GIS-supported People Tracking Re-Acquisition in a Multi-Camera Environment

Anastasios Dimou[1], Vasileios Lovatsis[1], Andreas Papadakis[2], Stelios Pantelopoulos[2] and Petros Daras[1]

[1]*CERTH-ITI, 6th kilometer Harilaou-Thermi, Thessaloniki, Greece*

[2]*SingularLogic, Athens, Greece*

{*dimou,lovatsis,daras*}*@iti.gr,* {*apapadakis*}*@ep.singularlogic.eu,* {*spantelopoulos*}*@singularlogic.eu*

Keywords: GIS, Re-Identification, Multi-camera.

Abstract: Modern surveillance systems consist of multiple, geographically dispersed cameras, increasing the technical and scalability challenges for person re-identification. In this context, the use of geographical information to boost the effectiveness of a state-of-the-art re-identification algorithm has been implemented and evaluated, by leveraging the prediction of an event evolution. It is argued that the estimation of possible target trajectories can limit the footage search space and allow focused application of the re-identification algorithm. This is reflected in performance, effectiveness and scalability. The parametrization of the interesting footage reduction mechanism allows using different profiles and a flexible trade-off between performance and robustness. Our work is verified and evaluated in a well known benchmark dataset for re-identification and a real-world dataset created in the framework of the EU-project ADVISE.

1 INTRODUCTION

Law enforcement agencies and private entities have increasingly relied on Close Circuit Television (CCTV) surveillance to enhance security in public spaces and their premises, respectively. Cities have extensive CCTV surveillance in private and public space, with some countries deploying open street CCTV for the purposes of crime prevention in their major cities. The awareness of the geographical distribution of the surveillance cameras, arises the need and opportunity to manage and exploit this type of information through a Geographic Information Systems (GIS) framework.

Surveillance cameras can have different characteristics and installation parameters, resulting in significant changes in the appearance of the people/objects tracked. Therefore, a re-identification (ReID) methodology is required to associate the target along multiple cameras and to observe the evolution of its trajectory. Re-identification is challenging since the appearance of a person can change significantly depending on the viewpoint, the illumination conditions, the camera type, or even random events like occlusions. Other parameters making re-identification more difficult include low quality of the videos, and that due to either fashion trends or dressing codes people tend to wear similar clothes. Due to these challenges, re-identification methods have low matching rates in real world scenarios, where a high number of possible matches is increasing the complexity of the problem.

The rationale of employing GIS information is the reduction of the examined footage to a small set of video camera- and time-wise excerpts, which, with high probability, include the potential re-appearances of the target. Such a reduction has to be based on the distribution of the cameras and the physical constraints of the target transition from one point of interest to another; in our case between the fields of view (FoVs) of cameras. Such constraints are translated into temporal information to limit the search space for object re-identification. While in the case of a limited number of cameras, this rationale may be trivial, in realistic deployments of a medium number of cameras, a robust, dynamic, methodology with a certain degree of automation, is needed.

2 RELATED WORK

In literature, various approaches have been presented to address the re-identification problem. A

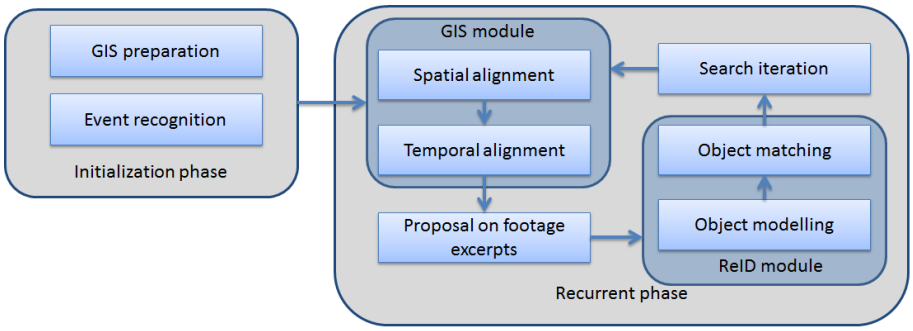


Figure 1: Framework pipeline.

wide variety of information cues have been proposed to describe the appearance of a person, including color, texture and their saliency information. It has been shown that color is the most important cue for re-identification. Various color spaces have been tested for invariance among different cameras (Kviatkovsky et al., 2013). The spatial layout of colors on the clothes has been also employed to enhance the process. Texture is also used to support re-identification using local descriptors, either on sparse or dense sampling points (Zhao et al., 2013). An evaluation of local features for ReID is found in (Buml and Stiefelhagen, 2011). Saliency maps have been also proposed as information cues in the re-identification process (Xu and Zheng, 2013).

Despite the wide range of techniques employed, re-identification performance is far from ideal. Preprocessing steps have been proposed to reduce the weight of background segments from the model (Hu et al., 2013; Farenzena et al., 2010). Color calibration has been also employed to enhance inter-camera re-identification. Transfer functions have been proposed (Avraham et al., 2012) to model the color variation between two non-overlapping cameras. However, these models have to be calculated for each camera pair and they are sensitive to illumination changes within a single camera. Furthermore, body part segmentation is proposed to divide the appearance model into semantically meaningful parts (Bak et al., 2010).

So far, only the appearance of each person has been utilized for matching. In this paper, we argue that it is possible to improve the performance of people/object re-identification in surveillance scenarios by employing information from the location and the viewpoint of the cameras. In literature, the relative positioning of the cameras has been employed to enhance the ReID procedure (Martinel and Micheloni, 2012) in a bidirectional approach but no absolute geographical information has been used.

GIS has been typically used for managing, analysing and decision making, combining both spatial and non-spatial data. Upon underlying maps, layers are created containing arbitrary information. Security has been a field, where video surveillance has been augmented by GIS functionality (Milosavljevic et al., 2010). While GIS usage is focused in camera/incident visualization and statistics, we investigate innovative usage of the geo-information, dynamically prefiguring routes in order to limit the footage employed in event re-acquisition scenarios. Predicting event evolution, in a micro-scale, leveraging upon geo-information is pursued, to our knowledge, for the first time. In terms of implementation, web-based GIS are gaining momentum; such system typically follow multi-tier approaches and consist of three major layers: (a) the presentation layer, (b) the application layer accommodating the geo-spatial middleware and (c) the persistency layer storing the information.

In the rest of the paper, the proposed ReID methodology is presented in Section 3, while in Section 4 experiments are performed to show the importance of trimming the employed search space and a use case is presented. Discussion on the results and conclusions are provided in Section 5.

3 METHODOLOGY

A video surveillance infrastructure consisting of multiple cameras, with known locations and extrinsic parameters and a security-related incident (e.g. bag theft) are considered. In order to identify the actors of the event, it is needed to reacquire the people in neighbouring cameras after or prior to the event time. Besides the challenge of re-identifying a person or object in a new camera due to changes in its appearance, there is also the computational burden of performing the re-identification procedure in a footage of extended duration, which also hinders the efficiency of automatic re-identification, as this is shown in the experimental results. This section describes the methodology followed in two phases: the initializa-

tion, where the required framework and the query incident are defined, and the iterative one, where the actual re-acquisition takes place. A high-level view of the framework pipeline is depicted in Figure 1.

The initialization phase is related to the introduction of the geo-information and event-related information, while the recurrent phase aims at identifying interesting footage excerpts, where the re-identification algorithm is applied. An iteration mechanism facilitates the identification of the route based on the interim re-identification results. The two phases are described subsequently.

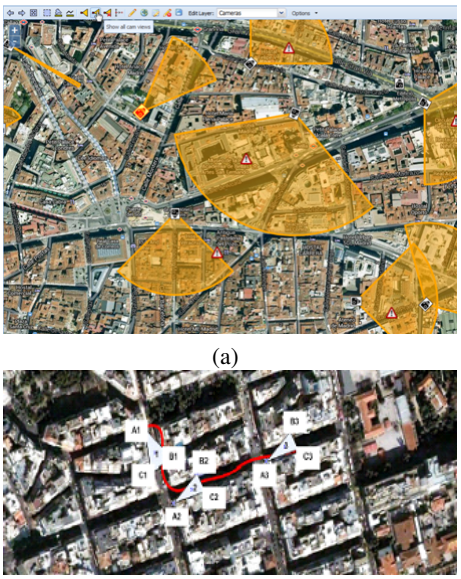
3.1 Initialization phase

The surveillance infrastructure consists of multiple cameras across different geographical locations. In the initialization phase, the framework collects all the necessary information for its operation, including camera setup details, possible routing options and metadata concerning the triggering incident.

3.1.1 Camera Setup

The geographical information of the cameras' position is available as we consider static cameras. The spatial and non-spatial information (metadata), handled by the system, pertains to the cameras included in the surveillance infrastructure. The camera geolocation is the exact location where the camera is installed, while the Field of View (FoV) is the part of the observable world where target detection can take place. Fields of view of different cameras can be overlapping, without that being necessary, as the geographical topology is arbitrary and typically depends on the selections of the surveillance infrastructure owner. Both the location and the fields of view of the cameras are regarded as static (time invariant). Camera metadata also include the elevation from the ground and the direction of the camera. The FoV depends on the maximum viewing distance and the angle that viewing is possible.

In our system, developed in the framework of EU-project ADVISE, camera registration is handled in a manual or programmatic (based on co-ordinates) way and their attributes can be edited using a GUI. The underlying GIS functionality is based upon OpenStreetMap, which is the openly licensed map of the world, the GeoServer, which is the open source server for sharing geospatial data, the PostgreSQL with the PostGIS extensions for persistency and Openlayers for the client side presentation functionality. Using the system, the user can perform typical GIS-related tasks such as calculating areas and distances among cameras. In Figure 2, example FoV of the available cameras are depicted while the user can fully control the visualization of the cameras, activating / deactivating them in an individual or batch manner. These features can facilitate the work of the investigator, while he is working in a typical, non-automatic way.



(a)

(b)

Figure 2: Visualization of the camera fields of view.

3.1.2 Route Calculation

The combination of the aforementioned features and the awareness of the topology surveillance assets can be employed to facilitate object re-identification in an automatic manner. While the absolute distance between each pair of cameras, in a straight line, provides a first indication for the effective distance between the cameras, the awareness of the underlying routing (streets and buildings) network can provide much more useful information. In this view we create all routes among the surveillance infrastructure cameras and calculate absolute distances between the origination and destination points. Each route is tagged on its access possibilities (pedestrian, bicycle / motorcycle, car). For each route, the information is provided in table 1.

3.1.3 Query Event Identification

Having defined the framework to track the evolution of an event (i.e. the route or trajectory of the person or object being followed), the context, namely a security-related event that took place in the geographical area, has to be defined. Metadata containing the actors of the event and their trajectories, including

Route ID	The ID of the route.
Origination point	The starting point of the predicted route.
Destination point	The destination point of the predicted route.
Distance	The distance of the route in meters
Accessibility	Accessibility options for each route: (a) on foot, (b) with bicycle or motorcycle, (c) by car and their combinations.
Direction	The direction of the route from the origination to the destination cameras.
Direction variability	The number of changes in the direction in the course of the route. This number is related to the variability of the direction.

Table 1: Information extracted for each possible route.

speed and direction, are required. The event detection is undertaken by a tracker capable of detecting and tracking objects of interest (pedestrians and vehicles) and analyse their motion patterns to estimate their speed and motion direction. While this information is input to the system described, its extraction is out of the scope of this work.

3.2 Recurrent Phase

Given that the geographical coordinates of the initial point (i.e. the point where the target has been identified for the last time), the direction of the motion and the motion profile of the target (as an estimation of the speed) are available, a set of hypotheses for the trajectory of the target is created. Subsequently, the spatial and temporal consistency of those hypotheses is recurrently tested to reduce the search space. The re-identification process is applied only on the footage verified for spatio-temporal consistency, significantly reducing the examined footage excerpts. It also involves a feedback mechanism allowing route (trajectory) identification based on the interim re-identification results in a causal and non-causal fashion.

3.2.1 Spatial Alignment

The spatial alignment targets at the reduction of the cameras that provide potentially interesting footage. We create a set of perimetric (bounding) boxes having their centers on the current point of interest. The perimetric boxes are rectangular and depending on the camera network topology, they can include a number of neighbouring cameras. Another approach, regarding the shape of the bounding box, is to consider a parallelogram (unequal side lengths) emphasizing on the length of side that coincides with the initial direction of the target (acknowledging the possibility that the object changes direction).

The length of the sides of successive boxes can follow specific relationships, such as length and perimeter doubling and surface quadrupling. Assuming a uniform distribution of cameras, the number of included cameras follow, in ratio, the surface of the box. An increase by one of the ratio (1st, 2nd etc.) of the perimetric box quadruples the number of the cameras. In order to define the length side of each perimetric box, we calculate the length of the maximum perimetric box (which includes the full set of the deployed cameras) and then the size is divided by 2, N times, where N is the number of the perimetric boxes. For each camera included in the perimetric box the direction and the route from other cameras are already calculated in the initialization phase. In case the origination point of interest does not coincide with the point of a camera, the routes towards the considered cameras are calculated.

3.2.2 Temporal Alignment

Depending on the movement profile of the target and the route distance between two cameras, included in the perimetric box, the transition time is calculated. The movement profile of the target includes moving on foot (running), on a bicycle or motorcycle and a car. The transition time is indicative and it depends on the conditions such as the traffic and the (potentially wilful) variability of the speed of the target during his departure. This has to be reflected on the footage excerpts propositions for each camera. The time frame of the excerpts are centred around the Estimated Time of Arrival (ETA) towards the Field of View of the destination camera, with a certain margin of the time needed for the transition. This margin can be optimistic, deviation of 10%, typical deviation of 25% and conservative 50%. This way for each of the surrounding cameras, excerpts of the video footage are proposed as the more probable for containing the suspect.

3.2.3 Object Modelling

In the previous step, a set of video segments has been identified employing the GIS framework, where the

target may reappear, considering its motion characteristics and profile. Subsequently, matching candidates are found in those segments and their appearance is modelled.

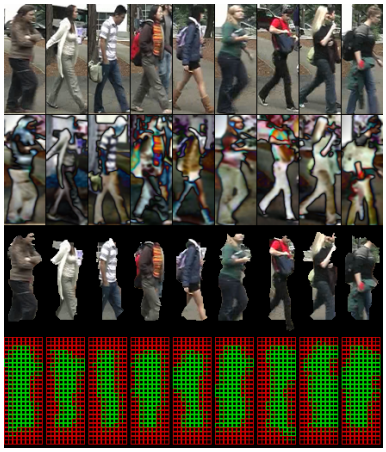


Figure 3: Analysis of the appearance modelling: the first row shows the raw frameshots, the second row shows the saliency maps, the third row shows the segmentation results and the last row shows the grid where green/red rectangles depict the foreground/background patches respectively.

For each candidate, a non-occluded frameshot is taken at the middle of its trajectory and it is defined by a tight bounding box where the person of interest is enclosed (Figure 3, 3*rd* row). It is divided into a dense grid of overlapping patches. Patch size and grid step were experimentally evaluated and the results (Zhao et al., 2013) confirmed that a size of 10×10 with a step of 5 pixels (producing 50% overlap between adjacent patches) is the optimal solution for far-field surveillance camera footage of human figures. In order to keep only the significant patches, i.e. patches that belong to the person of interest and not the background, an automatic foreground/background segmentation, described in (Lovatsis et al., 2013), is applied on the frameshot and patches with more than 50% of foreground elements are labelled as significant, while the others as irrelevant. As seen in Figure 3 (4*th* row) foreground patches are depicted with green color while the background ones with red.

For each foreground patch, a set of descriptors is extracted: 3 histograms of 32 bins on different colour spaces to robustly capture colour information (RGB, HSV, YUV), colorSIFT and, finally, colorSIFT on the saliency map. Saliency map is calculated in real-time without supervision and successfully reveals salient parts of an image (Montabone and Soto, 2010). As seen in Figure 3 (2*nd* row), saliency maps reveal information about the edges of the object, as well as the complexity of the clothing's texture. The length of all histograms together is $3 \times 3 \times 32 = 288$, and the two colorSIFT descriptors length is $2 \times 3 \times 128 = 768$. All descriptors are $L2$ normalized. The final descriptor $\mathbf{z}$ has a length of $288 + 768 = 1056$.

3.2.4 Object matching

The next step in the framework is matching between the target and the candidates, using the models created in the previous step. The list of spatio-temporally constrained candidates is extracted from the footage excerpts provided by the GIS framework.

The similarity function is based on comparing all the patches of the target with the patches of each candidate. To combine the efficiency of local descriptors with proper patch alignment, the search area of each patch to be matched is limited to its neighbourhood with a horizontal constraint rule. Each frameshot I from camera C_1 is divided into a grid of MxN patches, where each patch m,n is represented by a descriptor $\mathbf{z}^I_{m,n}$. The subset of patches belonging to the same row m is represented as:

$$R^I(m) = \{\mathbf{z}^I_{m,n} | n = 1,2,..,N\}. \quad (1)$$

The search space S for $R^I(m)$ in frameshot J from camera C_2 with the same size of MxN patches is the respective subset of patches $R^J(m)$. The horizontal rule can be relaxed in order to provide flexibility for pose variations and different camera viewpoints by widening the area vertically by a small factor r :

$$S(R^I(m), J) = \{R^J(m-r), ..., R^J(m+r)\} \quad (2)$$

The relaxation factor should be large enough to provide flexibility but not too large, to avoid erroneous matches between different body parts. In our experiments, a factor of $r = 1$ was chosen. Also, patches that enclose more than 50% of background data are discarded from S in order to avoid matching objects to background structures.

When comparing two frameshots, all foreground labelled patches from a frameshot I are compared against all adjacent foreground labelled patches from frameshot J. As a result, each patch m,n returns the nearest neighbour distance d from S. In order to amplify small distances and discard very large ones, a Gaussian function converts distances to similarities:

$$s(d) = e^{\frac{-d^2}{2\sigma^2}} \quad (3)$$

where $= 0.2$ is the bandwidth chosen for our experiments. The accumulation of maximized similarities has proven to be more efficient than aggregating minimized distances (Ma, 2012) and was validated in our experiments as well. The final similarity score between frameshots I and J is the mean similarity of all matched patches.

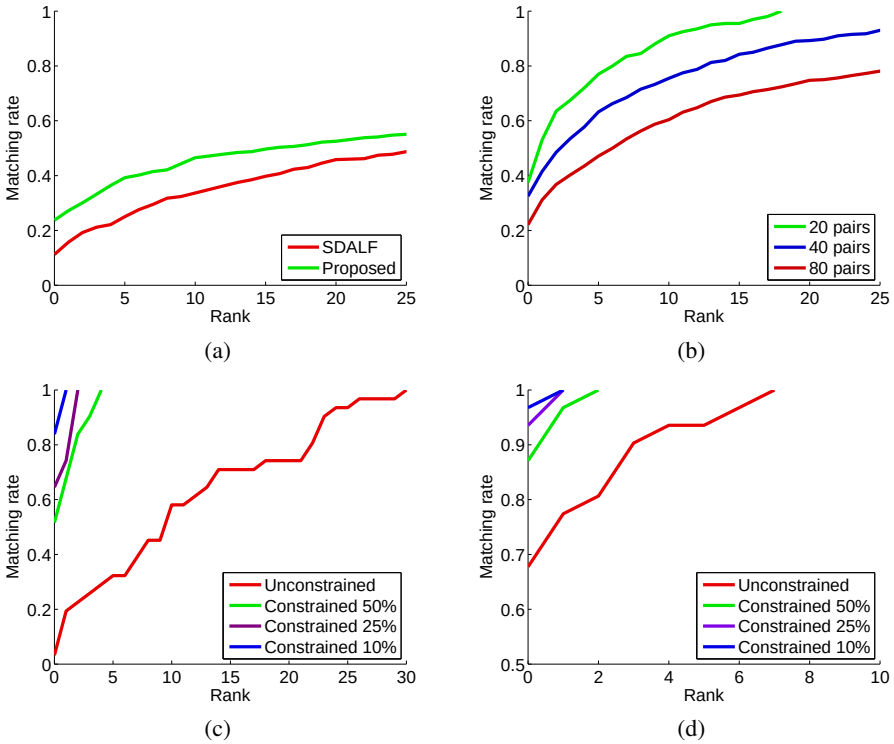


Figure 4: Comparison of the ReID CMC results on the VIPeR dataset for (a) 316 pairs between SDALF (Farenzena et al., 2010) and the selected methodology and (b) different number of pairs with the selected methodology. Comparison of the ReID CMC results on the ADM dataset with and without spatio-temporal constrains, using 3 different temporal prediction margins between (c) cameras 4 and 6 and (d) cameras 5 with 6.

3.2.5 Search iteration

As discussed the exploitation of the geographical information, in terms of spatial and temporal alignment takes place in an iterative manner, based upon the results of the re-identification. In principle we consider two types of results: (a) the person/object has not been re-identified and (b) it has been, indeed, re-identified. In the former, two types of parametric changes define each iteration: (a1) the increase of the dimensions of the bounding box with square increase of the number of the included cameras (in case of uniform distribution) and proposed footage excerpts; and (a2) the increase of the time margins (from 10% to 50% or even larger). The upper limit is to consider the full set of cameras and footage.

In the latter case (the person / object has been identified), we consider two (not conflicting) subcases: (b1) the person has been re-identified within the footage of a certain camera, without being identified in the footage of intermediate (in the sense of the route followed) cameras. This paves the way for re-identification in the intermediate cameras, gradually increasing the time margins. (b) the person has been re-identified in the footage of the full chain of cameras across the identified route and this set of cameras constitutes only a subset of the full set; i.e. the route can extended using currently unexploited footage. The iterative steps are repeated, setting as the point of interest the last point where the person has been identified and selecting spatial and temporal boxes.

4 EXPERIMENTAL RESULTS

In order to evaluate the proposed system, a series of tests is applied. First, the selected appearance-based ReID method is evaluated, following the literature protocol for standard ReID evaluation (Farenzena et al., 2010). Then, it is shown that the number of candidates plays a great role in the overall efficiency of the appearance-based method and that GIS can efficiently reduce the search space during matching.

Due to the lack of publicly available datasets, containing events happening in multiple cameras and geo-tagging, the validation of our experimental validation of our approach is performed in two stages. In the first stage, the impact in ReID performance and

efficiency of a reduced number of candidates is evaluated in the publicly available VIPeR dataset (Gray et al., 2007), which is intended for viewpoint invariant person recognition evaluation purposes.

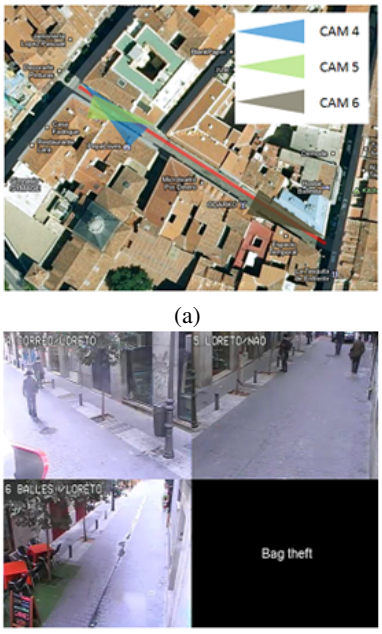
In the second stage, the proposed methodology is tested in a real-world use case, using a dataset created in the framework of the EU-project ADVISE, which contains multiple cameras and geo-tagging. This dataset has been used to evaluate the impact of the GIS module to the ReID results. Test results are measured and showed in standard Cumulated Matching Characteristic curves (CMC). The CMC curves represent the recognition rate in the n top ranked matches.

4.1 VIPeR dataset.

The VIPeR dataset[1] is comprised of 632 image pairs between two cameras under various pose and lighting conditions and it is considered as a very challenging one for ReID evaluation. Each pair represents a unique individual captured once in every camera. All frameshots are normalized to a size of 128×48 pixels. In literature, half of the pairs are used for training or as a reference set and the other half for testing. We follow the same protocol in order to be compared with other state of the art approaches like (Farenzena et al., 2010) and apply the proposed appearance based method using half of the pairs. As seen in Figure 4(a), our method achieves state of the art results that surpasses (Farenzena et al., 2010). The method of (Farenzena et al., 2010) was ported in C++ and the preprocessing step was replaced by the proposed automatic segmentation method resulting in a small drop in the overall efficiency compared to the reported results.

It is clear from the results, that these methods are not efficient under these circumstances and could only be useful to assist a user during a manual search, which is not our case. However, for semi-crowded surveillance footage where a few dozens of individuals are recorded during a reasonable amount of time, the applied test that includes hundreds of individuals is not valid. Using smaller subsets of the original pairs list, we repeat the evaluation test for smaller numbers of pairs this time. Each experiment was repeated 10 times on random VIPeR subsets to ensure the robustness of the results. As seen in Figure 4(b), the overall performance is significantly increased for smaller number of pairs.

[1]The VIPeR dataset can be downloaded at: http://vision.soe.ucsc.edu/node/178



(a)

(b)

Figure 5: The ADM dataset camera (a) topology and (b) frames from "Bag theft".

4.2 ADM dataset.

As mentioned, a dataset created in the framework of the EU-project ADVISE is also employed. The camera network is composed of 3 static cameras across a street with passing passengers and vehicles. Two of the cameras have overlapping fields of view and the third camera has non-overlapping field of view with the other two. The topology and sample frames of the cameras can be seen in Figure 5. For privacy reasons,the dataset has been recorded only with actors and their faces have been blurred in the images provided for this work. A total of 31 different actors were assigned a list of actions during recordings. Tracking results were obtained using (Lovatsis et al., 2013). Targets are represented by a non-occluded frameshot taken at the middle of their lifetime, accompanied with tracking metadata for direction, speed and time.

We apply the proposed framework to re-identify persons between the non-overlapping cameras. Initially, all targets are tested between two cameras without spatio-temporal constraints, like in the VIPeR test case. The combinations result to a total of $31x31 = 961$ comparisons. Then, the position of the targets (i.e. the camera ID) and their tracking metadata are fed to the GIS module and a prediction (evaluation of trajectory) is returned (place and time). Consequently, spatio-temporal constraints are imposed for each target and the candidate matches are drastically reduced.

As a result, the number of comparisons is drastically decreased, namely 124 for 50%, 84 for 25% and 53 for 10% margin.

Figure 6: Matching results for two persons. The first column is the query followed by ranked results based on appearance similarity. Due to spatio-temporal constraints, candidates were excluded from the valid results (darkened frameshots).

As seen in Figures 4 (c) and (d), the significance of spatio-temporal constraints is revealed since the efficiency of the ReID method increases drastically. Also, the smaller the temporal margin is, the higher the matching rate becomes. In both experiments, the first rank for the smallest temporal margin achieves around 90% recognition rate, making the ReID method applicable for unsupervised solutions.

The GIS module can be seen as a filter that excludes matching candidates in case of spatio-temporal inconsistencies with the query. Filtered results can be seen in Figure 6.

5 CONCLUSIONS

In this paper, the support of a GIS framework to enhance reacquisition of people/object tracking in real world surveillance scenarios was examined. Given that the trajectory of a target in a single camera and the absolute location and coverage of all the cameras are known, GIS can facilitate the inter-camera tracking of the target by creating a set of hypotheses concerning the evolution of its trajectory, taking into account the motion capabilities of the target and the implications of the terrain.

The rationale has been to reduce the footage that should be searched for the re-identification of the target. This reduction has a positive effect both on the computational burden of the re-identification functionality and its performance. By checking only the excerpts defined by the spatio-temporally consistent hypotheses the search space is greatly reduced. The speed of the system is increased, while the lower number of possible candidates for matching increases re-identification performance. This was experimentally confirmed both in a popular dataset for re-identification benchmarking and a use case scenario performed in the framework of the FP7 EU project ADVISE. While the primary objective is to reduce the examined footage, there is always the danger to miss the spatio-temporal window that the target appears. In order to safeguard the hypotheses, inserting an increased margin in the proposed excerpts, at the expense of performance should be considered; the margins can be parametrized based upon the profile of the target.

ACKNOWLEDGEMENTS

This work was supported by the European Community's funded project ADVISE under Grant Agreement no. 285024 (www.advise-project.eu).

The authors would also like to thank the Madrid Municipal Police (ADM, Ayuntamiento de Madrid) for their kind collaboration in the creation of the dataset.

REFERENCES

Avraham, T., Gurvich, I., Lindenbaum, M., and Markovitch, S. (2012). Learning implicit transfer for person re-identification. In *Proceedings of the 12th International Conference on Computer Vision - Volume Part I*, ECCV'12, pages 381–390, Berlin, Heidelberg. Springer-Verlag.

Bak, S., Corvée, E., Brémond, F., and Thonnat, M. (2010). Person re-identification using spatial covariance regions of human body parts. In *AVSS*, pages 435–440.

Buml, M. and Stiefelhagen, R. (2011). Evaluation of Local Features for Person Re-Identification in Image Sequences. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, page 6.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, San Francisco, CA, USA. IEEE Computer Society.

Gray, D., Brennan, S., and Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *Performance Evaluation of Tracking and Surveillance (PETS)*.

Hu, Y., Liao, S., Lei, Z., Yi, D., and Li, S. Z. (2013). Exploring structural information and fusing multiple features for person re-identification. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW '13, pages 794–799, Washington, DC, USA. IEEE Computer Society.

Kviatkovsky, I., Adam, A., and Rivlin, E. (2013). Color invariants for person reidentification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7):1622–1634.

Lovatsis, V., Dimou, A., and Daras, P. (2013). Introducing context awareness in multi-target tracking using re-identification methodologies. In *The 5th International Conference on Imaging for Crime Detection and Prevention (ICDP-13)*, London, UK.

Ma, K. (2012). Vector array based multi-view face detection with compound exemplars. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 3186–3193, Washington, DC, USA. IEEE Computer Society.

Martinel, N. and Micheloni, C. (2012). Re-identify people in wide area camera network. In *CVPR Workshops*, pages 31–36. IEEE.

Milosavljevic, A., Dimitrijevic, A., and Rancic, D. (2010). Gis-augmented video surveillance. *International Journal of Geographical Information Science*, 24(9):1415–1433.

Montabone, S. and Soto, A. (2010). Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing*, 28(3):391 – 402.

Xu, D. and Zheng, H. (2013). Person re-identification by multi-resolution saliency-weighted color histograms and local structural sparse coding. In *ICIG*, pages 477–482.

Zhao, R., Ouyang, W., and Wang, X. (2013). Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, USA.