learn these transition time probabilities. Javed *et al.* in [12] proposed a method for tracking people across non-overlapping cameras by learning the inter-camera relationship through exploiting the space-time cues between them. These relations are learned in the from of probability density functions of space time parameters using kernel density estimators. In [13] the camera images are represented as time series data and then segmented into regions of similar activity. Inter-region time delay are inferred using Cross Canonical Correlation Analysis. Loy *et al.* also followed a similar approach but modeled the dependency between the regional patterns as Time Delayed Probabilistic Graphical models in [14]. Mazzon *et al.* in [15] proposed Landmark-Based model (LBM) using a rough site map, made up of the projection of the camera's field of view, the unobserved regions, marked entry/exit zones of the cameras and crossing landmarks. Human trajectories are propagated along possible paths connecting the located landmarks. Using the initial observed velocity, an estimate of the time taken for traversal is calculated and used to filter the gallery set for re-ID.

In the proposed method, a complete 3D model of the camera network environment was constructed and the cameras in the real world were calibrated and then embedded as virtual cameras in the model. This steps eliminates the need for training to learn the camera network topography or relationship between the cameras. Furthermore, the trajectory forecasting model for propagating humans is based on observed human behavioral norms in contrast to LBM which employs a purely random approach.

III. Method

Let $G = \{g_1, g_2, ..., g_m\}$ be the gallery set of m known identities, and $P = \{p_1, p_2, ..., p_n\}$ be the probe set of n unknown identities. For every probe $p_i \in P$, the problem is to rank the gallery set as $\{g_{i1}, g_{i2}, ..., g_{im}\}, g_{ij} \in G$ based on their matching score to the probe p_i. Let $f_x = \{c_x, a_x\}$ be the features of the identity $x \in \{G \cup P\}$, where c_x are the contextual feature and a_x are the appearance features. Let $c_x = \{l_x, v_x, t_x\}$ be the contextual features of x observed at location l_x traveling with velocity v_x at time t_x. This paper describes a method for re-ID by leveraging contextual features to be used in conjunction with an existing appearance based method and hence a_x is described by the chosen method. The matching function $M_{ij} = M(p_i, g_j) = M(f_{p_i}, f_{g_j}) = M\left(s_c(c_{p_i}, c_{g_j}), s_a(a_{p_i}, a_{g_j})\right)$ calculates the matching score of the probe p_i to gallery item g_j, where s_c and s_a are scores estimated on the contextual and appearance features, respectively. The gallery items for the probe p_i are ranked as $\{g_{i_1}, g_{i_2}, ..., g_{i_m}\}$ such that $M_{i_1 j} < M_{i_2 j} < ... < M_{i_m j}$.

The core of the paper deals with estimating the score s_c. Let p be the probe with contextual features $c_p = \{l_p, v_p, t_p\}$ to be compared with gallery item g with feature $c_g = \{l_g, v_g, t_p\}$. Assuming that a human subject was first observed in gallery set and later again in the probe set, let the trajectory that the subject has taken from the location l_g at time t_g to reach the location l_p at time t_p be $T_{gp} = \{({l_1}^{gp}, {t_1}^{gp}), ({l_2}^{gp}, {t_2}^{gp}), ..., ({l_r}^{gp}, {t_r}^{gp})\}$ such that $({l_1}^{gp}, {t_1}^{gp}) = (l_g, t_g), ({l_r}^{gp}, {t_r}^{gp}) = (l_p, t_p)$. If the trajectory was known, the probe p can be associated with the correct gallery item $g^{'}$ by traversing it in space and time. It is not possible to observe the trajectory across non-overlapping cameras. Hence, given the geometry, the idea is to predict the trajectory using CTF from the gallery set to the probe, to find the best match in space and time. CTF provides a prediction for $T^{'}_{gp}$ from which the contextual score $s_c(p)$ can be calculated.

A. Contextual Trajectory Forecasting

Given the 3D geometry of the environment and the starting point and destination of a human, CTF is assembled on two assumptions. First, the person would follow a path that requires the shortest time to reach the destination, and second, the person would adhere to certain behavioral norms that are observed when walking. Let $L = \{l_1, l_2, l_3...\}$ be set of all points on the ground plane on which the trajectory forecasting is being performed. Given $(l_g, l_p) \in L$ the starting and destination points, CTF assigns probabilities to $l_i \in L$, such that consecutive points can be sampled from l_g to l_p, to form a trajectory that represents the shortest path while conforming to observed behavioral norms.

1) Distance Map: The CTF algorithm takes as input a distance map to find points that are closer to the destination l_p. This map calculates the distance to the destination l_p from every other point on the floor. Euclidean distance between two points is not altered by the presence of inaccessible areas in the path. Hence using Euclidean distance can potentially be erroneous. Martinez *et al.* defined geodesic distances in [16],

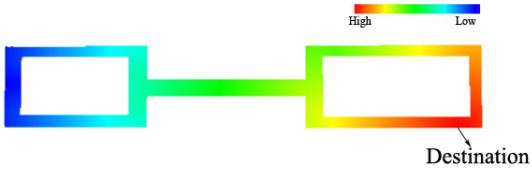

Fig. 1. Distance map for geometry A with a given destination.

which is used instead. Geodesic distance is measured around the inaccessible areas along the ground plane and gives a more accurate sense of distance for human navigation. A rendering

of the distance map for geometry A with a given destination is shown in Fig. 1.

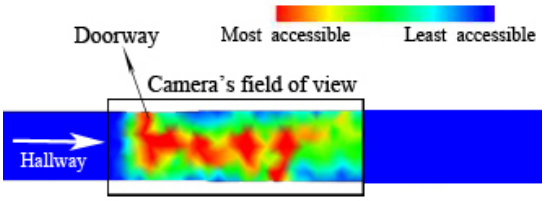

Fig. 2. Observed occupancy map of a hallway in a building from a video observed over 5 days.

2) Human Occupancy Map: CTF also takes a human occupancy map as a second input. Hypothetically, if a large number of trajectories followed by human subjects from l_g to l_p are observed, it is possible that certain points on the ground plane are accessed more often then other points. This might imply the existence of a certain distribution or an occupancy map to the points on the ground plane. Estimating this map can assist the CTF in choosing points that are accessed often by complying with behavioral norms. Human motion is influenced by a multitude of factors, many of which are driven by perception. CTF specifically focuses on modeling this human motion by taking into account the constraints imposed by 3D geometry and the physical world. When traversing on the ground plane, the immediate decision of movement is influenced by the objects in the path and the surrounding geometry like walls. For example, the way humans navigate around tables and chairs when moving from one corner of a classroom to the opposite corner. Since the human behavior is assumed to be influenced by the 3D geometry, the aim is to model the relationship between them. This model would provide a means to estimate the human behavior or occupancy map for any novel location based on its 3D geometry.

This relationship between them was modeled based on empirical data. The model first represents a point on the ground plane using a set of geometric features that capture the 3D geometry of the environment surrounding that point. Then establishes a linear relationship between geometric features of the point and its observed occupancy. Initially, the occupancy map of a known geometry was observed. Consider a dataset of humans traversing the ground plane whose surrounding 3D geometry is known. The occupancy of the point l_i is proportional to the number of times the humans in the dataset has accessed that point. The observed occupancy map in a hallway over a period of 5 days is shown in Fig. 2. Since CTF assumes that the occupancy of a point on the ground plane is influenced by the 3D geometry surrounding it, the geometric features gf_i of any point l_i on the ground plane in the 3D model are represented as a set of numbers $\{d_{i1}, d_{i2}, d_{i3}...\}$, which are its distances from the walls and objects surrounding. So, to obtain the geometric features the distances are measured to walls or objects in the hallway along vectors pointing at a certain inclination from the ground plane at regular interval spanning an entire circle with its tail fixed at the point l_i as shown in Fig. 3. The distances are measured consistently in either clockwise or anti-clockwise direction always starting from the closet object or wall. In order to confine the effect to only objects with in the close vicinity of the point, the distances are thresholded by a hemisphere as shown in Fig. 3.

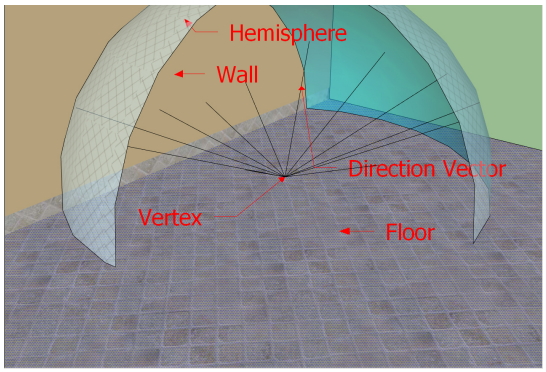

Fig. 3. Geometric features.

The radius of this hemisphere is inferred from the Theory of Proxemics [17]. This is a theory based on observation that defines how human beings unintentionally make use of physical space around them. Proxemics classifies the space close to a human subject into four broad regions, intimate, personal, social and public distance. It is assumed that the interaction between human subjects in closed hallways take place within the social distance.

We assume a linear relationship between the geometric features $gf_i = \{d_{i1}, d_{i2}, d_{i3}...\}$ of the point l_i and its observed occupancy o_i and can be modeled as

$$\begin{aligned} o_i &= \beta_1 d_{i1} + \beta_2 d_{i2} + ... + \beta_n d_{in} + \epsilon_i = GF^T{}_i \beta + \epsilon_i \\ O &= GF\beta + \epsilon \end{aligned} \tag{1}$$

To estimate the values of β we minimize the sum of squares of the error term ϵ, which would give us.

$$\beta = (GF^T GF)^{-1} GF^T O \tag{2}$$

To determine the occupancy of any point on the ground

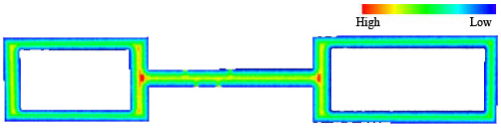

Fig. 4. Estimated occupancy map.

plane in a new geometry, first the geometric features of that point are computed and then the estimated β values are used to estimate occupancy using Equation 1. Fig. 4 depicts the estimated occupancy.

It can be observed how the occupancy of the points in the center of the hallway is higher than those along the edges. The rotational invariance of the features allow for the expected estimation of the occupancy even along curved hallways.

3) Trajectory Forecasting: CTF combines these two maps and assigns an energy value to every point on the ground plane. Let O be the occupancy map function and let D be the distance map function. Then the energy of the point l_i is defined by the function E as:

$$E(l_i) = -D(l_i)/O(l_i) \tag{3}$$

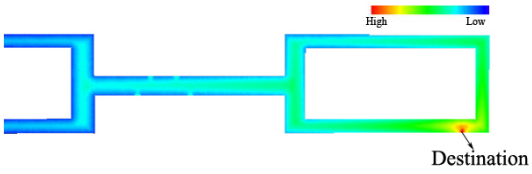

Fig. 5. Energy map.

The energy function for geometry A is shown in Fig. 5. The energy is higher in the center of the hallway than along the edges, and the energy increases as the points get closer to the destination. To forecast the trajectory from the starting point l_g to destination point l_p, points are sampled consecutively with a probability defined by the energy map. If the current state is l_c, the point l_i is chosen if and only if it is closer to the destination, that is $D(l_i) <= D(l_c)$. This ensures the trajectory propagation, without getting stuck in local maximums. The points closer to the destination are sampled with a probability which is proportional to the difference in there energies. So $P(l_i|l_c)$ is

$$\propto \left\{ \begin{array}{cc} E(l_i) - E(l_c) & \text{if} \quad D(l_i) - D(l_c) \leq 0 \\ 0 & otherwise \end{array} \right\} \tag{4}$$

Let the points predicted by CTF from l_g to l_p be $\{l_g, l_2, ..., l_p\}$. Assuming that the human subject moves at a constant velocity v_g, the time t_i taken to reach location l_i from l_g can be estimated as $t_i = t_g + \frac{d(l_i,l_g)}{v_g}$, where $d(l_i, l_g)$ is the length of the trajectory from l_g to l_i. CTF predicts an estimate of the trajectory from gallery g to probe p as $T^{'}_{gp} = \{(l_g, t_g), (l^{'}_2, t^{'}_2), ..., (l_p, t^{'}_r)\}$. The contextual score of the probe p and gallery g are defined as

$$s_c(p) = t_p; s_c(g) = t^{'}_r \tag{5}$$

4) Re-Identification using SDALF and CTF: SDALF is a symmetry based description of the human body. In SDALF, the asymmetry principles allows the segregation of meaningful body parts (head, upper body and lower body). The symmetry criteria helps in extracting the actual appearance features. SDALF uses three different appearance features. First a HSV histogram is used to capture the global chromatic content, second, Maximally Stable Color Regions (MSCR) is used to capture the pre-region color displacement and finally Recurrent Highly Structured Patches (RHSP) are estimated by a per-patch similarity analysis. Let $s_a = \{s_a^{WHSV}, s_a^{MSCR}, s_a^{RHSP}\}$ be the appearance score values. If $\{d_{WHSV}, d_{MSCR}, d_{RHSP}\}$ be the distance functions that calculate the HSV, MSCR and RHSP distance between the probe and gallery items, then SDALF matching distance is defined as convex combination of these features.

$$\begin{aligned} d(p,g) = \gamma_{WHSV} \cdot d_{WHSV}(s_a^{WHSV}(p), s_a^{WHSV}(g)) + \\ \gamma_{MSCR} \cdot d_{MSCR}(s_a^{MSCR}(p), s_a^{MSCR}(g)) + \\ \gamma_{RHSP} \cdot d_{RHSP}(s_a^{RHSP}(p), s_a^{RHSP}(g)) \end{aligned} \tag{6}$$

Where γ are the weighting parameters. The contextual distance function is defined as $d_{CTF}(p,g) = d_{CTF}(s_c(p), s_c(g)) = |t_p - t^{'}_r|$, $t^{'}_r$ and t_p are as defined in Equation 5. The CTF distances were normalized such that $d_{CTF} \in \{0, 1\}$. The CTF score is embedded in Equation 6 as:

$$\begin{aligned} d(p,g) = \gamma_{WHSV} \cdot d_{WHSV}(s_a^{WHSV}(p), s_a^{WHSV}(g)) + \\ \gamma_{MSCR} \cdot d_{MSCR}(s_a^{MSCR}(p), s_a^{MSCR}(g)) + \\ \gamma_{RHSP} \cdot d_{RHSP}(s_a^{RHSP}(p), s_a^{RHSP}(g)) + \\ \gamma_{CTF} \cdot d_{CTF}(s_c(p), s_c(g)) \end{aligned} \tag{7}$$

In our experiments, we fix the values of the parameters as follows: $\gamma_{WHSV} = 0.03, \gamma_{MSCR} = 0.03, \gamma_{RHSP} = 0.03, \gamma_{CTF} = 0.9$. These values seems to provide the best performance. The high value of γ_{CTF} compared to other parameters allows for temporally constraining the data and then trying to find the best match using SDALF within the temporally constrained data. The matching function ranks the gallery items for probe p as $\{g_{p_1}, g_{p_2}, ..., g_{p_m}\}$ such that $M_{g_1 p} < M_{g_2 p} < ... < M_{g_m p} \equiv d(p, g_1) < d(p, g_2) < ... < d(p, g_m)$.

IV. Experiments

A. Implementation

This section describes how a complete 3D model of the environment can be constructed, and furthermore how cameras in the real world are calibrated and then embedded as virtual cameras in the model.

1) Modeling 3D environment: The 3D geometry of the environment like floors, walls, hallways, etc. are modeled using Google Sketchup, a 3D modeling tool. Figure 6 depicts the 3D model of a building constructed using existing floor plans to

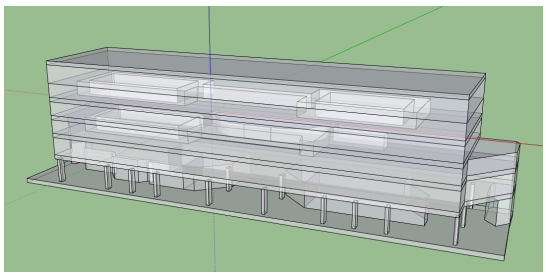

Fig. 6. Model of a building using Google Sketchup.

obtain the measurements and dimensions. The 3D model is then exported using a *common digital asset exchange format* [18] called COLLADA file format. COLLADA Document Object Model (DOM) library is used to load and save this 3D model into an application, and then OpenGL is used to interact with this 3D data in the application.

2) Embedding virtual cameras and calibration: To create virtual cameras in the 3D model that represent cameras in real world. First the internal camera parameters of the existing real world camera are determined by using a general calibration approach with a checkerboard. These parameters are used to create virtual cameras which render perspective projections of

the 3D model that are conceptually equivalent to the real world cameras. Now in order to determine the location and orientation of the camera in the 3D model, the image from the real world camera and manually registered with the corresponding camera's perspective projection in the 3D model, by manually changing the parameters in the transformation matrix using OpenGL. When the images register as shown in Figure 7, the transformation matrix of the camera is extracted which gives us the approximate location and orientation of the camera in the 3D model [19].

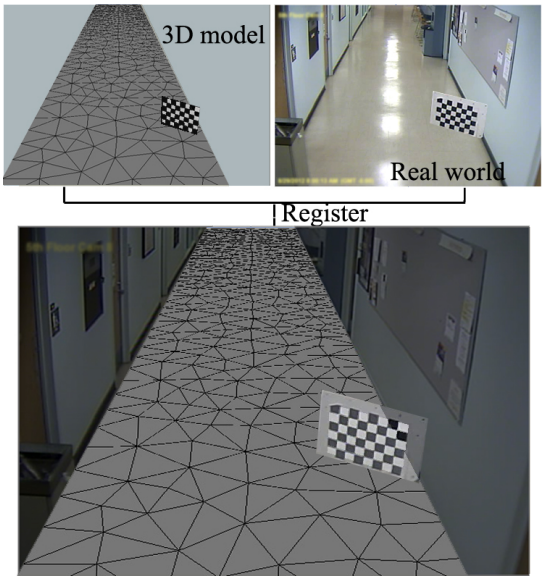

Fig. 7. Manual registration of an image from a camera with the perspective rendering of the 3D model to extract the transformation matrix. The floor is represented by a uniform triangle mesh obtained by Delauney triangulation.

3) Delaunay triangulation of the floor mesh: The ground plane is represented as a triangular mesh though other representation are possible. Delaunay triangulation is used to obtain a uniform triangular mesh as shown in Figure 7. An implementation of the Delaunay triangulation is available in the Computational Geometric Algorithms Library (CGAL) [20]. The centroids of the triangles are considered as points on the ground plane.

4) Projecting points on the image into the 3D geometric model: Human subjects captured from videos are projected into the ground plane in the 3D model, to obtain their global position. The location and orientation of the camera is available in the transformation matrix. The point on the image where the humans feet touches the ground plane is located and using the cameras parameters are projected on the ground plane.

B. Experiments

Over the years many datasets like CAVIAR [21] and VIPeR [6] have been used for evaluating re-ID algorithms, but none of these datasets are equipped with the environments geometry and camera calibration. To evaluate the performance of the proposed method, real world data was collected from two different geometries. Each geometry consisted of three cameras with non-overlapping views in a hallway as shown in Fig. 8 (geometry A) and 9 (geometry B). Human subjects were allowed to walk down the hallway starting from camera 1 and are allowed to randomly choose between making either a left or right to show up in either camera 2 or 3 respectively. The images from camera 1 were used to create the gallery set and camera 2 and 3 were used to create the probe set. To simulate a real world environment, groups of subjects were allowed to start walking at the same time from camera 1. The evaluation was performed on 38 subjects, 26 of which were used in geometry A and 12 in geometry B. In geometry A, 10 groups containing two subjects started at the same time and the rest started individually, and in geometry B, 2 groups of 4 subjects and 1 group of 3 subjects started at the same time and the rest individually. In both the geometries, half of them were captured in camera 2 and the other half in camera 3 to create the probe set. For each ID, 5 shots were captured in all cameras. So the gallery set in geometry A consisted of 130 images and geometry B consisted of 60 images.

To perform re-identification for a given probe, we perform CTF from every image in the gallery. The starting position is determined by the position of the subject in the gallery and the end being determined by the position of the subject in the probe. These points are re reprojected into the 3D model as described in IV-A.

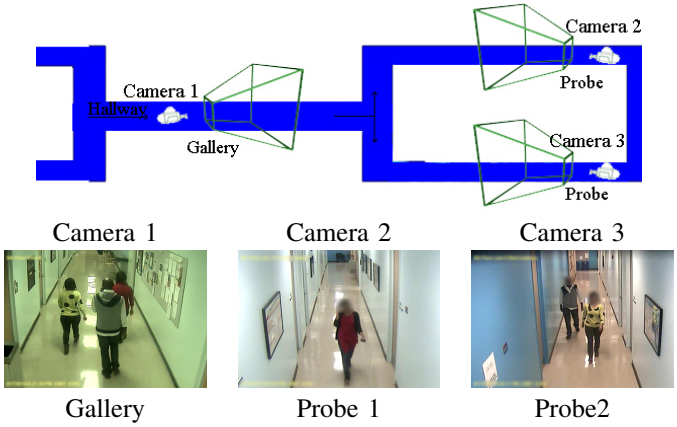

Fig. 8. Geometry A experimental setup.

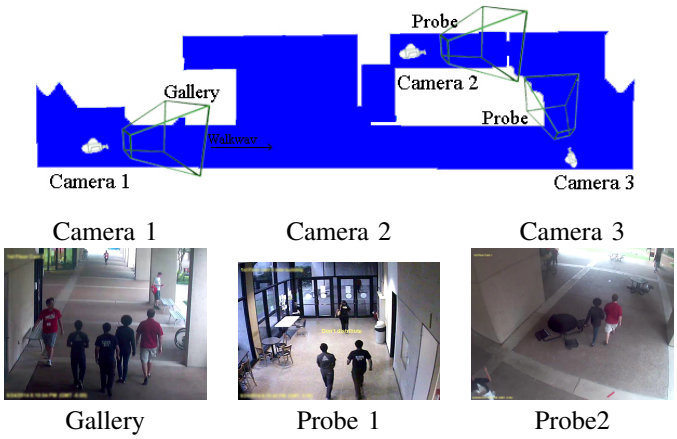

Fig. 9. Geometry B experimental setup.

Experiments were performed in four different modes based on the number of shots used for calculating the scores. In

single-shot vs single-shot (SvsS), each image in a set represented a different ID, in single-shot vs multiple-shot (SvsM), each image in gallery set is different ID but in the probe set, the scores from multiple shots of the same id were average out. In multiple-shot vs single-shot (MvsS), every shot in probe was compare to multiple shots belonging to the same ID in the gallery and the scores were averaged out, finally in multiple-shot vs multiple-shot (MvsM) multiple shots were used in both the gallery and probe set for matching. The results are presented in the form of recognition rate using Cumulative Matching Characteristic (CMC) curves.

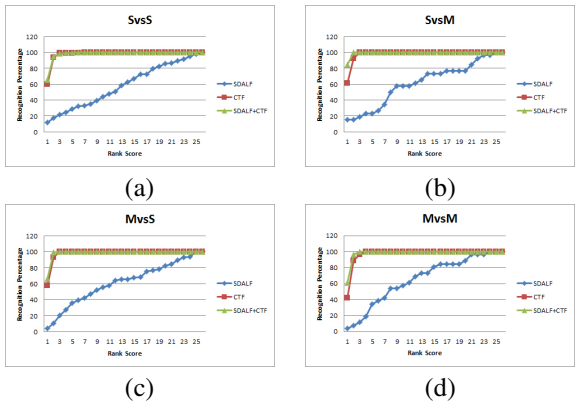

Fig. 10. CMC curves: Geometry A

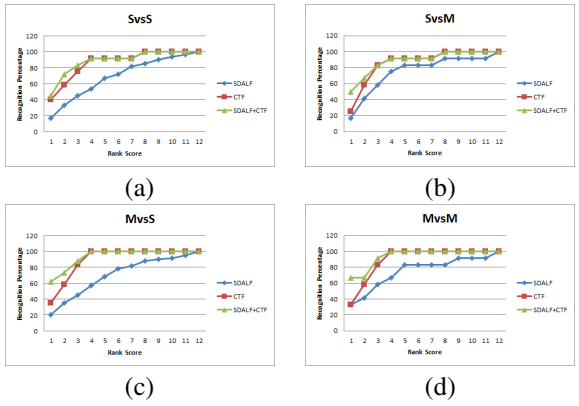

Fig. 11. CMC curves: Geometry B

In geometry A at most two subjects were allowed to start at the same time and hence a 100% recognition was obtained within the first two ranks in all the modes. Similarly in geometry B at most four subjects were allowed to start at the same time and hence a 95-100% recognition was obtained within the first four ranks in all the modes. In all the cases, in was observed that using CTF alone generated a significant boost in recognition over SDALF, and embedding CTF in SDALF generated a further enhancement in recognition performance over CTF.

V. Conclusion

We have implemented a model to construct 3D geometry of the environment and embed virtual cameras for the purpose of surveillance. We have implemented a methodology to predict the future position of human subjects using contextual trajectory forecasting. Finally we have successfully embedded the CTF into a traditional appearance based re-identification algorithm. Preliminary results show that using the 3D geometry and contextual trajectory forecasting can enhance re-identification performance significantly. A Large scale study will be taken into consideration in the future.

References

[1] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, 2013.

[2] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, 2014.

[3] R. Satta, "Appearance descriptors for person re-identification: a comprehensive review," *CoRR*, vol. abs/1307.5748, 2013.

[4] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, 2010.

[5] M. S. L. B. Dong Seon Cheng, Marco Cristani and V. Murino, "Custom pictorial structures for re-identification." BMVA Press, 2011.

[6] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision ECCV 2008*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008.

[7] W. Schwartz and L. Davis, "Learning discriminative appearance-based models using partial least squares," in *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, 2009.

[8] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2010.

[9] R. Layne, T. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *Computer Vision ECCV 2012. Workshops and Demonstrations*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012.

[10] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[11] D. Makris, T. Ellis, and J. Black, "Bridging the gaps between cameras," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2004.

[12] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views," *Computer Vision and Image Understanding*, 2008.

[13] C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *International Journal of Computer Vision*, 2010.

[14] C. Loy, Chen, T. Xiang, and S. Gong, "Incremental activity modeling in multiple disjoint cameras," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012.

[15] R. Mazzon, S. F. Tahir, and A. Cavallaro, "Person re-identification in crowd," *Pattern Recogn. Lett.*, 2012.

[16] D. Martínez, L. Velho, and P. C. Carvalho, "Computing geodesics on triangular meshes," *Comp. Graph.*, 2005.

[17] E. T. Hall, *The Hidden Dimension*. Anchor Books. ISBN 0-385-08476-5, 1966.

[18] R. Arnaud and M. C. Barnes, *Collada: Sailing the Gulf of 3d Digital Content Creation*. AK Peters Ltd, 2006.

[19] P. Shirley and M. Ashikhmin, *Fundamentals of Computer Graphics, Second Edition*, ser. Ak Peters Series. Peters, 2005.

[20] "CGAL, Computational Geometry Algorithms Library," http://www.cgal.org.

[21] M. Baeuml and R. Stiefelhagen, "Evaluation of Local Features for Person Re-Identification in Image Sequences," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2011.