# Urban 3D Reconstruction and Place Recognition

A brief review of the current eﬀorts on Urban 3D reconstruction and visual place recognition is given in this section. The algorithms are grouped by the type of data they use and short examples of the recent work in the area is mentioned.

# Urban 3D reconstruction

One of the problems studied in computer vision is that of accurate reconstructions of 3D objects from series of 2D images. As early as the 1990's algorithms and methods were developed to convert sequences of pictures into 3D points. Of these, one of the most widely used is Structure-from-Motion (SfM). This algorithm takes a (possibly unordered) sequence of images of an object or scene, calculates features in each image and matches them. From the resulting matches the algorithm estimates the location in the real world of points in the image. Generally, the 3D reconstructions obtained from SfM are sparse, i.e. there are very few points in 3D space when compared to what a dense reconstruction would yield. Further refinements of SfM and other Multi-View Stereo techniques like Semi Global Matching have led to dense reconstructions, resulting in better 3D models. Due to its success, researchers have turned to modeling entire cities using these methods.

However, there are additional challenges when trying to model cities. Some these include: images taken far apart, occlusions, objects not being fully photographed and, needless to say, the volume of data that needs to be processed. Many diﬀerent approaches can be used to tackle these challenges. We will describe a few of these in the next subsections. They are grouped by the type of data they use.

**Street side city modeling**

Acquisition of images from street level is arguably the most common approach. Usually, this is done by outfitting a vehicle with cameras and driving around the place to be reconstructed. The main advantage that such methods have is the level of detail of the reconstructed buildings. Acquiring the imagery from

street level allows for mapping the textures from the images resulting in a more realistic reconstruction of the original object. The downside to these methods −and to a lesser extent also for the other two methods− is that objects can be occluded or not fully imaged resulting in incomplete or erroneous reconstructions. A common example of the results these methods produce is Google's Street View. A more recent example is the work of Menzel et al. [5] where publicly available data sources are used as the input to the reconstruction pipeline. However, this requires the availability of city images and their maps.

Another option regarding data acquisition is the usage of planes that y over the areas of interest and acquire pictures of the region. These methods are described in the following subsection.

**Aerial city modeling**

An alternative approach to generating 3D reconstructions of cities is to acquire aerial images or point clouds and use these as the input to the algorithms. These images allow to model especially the rooftops of the structures and do a bad job on generating building footprints, even when instead of nadir aerial imagery oblique imagery is being employed. A limitation of nadir aerial imagery is the level of detail of facades that can be obtained. Here oblique aerial imagery offers a solution, but only when it is sufficient geometrically correct. Most research in these methods deals with either the correct generation of the building's footprint or the correct identification and modeling of the rooftop type and structure. An example is the work of Huang and Mayer [3] where point clouds obtained from the air are used to break-down a building model into primitives after identifying ridge lines in the rooftops. While the method performs well, the models are a combination of geometrical primitives and have no additional details.

Methods based exclusively on one type of input data are generally outperformed by methods using two or more imaging modalities. We will refer to these methods as hybrid and they are described in the following subsection.

**Hybrid city modeling**

The limitations of each of the previous approaches become evident when building must be as accurately modeled as possible. Streetside imaging cannot capture the rooftops of tall buildings and aerial images don't have the details of the building facades. Most approaches now combine one or more imaging modalities with other sources of data such as LIDAR or pre-existing maps to further refine the reconstructions. Szomoru et al. [2] use point clouds obtained from the air and at street level to improve the meshes of reconstructed buildings allowing for more accurate representations of their facades. Other methods are concerned with accurate geo-location of buildings such as the work of Yuan and Cheridayat [8] which attempts to accurately tag buildings given a vehicle's GPS location, the city map and images taken from a vehicle-mounted camera. While hybrid methods solve some of the problems present when using only one type of data, they also require additional techniques for managing the fusion of the data.

# Place recognition

Visual place recognition, or place recognition for short, describes the task of identifying the location of a picture based on the comparison of the scene depicted in the query image against a large database of geo-tagged images. Recognizing that two images were taken in each other’s vicinity is challenging since images may be taken with significant viewpoint and position diﬀerences. Furthermore, weather conditions can alter the scene increasing the complexity of the task.

Two approaches are commonly seen in literature when trying to solve this problem. One focuses on describing the image through landmarks or local objects/points of interest, while the other generates a single descriptor for the whole image.

**Local description**

Methods belonging to this category attempt to break down the image into interesting objects, points or landmarks. These are later used for matching between the landmarks of the query image and those of the database. There exist many descriptors, from the traditional (such as SURF and ORB) to more recent ones such as Edge Boxes. Furthermore, the spatial relationships among the descriptors can also be leveraged to help identify a scene such as in the work of Stumm et al. [7], these methods may fail due to recurring patterns in the images, occlusions or changes in the feature vectors that represent the patches.

An alternative approach is to consider the whole image and match entire images to each other. We denote those as global descriptions.

**Global description**

This category groups methods that use all the information in the image to find a suitable match. Examples of these are NetVLAD [1] and PoseNet [4] which use CNNs to identify, respectively, the location and the pose of a query image. These methods are accurate and fast at inference time. However, they require abundant data in order to train the models. Other approaches are based on image retrieval techniques and also leverage recent advances in CNNs. A concrete example is the work of Iscen et al. [6] where Google street view panoramas are generated in either image space (by image stitching) or in feature space (by constructing memory vectors from a collection of images). This approach has near perfect top-5 accuracy in the authors' tests. Nevertheless, further evaluation of the method is required. Specifically the ability to generate correct matches from images acquired from diﬀerent cameras and the search time required in large databases.

# Conclusion

From this brief review of techniques both for 3D reconstruction and place recognition it becomes clear that there are still many challenges to be tackled before either of these problems can be considered solved. Regarding 3D urban reconstruction we observe a trend towards hybrid inputs and data fusion, thus covering some of the weaknesses inherent to each imaging mode. In the case of place recognition the influence of CNNs cannot be ignored. However, it is yet to be seen whether local or global description oﬀers the best performance.

# References

1. Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern* *Recognition*, pages 5297{5307, 2016.
2. Andras Bodis-Szomoru, Hayko Riemenschneider, and Luc Van Gool. Eﬃcient volumetric fusion of airborne and street-side data for urban reconstruc-tion. In *Pattern Recognition (ICPR), 2016 23rd International Conference* *on*, pages 3204{3209. IEEE, 2016.
3. Hai Huang and Helmut Mayer. Towards automatic large-scale 3d building reconstruction: Primitive decomposition and assembly. In *International* *Conference on Geographic Information Science*, pages 205{221. Springer,2017.
4. Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings* *of the IEEE international conference on computer vision*, pages 2938{2946,2015.
5. J Robert Menzel, Sven Middelberg, Philip Trettner, Bastian Jonas, and Leif Kobbelt. City reconstruction and visualization from public data sources. In *Proceedings of the Eurographics Workshop on Urban Data Modelling and* *Visualisation*, pages 79{85. Eurographics Association, 2016.
6. Pilailuck Panphattarasap and Andrew Calway. Visual place recognition using landmark distribution descriptors. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, *Computer Vision { ACCV 2016*, pages 487{502, Cham, 2017. Springer International Publishing.
7. Elena Stumm, Christopher Mei, Simon Lacroix, and Margarita Chli. Lo-cation graphs for visual place recognition. In *Robotics and Automation* *(ICRA), 2015 IEEE International Conference on*, pages 5475{5480. IEEE,2015.
8. Jiangye Yuan and Anil M Cheriyadat. Combining maps and street level images for building height and facade estimation. In *Proceedings of the 2nd* *ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, page 8.ACM, 2016.

# Person Re-identification over Multi-Camera Network

Currently, the literature study phase of the Person Re-identification Project has been completed and the first steps towards implementing an algorithm has been set. During the literature study phase, many previous works have been reviewed. As this project will base further research on these works, this section will give an overview of the most important works encountered that aim to solve person re-ID. Many more works have been viewed, but cannot all be discussed while providing an in-depth overview.

# 2.1. Common aspects in the approaches

Although research started on person re-ID many years ago, Person re-ID is still a challenging task. As such, to get the research started in the early days, this challenging task was simplified by using a predefined type of dataset. To aid comparability among published work, practically all person re-ID papers try to solve the problem with the resulting same general approach using the same or a similar set of publicly available datasets. Only just recently, publications start to report respectable scores. However, all works are still based on the initial simplified kind of datasets.

Yet, all these public datasets aim to be as representative to practical circumstances as possible, e.g. capture persons in a wide variety of environments, with realistic illumination- and pose variations, with realistic capturing angles, etc. However, due to certain assumptions that are inherently present in all encountered public datasets, the truth is that person re-ID is yet far from being practically employable. These assumptions being, that it is common to extract the bounding boxes of the captured persons and to save and process only these image cutouts. Before processing can start, the person cutouts are created for all imagery from multiple cameras to obtain the imagery part of the dataset. Once the images are annotated, i.e. determined what images belong to the same person, the predefined type of dataset is complete. The aforementioned general approach, in answer to this type of dataset, is then to subdivide the dataset in two sets of images, where one set forms a database of known people for the algorithm. The other set acts as a queue of query images on which the algorithm must decide who that person most likely is by comparing it with the database.

To be able to compare all published algorithms, the overall performance of the algorithm must be determined with a consistent method along algorithms. As such, all algorithms use the same datasets (typically a selection of datasets) and for each provide a ranked list of candidates, which is needed to measure this overall performance. More specifically, by counting the relative number of correct matches. That is, to determine a rank N score, where a query is counted as a correct match if the correct person is found among the first N candidates (i.e. the N most likely matches). We figured that the rank 1 score is the most important, as higher rank N scores were found to be highly correlated to the rank 1 score.

It may be clear that, if such a system is to be implemented in practice, there are lots of caveats to this general approach, like (1) no real-time execution constraints, (2) a fixed number of people in the database, (3) unknown influence of continuously adding more people to the database, (4) unused metadata contained in the original frames from which the people cutouts originate, etc.

# 2.2. SoTA overview

As is quite common in the field of computer vision nowadays, it is also the case for person re-identification (re-ID) that since Convolutional Neural Networks (CNNs) gained popularity, a rather large increase of re-ID performance could be observed. Right before CNNs appeared, it was common that researchers selected the applied features themselves and then build some specific algorithm around that. Therewith, the algorithms of that time did not reach higher rank 1 scores than ±45% (in [3]) on the hardest public datasets. In person re-ID, CNNs were first put on this task around the year 2015. To emphasize how challenging re-ID is, this ±45% score was achieved just before the first CNN appeared, while in earlier years it was even common that algorithms achieved around 20% performance. Luckily, once increasingly better CNN architectures were developed, the scores started to get higher quickly, nowadays around 95% accuracy (in [12]) is achieved on similar datasets.

Some of the more promising algorithms of the pre-CNN era are [1][2][3], these will be discussed in more detail in this paragraph to give an idea of what this era involved. In [1], every person is divided into lower- and upper body regions by detecting the horizontal axes of asymmetry. For every such region, also one vertical axis of symmetry is determined, which results into a total of 4 regions. For every such region, three manually selected features are taken, resulting in one feature per person cutout. The three individual selected features are (1) Weighted Color Histograms, (2) Maximally Stable Color Regions, and (3) Recurrent High-Structured Patches. In [2], every person cutout is resized several times and the Gabor filter is applied on all these scales. These results are then grouped in pairs of neighboring scales, on which magnitude images are produced. Finally, the difference of the covariance descriptors between every two consecutive bands are used as the image feature. In [3], an implementation based on Support Vector Machines (SVMs) is created, where the SVM filter learning process uses the in patches divided cutout images. In this process, matches are made within horizontal image strips only, which aims on consulting corresponding body parts.

The oldest proposed algorithm with an applied CNN that appeared in our literature research is [4]. It was the first paper which reported accuracies of over 50% rank 1 scores on some of the then existing public datasets. It proposes a CNN architecture, specifically build for person re-ID, where the network should determine whether inserted pairs of images shows the same person or not. However, the gain in accuracy, as compared to its previous work, was still limited.

On the other hand, the CNN included algorithm [5] that was published one year later did show a particularly convincing increase in performance. One of the issues of applying CNNs on person re-ID is that the person cutouts typically contains too few resolution, as such the ImageNet networks could not be applied directly. Another general issue is that the datasets are quite small, due to the time involved in annotating the images. This work [5] showed a successful method of combining multiple public datasets to increase overall performance. The first step of this algorithm is to train the CNN on all public datasets together, after which deterministic Domain Guided Dropout is applied, where a dataset is referred to as a Domain. In the last step, the CNN will be finetuned by applying stochastic Domain Guided Dropout. It achieved at that time state-of-the-art performance on most of the datasets it used, with a gain of approx. 20% with respect to the previous CNN work.

In the meantime, occasionally, non-CNN works were still published. One that stands out is [6], it proposes a rather effective way of combining a so-called sequence/track of captured images for re-ID. Thereby, it addresses one other issue of person re-ID in literature. That is, as a person is walking while being captured, it results in multiple images of the subject, i.e. the mentioned sequence. Yet, the bulk of public datasets contains just a few cutouts of a person in such a sequence. As this work obtained quite substantial performance at the time, it may indicate that utilizing these sequences may be promising. However, it could not report results on many datasets, because of its requirement of video sequences.

Next, previous work [7] is also worth mentioning, as it uses the GoogleNet CNN architecture, pretrained on the ImageNet dataset, and next finetuned on a few Person re-ID datasets using a two-stepped finetuning strategy. Furthermore, it also applies unsupervised learning to aid in the fine-tuning process, which may be a promising method for real-world applications, as typically only unlabeled data is abundantly available.

The next paper that stands out is [8], as it stresses the importance of the triplet loss, but also the impact it may have on performance. It allows for end-to-end learning between the input image and the desired embedding space. One especially promising aspect of this paper is that it proposes a variant for data mining using the triplet loss. Where the usual hard mining drops most of the easier samples to prevent overfitting, this variant utilizes all training samples the dataset has to offer, while still putting more emphasis on the hard examples, thereby effectively preventing overfitting. As such, it is achieving competitive and improved results. Additionally, it also pointed out that pretrained models, nowadays, more often obtain great scores for person re-ID, while ever fewer top-performing approaches use networks that have been trained from scratch. Finally, it also investigates how different resolutions affect performance and concludes that a model retrained from scratch is better able to handle this than a further refined pretrained model.

The work presented in [9] is also worth discussing. Interesting enough, it actually does not so much focus on person re-ID, but presents a new way of data augmentation for deep learning, which became quite common for deep learning. It is the Random Erasing Data Augmentation type, which was presented with a test case on a person re-ID dataset. In short, Random Erasing is a way of artificially increasing the dataset size by duplicating every image several times with a different region that is randomly masked away, for different region sizes and orientations. It is an attractive way of improving performance by making the network more invariant towards occlusion. The paper has shown that Random Erasing is also beneficial for person re-ID. Some other interesting conclusions are that, once Random Erasing is applied, applying Dropout and Random Noise no longer is complementary. On the other hand, applying Random Erasing in combination with both Random Flipping and Random Cropping is shown to be beneficial.

Also interesting is how researchers try to apply Generative Adversarial Networks (GANs) to improve detection performance since GANs have been introduced in deep learning. An early example of this on person re-ID is [10]. Nowadays GANs have shown to be capable of producing impressive results on datasets of people faces [11]. As GANs become better, they probably also become better suitable for data augmentation applications. How to best achieve this and whether it is practical at all for person re-ID is yet to be discovered, but it is likely promising for future investigation.

The final paper we want to discuss is [12]. Currently, this is the most promising algorithm encountered so far. This paper reports the current state-of-the-art results, for example, 96% rank 1 Score on CUHK03 and 94% on Market-1501, two of the more important person re-ID datasets. Its algorithm applies local feature learning followed by global feature learning, where the latter uses the local features. Part of this process is automatic part alignment, meaning that body parts are aligned prior to determining a matching score for a query image. The method to achieve the alignment does not require extra supervision nor explicit pose estimation and seems one of the promising aspects of this work. Another outstanding element of this work is that it compares its performance with human annotators and concludes that it is overall outperforming them, although humbly noticing that the type of errors the algorithm makes does not often confuse humans.

# References

1. M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. *Person Re-Identification by Symmetry-Driven Accumulation of Local Features.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2010.
2. B. Ma, Y. Su, and F. Jurie. *BiCov: a novel image representation for person re-identification and face verification.* In International Conference on Advanced Video and Signal-Based Surveillance, September 2012.
3. R. Zhao, W. Ouyang, and X. Wang. *Learning Mid-level Filters for Person Re-identification.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.
4. E. Ahmed, M. Jones, and T.K. Marks. *An Improved Deep Learning Architecture for Person Re-Identification.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.
5. T. Xiao, H. Li, W. Ouyang, and X. Wang. *Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
6. S. Yoon, F.M. Khan, and F. Bremond. *Efficient Video Summarization Using Principal Person Appearance for Video-Based Person Re-Identification.* In The British Machine Vision Conference (BMVC), September 2017.
7. M. Geng, Y. Wang, T. Xiang, and Y. Tian. *Deep Transfer Learning for Person Re-identification.* In CoRR, November 2016.
8. A. Hermans, L. Beyer, and B. Leibe. *In Defense of the Triplet Loss for Person Re-Identification.* In CoRR, March 2017.
9. Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. *Random Erasing Data Augmentation.* In CoRR, August 2017.
10. Z. Zheng, L. Zheng, Y. Yang. *Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro.* In CoRR, January 2017.
11. T. Karras, T. Aila, S. Laine, J. Lehtinen. *Progressive Growing of GANs for Improved Quality, Stability, and Variation.* In CoRR, October 2017.
12. X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. *AlignedReID: Surpassing Human-Level Performance in Person Re-Identification.* In CoRR, November 2017.

# Object Detection in Street Traffic

* 1. **Introduction**

There are thousands of surveillance cameras installed along highways which are mainly used for traffic management and law enforcement. Continuous manual inspection is not feasible, as this requires enormous manual effort involving high costs. Automatic visual interpretation enables detection, tracking and classification of all traffic. The discussion here concentrates on the detection of cars as an example of rigid traffic objects. The given discussion largely applies also to motorbikes, bikers etc. One specifically important concept is visual Make and Model Recognition (MMR). Make and model information of vehicles can be used to find vehicles with stolen license plates, when comparing the observed vehicle model information with the registered information associated with the license plate. An additional application is to find specific vehicles after a crime when only a vehicle description is available without license plate number. In such cases, make and model (the brand and vehicle type) of the vehicle needs to be obtained visually by means of computer vision techniques. These challenges are typical for state-of-the-art object detection in traffic.

Recognition of the vehicles in the above applications is now performed by an Automatic Number Plate Recognition (ANPR) system in combination with a lookup in the national vehicle registration database. Although this works for most cases, it is easy to circumvent this database matching by altering the license plates. Moreover, it does not work for vehicles without a license plate, foreign vehicles or for motorcycles (when considering a frontal viewpoint). We discuss here also to solve the mismatch and missing license plates cases with an accurate visual analysis system. To this end, we present an MMR system developed for the National Police, in which vehicles are observed from a camera mounted in an overhead sign structure on the highway, with the purpose to extract accurate make and model information. The extracted information may be combined with existing ANPR information. It should be noted that the license plate is an attached object with color and texture like a person would carry a shopping bag with him. Hence, the car with plate stands for typical computer vison cases in traffic participants end their detection with semantic understanding.

The system implementation has a focus is on observing a single lane. This existing camera is used to feed the training process of our recognition system. The recognition model is trained to recognize vehicles from a large training set of vehicle images and make and model labels. Due to bandwidth restrictions between the camera (online) and our training and testing facilities (offline), we have tooptimize the gathering of training and testing samples. Another challenge is the automated handling of new and rare vehicle models as registered in the vehicle registration database, for which it is hard to collect training and testing images. For these reasons, we propose a semi-automatic system to create a vehicle dataset. The sampling and their annotation in this system are automated, while the updated training still needs manual control. This approach enables the construction of an initial dataset and allows to incrementally collect new vehicle samples over time, so that the best system performance is ensured at all moments.

The MMR system consists of a detection and a classification stage, to localize and recognize vehicles in a full-frontal view. The aim is to find the vehicle make and model information without being dependent on an ANPR system. Our two-stage approach enables detection of vehicles in every video frame and performs classification once a vehicle is found. This paper extends our initial work1 by providing extensive insight in our MMR classification performance and discussing the evaluation of the MMR system in high detail. First, a comparison between different convolutional neural networks for vehicle model classification is reported. Second, we give more insight in the classification performance by finding the most informative region for MMR classification and measure the robustness against occlusions. Third, the false classifications are further investigated to find shortcomings in the system and information handling.

* 1. **Related state-of-the-art work in traffic object detection**

Our vehicle recognition system consists of a detection and a classification stage, to localize and recognize vehicles in a full-frontal view. The first detection stage can be solved with different approaches. The full vehicle extent is detected using frame differencing by Ren and Lan2 or background subtraction by Prokaj and Medioni.3 Siddiqui *et al.*4 and Petrovi´c and Cootes5 extend detections from a license-plate detector. Wijnhoven and De With6 propose Histogram of Oriented Gradient (HOG)7 to obtain contrast-invariant detection. Recent work by Zhou *et al.*8 reports on a Convolutional Neural Network (CNN) to obtain accurate vehicle detection. When the vehicle is detected, the vehicle region of the image is used as input for the classification task of MMR.

Image classification has been also broadly reported. CNNs are state-of-the-art for image classification and originate by work from LeCun9 and gained popularity by Krizhevsky,10 who used a CNN (AlexNet) to achieve top performance in the 1000-class ImageNet Challenge.11 For MMR, Ren and Lan2 propose a modified version of AlexNet to achieve 98.7% using 233 vehicle models in 42,624 images. Yang *et al.*12 published the CompCar dataset which contains different car views, different internal and external parts, and 45,000 frontal images of 281 different models. They show that AlexNet10 obtains comparable performance to the more recent Overfeat13 and GoogLeNet14 CNN models (98.0% vs. 98.3% and 98.4%, respectively). Siddiqui *et al.*4 show that for small-scale classification problems, Bag of SURF features achieve an accuracy of 94.8% on a vehicle dataset containing 29 classes in 6,639 images.

Other work extends full-frontal recognition towards more unconstrained viewpoints. Sochor *et al.*15 use a 3D box model to exploit viewpoint variation, Prokaj and Medioni3 employ structure from motion to align 3D vehicle models with images, and Dehghan *et al.*16 achieve good recognition results but do not reveal details about their classification model. In conclusion, detection methods involving background subtraction or frame differencing are sensitive to illumination changes and shadows. Therefore, we select Histogram of Oriented Gradients to obtain accurate detection.

We have found that detection performance in this constrained viewpoint is sufficient, whereas complex detection using CNNs8 is considered too expensive in terms of computation. Given the previous work, we have adopted the AlexNet10 network as classification model and focus on an extensive evaluation of the large-scale Make and Model Recognition problem. As shown by Yang *et al.*,12 AlexNet achieves state-of-the-art performance, is one of the fastest models at hand and suitable for a real-time implementation.17 Our experiments are performed on our proprietary dataset, which contains 10 times more images and the double amount of vehicle models than the public CompCar dataset,12 but focuses on a single frontal vehicle viewpoint.

We do not evaluate on the CompCar dataset because classification results are presented by Yang *et al.*12 and we specifically aim at a large-scale evaluation.

The vehicle recognition system is shown in Fig. 2 and consists of two main components: detection and classification. The input of the detection component is a video stream from a camera mounted above the highway focusing on a single lane. The detection component localizes vehicles in each video frame. If a vehicle is found, the vehicle sub-region is extracted from the video image. This cropped image is then processed by the classification component recognizing the make and model of the vehicle. During normal operation, all images from the camera are directly down-sampled so that license plates are not readable anymore, while preserving sufficient resolution for classification. During training and validation, the original image resolution is used because the license plate information needs to be processed by an ANPR engine to automatically annotate the vehicle make and model label for our experiments. The detection and classification components are discussed below in more in detail in [1].

# References

1. M. H. Zwemer, G. M. Y. E. Brouwers, R. G. J. Wijnhoven, *et al.*, “Semi-automatic training of a vehicle make and model recognition system,” in *Image Analysis and Processing – ICIAP 2017 : 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings,Part II*, S. Battiato, G. Gallo, R. Schettini, *et al.*, Eds., 321–332, Springer International Publishing, (Cham) (2017).
2. Y. Ren and S. Lan, “Vehicle make and model recognition based on convolutional neural networks,” in *2016 7th IEEE International Conference on Software Engineering and Service* *Science (ICSESS)*, 692–695 (2016).
3. J. Prokaj and G. Medioni, “3-d model based vehicle recognition,” in *Workshop on Applications of Computer Vision (WACV)*, 1–7 (2009).
4. A. J. Siddiqui, A. Mammeri, and A. Boukerche, “Real-time vehicle make and model recognition based on a bag of surf features,” *IEEE Transactions on Intelligent Transportation Systems*, 3205–3219 (2016).
5. V. S. Petrovic and T. F. Cootes, “Analysis of features for rigid structure vehicle type recognition., in *BMVC*, 2, 587–596 (2004).
6. R. G. J. Wijnhoven and P. H. N. de With, “Unsupervised sub-categorization for object detection: Finding cars from a driving vehicle,” in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2077–2083 (2011).
7. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, 886–893 vol. 1 (2005).
8. Y. Zhou, L. Liu, L. Shao, *et al.*, “DAVE: A unified framework for fast vehicle detection and annotation,” *CoRR* abs/1607.04564 (2016).
9. Y. LeCun, B. Boser, J. S. Denker, *et al.*, “Backpropagation applied to handwritten zip code recognition,” *Neural computation* 1(4), 541–551 (1989).
10. 10 A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Adv. Neural Inform. Proc. Systems*, (2012).
11. O. Russakovsky, J. Deng, H. Su, *et al.*, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision* 115(3), 211–252 (2015).
12. L. Yang, P. Luo, C. Change Loy, *et al.*, “A Large-Scale Car Dataset for Fine-Grained Categorization and Verification,” *ArXiv e-prints* (2015).
13. P. Sermanet, D. Eigen, X. Zhang, *et al.*, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” *CoRR* abs/1312.6229 (2013).
14. C. Szegedy,W. Liu, Y. Jia, *et al.*, “Going deeper with convolutions,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015).
15. J. Sochor, A. Herout, and J. Havel, “Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3006–3015 (2016).
16. A. Dehghan, S. Z. Masood, G. Shu, *et al.*, “View independent vehicle make, model and color recognition using convolutional neural network,” *CoRR* abs/1702.01721 (2017).
17. A. Canziani, A. Paszke, and E. Culurciello, “An analysis of deep neural network models for practical applications,” *CoRR* abs/1605.07678 (2016).
18. R. G. J. Wijnhoven and P. H. N. de With, “Fast training of object detection using stochastic gradient descent,” in *Proc. IEEE ICPR*, 424–427, IEEE Computer Society, (Washington, DC, USA) (2010).
19. J. Shi and C. Tomasi, “Good features to track,” in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 593–600 (1994).
20. K. He, X. Zhang, S. Ren, *et al.*, “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385* (2015).
21. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR* abs/1409.1556 (2014).
22. F. N. Iandola, M.W. Moskewicz, K. Ashraf, *et al.*, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size,” *arXiv:1602.07360* (2016).

# Public safety management and Smart Lighting

## 4.1 Market innovation need

Global trend studies of research institutes (e.g. Frost & Sullivan, 2012[[1]](#footnote-1)[[2]](#footnote-2)) show that urbanization and decreasing city budgets have a negative effect on the actual and perceived safety of citizens. This has a negative impact on the economic growth of a city. Detecting and solving incidents in early stages would benefit and increase the perceived and actual safety of a city.

Currently, the dominant technology used by city authorities to monitor and safeguard public safety as well as detect incidents in early stages, is camera surveillance. Based on research with various cities, insights demonstrate that although camera surveillance has been proven to significantly improve the operational efficiency of law enforcement and other emergency services, it has several drawbacks. First, recorded video footages often cannot be used by video surveillance operators and law enforcement. The quality of the images gathered during night is often too low due to lack of sufficient artificial lighting 3. Second, due to the large amount of different camera views and high operational costs to have multiple observers, incidents or unsafe situations are easily missed by video surveillance operators or are identified too late. On top of the two operational drawbacks, camera systems are expensive for cities to install. Interview with video surveillance consultant showed that in the UK average costs to install a new outdoor IP camera is approximately £15,000 per camera. If fixed fiber optic transmission cabling and equipment is required. Research shows that on average these installation costs are £25,000[[3]](#footnote-3).

## 4.2 State of the art

To successfully play on the smart cities market, companies often collaborate with other verticals to offer complete solutions to the customers. In this chapter we describe participants from the verticals that would potentially form an ecosystem for smart public safety solutions.

## CCTV surveillance

Cameras technology shifted from analogue to IP. The shift enabled various functionalities for end users, one of them being communication with other connected devices. The cameras can communicate with smartphones, traffic signals, windows to recycle bins, or other connected devices that generate new sources of data. Smart management and intelligent analysis enables for officers attending a scene to receive relevant video information.

On the other hand, IP-based surveillance presents various challenges for end users. For example, network bandwidth optimization, large bandwidth requirements to sustain video surveillance on the network, inability to use wireless video surveillance solutions, high cost of private networks for security and others.

In 2008, the market of hardware for IP cameras was strongly dominated by four players, which hold 70% market share all together: Axis with 38.9%, Mobotix 14.5%, Sony 9.0% and Bosch 6.8% [[4]](#footnote-4).On the opposite, the market of end-to-end (E2E) solutions for video surveillance was highly fragmented, meaning that Axis Communications holds the highest share of 12.5%, followed by Panasonic with 5.8% 4.

## Smart sensor solutions for public safety

As already mentioned, IP technology allows to enable intelligent analysis on the video or other devices. At this moment, there are algorithm solutions for both audio and camera that classify events like graffiti, gunshots, explosions etc.

NLSS provides solution to send audio alerts to notify personnel of an event identified as e.g. gunshot or break-in on parking areas. ShotSpotter and Information System Technologies Inc. offer solutions to detect gunshots and provide location and time stamp of the gunshots. Iomniscient provides algorithms for smoke and fire detection, vehicle speeding, illegal parking, detection of suspicious behaviours (e.g. man-down/ fighting, loitering, running, crowd gatherings/ formation etc), gun shots, serious car accidents and screaming sounds.

## Smart lighting

Public lighting has always played an important role in cities, but now with LED and connected lighting it can fulfil its role as never before. By digitizing and connecting traditional street lighting, cities can deliver the right light when and where needed, optimize operations, save money, and help people feel safer and more comfortable. But the value of digitized public lighting doesn’t stop there. As part of the IoT environment, the connected street lighting infrastructure can become an enabling platform for smart city applications, hosting sensor networks and wireless communications for smart parking, noise and air quality monitoring, incident detection, and more. Because street lighting is already installed throughout the city, a connected system can serve as an integration point for many other city services. A smart connected lighting system will contribute significantly to safety applications by means of lighting control or ability to continuously collect, aggregate, and analyse data from various sensors like cameras, noise sensors, smart phone apps, etc.

Several lighting companies active in the area of smart lighting solutions and their applicability as a smart city platform are Acuity Brands, Telensa, Current by GE, Cree, Enlighted

## Safe city solutions and Smart city platforms

Safe city solutions typically focus on improving the safety of citizens. Factors that normally drive these projects are strong city economy, high level of safety and security threat, the availability of internet protocol, susceptibility to natural disasters, and improved private and public partnership (European Video Surveillance Cameras Market , 2010)8. As already mentioned, IoT is increasing its role on the public safety market with IP solutions. This is why horizontal IT players, who offer and enable cloud-based solutions, play an important role in offering smart city platforms to enable a range of application, public safety being one of them.

IBM offers solutions around an Intelligent Operations Center (IOC) and intelligent video analytics. IOC is a major component in most of their smart city solutions offerings and a fundamental piece of IBM’s law enforcement suite, i2 Integrated Law Enforcement [[5]](#footnote-5).

Cisco offers a solution called Connected Public Safety. The solution is scalable to serve operations and requirements across the public safety user landscape: they provide network to interconnect agencies and field personnel, host information in a secure cloud, support the operations centres with situational awareness tools and integrate sensor platforms to capture data 5.

Other players are also Microsoft, SAP, Oracle, HP, Accenture, NEC, Hitachi, Huawei, Thales, AGT, ZTE, NICE etc.

## 4.3 Philips Lighting offering

Philips Lighting has shaped the public realm with public lighting for more than a century. Innovation has always been at the heart of our rich history, having pioneered a number of transformations the lighting industry has witnessed. City Touch, a connected lighting platform, since its launch in 2012 has been installed in 1000 projects over 37 countries in the world, which when combined with LED lighting can deliver up to 70% of energy savings. Transitioning from conventional lighting to energy-efficient LED lighting, to leading the connected lighting journey and shaping what it means to cities, we have embarked on the next wave of integration of lighting with the internet-of-things revolution – taking light beyond illumination.

Leveraging on the increased public safety that LED lighting already brings to cities, connected lighting systems offered by Philips Lighting enables emergency services to turn on additional light when needed in a particular area. By providing the right light at the right moment, they can use the lighting to de-escalate incidents. Additionally, the connected lighting infrastructure is a viable IoT platform for cities to host sensors collecting fine-grained and city-wide data and enabling smart city applications like public safety. By integrating cost-efficient microphone sensor solutions to the lighting grid, city-wide acoustic real-time safety monitoring (aRTSM) can be done to automatically detecting various events such as sound peaks, and public safety incidents (gunshots, breaking glass, aggression, etc.) and the emergency services are alerted on the areas where incidents occur. By receiving the information about the sounds levels in the area, the emergency services can respond faster, operate more efficiently, save costs and prevent escalations of incidents. The system supports them to reduce collateral damage and it also supports the municipality to reduce societal damage (e.g. financial damage such as consequential costs for society and emotional damage such as perceived safety and city branding). The aRTSM can be deployed in cities to either supplement existing video surveillance infrastructure or as an acoustic surveillance solution on its own.

# Video Analytics for Public Safety

This overview of the state-of-the-art for public safety with respect to video analytics is split into two parts. The first part focuses on Video Analytic technologies and their use in public safety applications. The second part focuses on two use cases that are targeted within PS-CRIMSON to improve public safety with video analytics, and compares the use of video analytics with other sensor types.

## 5.1 State-of-the-art, video analytics oriented

In this section, the state-of-the-art is discussed with respect to video analytics for the public safety use case. This is split into four parts: The first section will showcase some whitepapers that are published about the importance and growth of video analytics in the public safety domain, and the challenges that video analytics face. The second section will discuss the state-of-the-art for end-to-end solutions, which use video analytics for city surveillance in a complete Smart City environment/CCTV installation. The third section discusses the use of embedded video analytics, which is video analytics that is performed on the edge, without the need of transmitting video to external servers/over the network or cloud. The last section briefly discusses the development of advanced public safety algorithms, to detect abnormal behaviors.

* + 1. Overviews about video analytics

There are several whitepapers and overviews published that discuss the importance of video analytics with respect to the public safety domain. Whitepapers are published by video analytic providers, camera manufacturers, but also by independent consultancy agencies and consortiums.

Accenture has published some insights about video analytics in the public safety/CCTV domain.[[6]](#footnote-6) They mention that the biggest challenge of current CCTV installations is that CCTV operators struggle to remain alert to view all the footage. “Known limitations in human concentration tell us 95 percent of incidents are likely to be missed after 20 minutes in the command and control room. The sheer volume of footage that needs to be analyzed is also an issue when CCTV is used reactively, following an incident. Ninety-eight percent of video footage is not seen by anyone, let alone acted upon.”

They mention that video analytics can improve results by being able to quickly analyze thousands of video in real-time, or by being applied as a part of post-incident analysis. Especially the last five years are mentioned as immense improvements have been made in video analytics. They mention the Singapore safe city test bed as an example where the government harnesses the latest video technologies to test during two major public events. This predicted crowd behavior, coordinated resources and let them respond to incidents more quickly.

The whitepaper of Axis[[7]](#footnote-7) mentions that video analytics have the following benefits with respect to traditional CCTV installations: 1) More efficient use of manpower, as mentioned before; 2) Faster retrieval of stored video, as video analytics can ensure that only relevant video footage is stored, and in addition can automatically filter the video stream with appropriate labels; 3) Reduced network load and storage needs, by only recording relevant video and by the option of placing intelligent video at the edge; and 4) new business opportunities, to combine the video analytics in other domains next to city surveillance.

They further discuss two categories of systems that are used for intelligent video implementations: centralized and distributed. They mention that the most scalable and flexible architecture is based on distributed computing where video is processed as much as possible near the network camera’s without transmitting video. This limits the required bandwidth, reduces the cost and complexity of the computing network and can ensure privacy by only sending the required data instead of the video footage. However, when employing edge computing, the video analytic providers and users should make sure that the software is compatible with the Smart City architecture and network.

Cisco also published a whitepaper[[8]](#footnote-8) that mentions the same advantages of video analytics, but also elaborates on some applications of video analytics such as perimeter security/asset protection, automatic license plate recognition and port security. Furthermore, Cisco describes the critical factors of video analytics that have led to its advancement and increased potential in public safety scenarios. They mention that the camera technology and sensor capabilities have tenfold, that the public has grown accustomed to video surveillance, and that video analytics have greater accuracy than before, thereby reducing false alarms and increase their acceptance rate.

The most interesting development is the formation of a coordination group in the USA to foster federal interagency collaborations in the emerging area of Video and Image Analytics (VIA). VIA is a working group starting in 2016 and is made up of over 20 organizations: Federal, industrial and academical. They hosted a first workshop on Video Analytics in Public Safety (VAPS) in June 2016, organized by the NIST Public Safety Communications Research Program.[[9]](#footnote-9) The workshop started by explaining the different applications of video analytics, such as mentioned above but also scene reconstruction, object left behind detection and activity/event detection. They also mention some R&D growth areas, such as spatial analysis of large areas, at-scale video analytics incorporating embedded/edge analytics, and multi-camera/multi-sensor processing. Developments in computing power, broadband networks and maturing video analytics technology are mentioned as main breakthroughs that play a critical role in future public safety video applications.

Prior to the workshop, five focus group panels were established (each between 10-20 experts) to discuss different needs and issues of different stakeholders. These were: 1) Public safety and transportation safety video use and analysis; 2) Industry video analytics R&D and related technologies and standards; 3) Academic research in areas relevant to public safety video technologies; 4) Human factors, HCI, and visualization research; and 5) Legal, policy and social considerations. The most interesting conclusions of this workshop and the focus group sessions will be further elaborated in the following sections.

5.1.2 End-to-end solutions for city surveillance

This section explains the current state-of-the-art for end-to-end solutions for city surveillance. Three main suppliers are briefly mentioned. The first is IBM Intelligent Video Analytics[[10]](#footnote-10), which provides facial recognition, people search in databases, pattern change detection, redaction, and video to public safety insight. The Cisco Video Analytics[[11]](#footnote-11) provides a security package, that classifies objects, detects camera tampering and loitering events, and can trigger tripwire and left behind objects alerts. Also, they provide people/vehicle counting, occupancy and dwell-time detection. The Hitachi Video Analytics[[12]](#footnote-12) provides counting, pedestrian classification including wheelchairs, walking poles, windingly walking patterns and loitering, objects left behind, and people searching in a database.

All these companies supply end-to-end solutions, where they apply video analytics in CCTV networks and provide dashboards to the end user/government. However, the video analytics require server computing and do not provide edge processing. Also with respect to the offered video analytics, a gap still remains when compared to the needs of the public safety community as discussed in the VAPS workshop. Among these needs are edge computing, video analytics interoperability, video redaction, more advanced analytics on object behavior, and security.

5.1.3 Embedded video analytics

Other suppliers of video analytics software focus on providing video analytics in embedded forms with edge computing. Example providers are Bosch[[13]](#footnote-13) with their integrated VCA software in their cameras; RIVA[[14]](#footnote-14) that delivers IP network cameras with onboard video analytics; and FLIR, which provides the product Brickstream[[15]](#footnote-15) that is supplying 2D and 3D people counting/tracking with edge computing and wireless networking.

However, all these embedded video analytics do not provide the advanced video analytics required for public safety applications as mentioned by the VAPS workshop, such as video search for target type, object path analysis or activity mapping.

* + 1. Advanced public safety algorithms

Several companies are specialized in the development of video analytic software and can provide advanced algorithms. Intelli-Vision[[16]](#footnote-16) is one of those companies and provides intelligent video motion detection, camera tamper detection, intrusion detection, line crossing, object left/removed, loitering, vehicle counting, face detection/recognition/watchlist, license plate detection/recognition/watchlist and video search and summary.

Another provider is PureTech systems[[17]](#footnote-17) that provides scan-to-target, man overboard detection, car counting, camera tampering, loitering, crowd detection, stopped vehicle detection, dropped/thrown objects, automatic PTZ following and wrong direction detection.

However, these specialized companies only offer these advanced algorithms in server-based installations and not edge-based.

Conclusions of the VAPS workshop about the current state-of-the-art of advanced algorithms are that algorithms are either 1) Forensic – what has happened; 2) Predictive – what will happen; and 3) prescriptive – what actions should be taken. Common algorithms are people tracking, vehicle tracking, tripwires, motion detection, face detection, license plate reading, left object detection, slip and fall detection, crowd formation, and loitering.

The VAPS workshop also discussed about academical development of advanced public safety algorithms. They mention that in general there is a pressing need for more public datasets and funding, and that video analytics is in its infancy. The main hard problem that was discussed, was the detection of anomalous events, context-driven privacy and real-time analysis, under de constraints of crowded scenes and energy efficiency. The topic is considered so new that research funding is needed in all areas, such as research programs, technology transition efforts, data collection, annotation, and benchmarks. Also it is mentioned that consortia-based collaborations with industry in addition to federal funded R&D might help foster more open research and synergy.

* + 1. Conclusions

As discussed in the previous sections and also pointed out by the VAPS workshop: the main challenge of video analytics is to provide the complete package, which can be summarized in the following goals:

* Provide edge computing to limit bandwidth requirements and network costs/installation;
* Provide advanced video analytics to support public safety regulators and CCTV operators in their daily tasks. Some of the most frequently mentioned algorithms/challenges are:
  + Loitering
  + Anomaly detection
  + Comply with low-quality cameras and crowded environments
  + Object path analysis
* Sensor interoperability and connectivity with other platforms

As shown, the current state of the art cannot provide on all of these goals. PS-CRIMSON aims to solve these challenges in the public safety space by developing embedded video analytics solutions, which can communicate with other platforms using IoT protocols such as MQTT and at the same time provide highly advanced algorithms to properly support city surveillance.

## 

## 5.2 State-of-the-Art, use case oriented

Video analytics can be used to improve public safety. This can be split into city surveillance and traffic management applications. These use cases are further explained below and video analytics is compared with the state-of-the-art techniques and sensors that are used at the moment.

* + 1. City surveillance

The area of city surveillance has grown continuously over the years, of which visual surveillance is clearly the dominant factor. The application areas are split into people and crowd monitoring, people counting and flow measurement for safety management, parking lot surveillance, and monitoring of critical infrastructure. A growth area that strongly emerges is shop and retail surveillance both for people safety and/or shopping behavior.

1. People and crowd monitoring is certainly one of the most important applications in urban areas. This is particularly important for events in cities like big sports events, trade fairs, city festivals and others. The surveillance is deployed with an existing – usually fixed mounted - infrastructure of surveillance cameras, which are connected to a network (CCTV or fiber, or similar). The information is typically video compressed and gathered in a local control room, operated by city government and/or police or security companies. The analysis is strongly based on the detection of individual persons and the tracking thereof. The control room contains many video displays arranged in a matrix and requires personnel to actively observe the live video images and then signal situations to local police and other services. Experiments in the R&D area also report besides the detection of persons, the behavior of persons by analyzing their profile and motion actions over time, also the behavior analysis of groups of people.
2. People counting and flow measurement is closely connected to the previous area of monitoring. For large festivals or city events, the counting of people in certain areas is crucial for safety reasons and organizational guidance. Conventional surveillance cameras are not always good enough for accurate measurements, resulting in that regularly a temporal and/or removable system is deployed with a specialized camera. The advantage of such a system is that the sensor is optimized for the purpose, such as a high sensitivity for low or fluctuating light conditions. The corresponding video analysis is a highly tuned person detector and classifier that is optimal for specific mounting positions of the camera. The use of stereo cameras and range finding sensors has been reported in literature for such experiments. Also, tracking algorithms for persons are available, based on various principles, such as mean shift, motion analysis, and various filtering techniques like particle filtering. New sensors can be optimized even for measuring heart beat and so on, but this is not broadly used and infrared sensors may be used for night-time surveillance.
3. Monitoring of critical infrastructure (incl. parking lots). This is a field where typically more sensing principles are jointly used, because of the value of the construction. For example, in a harbor, radar is often used for ship guidance, and this may be extended with visual sensors. For vehicles, magnetic coils or inductive loops are used for detection and supplemented with visual sensing. For entrance control, applications of face detection and recognition are deployed and can be enhanced with eye detection or iris scanning, or biometric features like fingerprint processing and recognition. With respect to analysis, the world is diverse in that area. Human face recognition exists, but the size of the database may become critical at larger scale and for fast identification. The detection of vehicles and other objects with different sensors is may be enhanced with different information such as identity tags, and is typically shown in large control rooms or on a table or screen. In urban areas, object detectors and classifiers exist for cars and persons and initial algorithms are there for following an object through the infrastructure from one camera view to the other. Tunnels in a city have typically a similar infrastructure for surveillance.
4. Shop and retail surveillance. This is an emerging area that is initially starting from safety but now increasingly is used for shopping behavior as well, so that the investment for the retail becomes more interesting. The typical mode is visual surveillance and this is regularly extended with RFID tags and magnetic/inductive loops for good flow and detection. The analysis involves high-quality person detectors and pose/ posture descriptors and trackers for recording the walking paths of customers in a shop. It is obvious that the people’s presence and the type of goods may be well combined with specific light types and colors of light to stimulate customer behavior.
   * 1. Traffic management

We distinguish two types of systems for intelligent traffic management as state-of-the-art: Intrusive sensors and non-intrusive sensors. The first ones include devices such as pneumatic road tubes and inductive loop detectors. These systems are installed directly on the pavement surface of a road via saw-cuts, holes, or by anchoring directly to the pavement surface as is the case with pneumatic road. In general, the advantage of using intrusive sensors is related to the maturity: It is well understood and cheap to implement. On the other side, the drawback is related to the costs due to disruption of traffic during installations, and damage to the roads every time such systems are deployed.

In locations where road pavement work and interruption of traffic should be minimized, non-intrusive sensors can be used. These sensors can be mounted above-ground, above the traffic lane they are monitoring, or on the side of a roadway where they can view multiple lanes of traffic at angles perpendicular to or at an oblique angle to the flow direction. The non-intrusive technologies include Global positioning systems (GPS), PIR sensors and video content analysis based camera sensors. Like intrusive sensors, the non-intrusive sensors measure vehicle count, presence, and passage. However, sensors like video content analysis based cameras deliver more broad information such as multiple-lane, multiple-detection zone coverage, and vehicle speed/type classification.

In the following subsection we describe the principle of functioning, the use, the advantages and disadvantages of each intrusive and non-intrusive sensor:

1. Pneumatic road tubes consist of a rubber tube with air that signals a vehicle when pressured. This method is cheap and easy to install. However, measurement accuracy is not very high due to high temperature sensitivity of the air switch and the unknown relation between the number of wheels and the vehicle. Moreover, it cannot classify the traffic, does not work for pedestrian and bicycles and is sensitive for vandalism and wear produced by truck tires.
2. Inductive loop detectors are devices composed by a coil of wire embedded in the road and a detector. Such systems are used to assess vehicle passage, presence, count, and occupancy. The main advantage of such technology is the low costs involved purchasing the sensors, however, the drawbacks include disruption of traffic for installation and repair and failures associated with installations in poor road surfaces. Also, inductive loops are not suitable for pedestrians and bicycles.
3. Passive Infrared (PIR) sensors are made of thermoelectric materials and usually contain lenses or mirrors in order to focus the infrared light for maximum reception in its field of view. PIR sensors are often used in the construction of PIR-based motion detectors. Apparent motion is detected when an infrared source with one temperature, such as a human, passes in front of an infrared source with another temperature, such as a wall. Infrared sensors are used for estimation of object speed, as well as detecting pedestrians in crosswalks. A PIR sensor can provide accurate measurement of object position and speed although it cannot distinct different traffic participant and is not able to classify the type of participant. Consequently, if the traffic situation is at some points a bit dense, the sensors fail.
4. Cameras and video content analysis. A smart camera usually consists of several components like the image sensor device and an image processing unit. In literature, extraction of moving objects from a static camera is typically carried out by calculating a background model of the scene. Usually foreground objects are detected by evaluating the difference between the current image and the background image. However, the maintenance of a background model is difficult, because of issues related to lighting change and moving background (e.g., waving leaves and flowing water and cast shadows). A different technology, used also by ViNotion, is a technology based on object recognition carried out employing knowledge of object shape and texture properties. Such approach is independent of motion, illumination changes and a background model. Other advantages with respect to the other methods are that video content analysis techniques are general and can serve other purposes such as intelligent video surveillance. For instance, a camera network on a street and the video content methods could be used by public authorities such as the police to monitor a certain area. Because a video camera solution is non-intrusive for the road infrastructure and can serve multiple more city services such surveillance and lighting control, the solution is more cost efficient.

Regarding video detection for road traffic application the challenges of a deep integration with traffic lighting will be:

1. The capacity to work under changing lighting condition (as of today road traffic detection relies on the fact that illumination changes over a scene are slow);
2. Installation constraints allowing easy deployment compatible with public lighting deployment constraints. Especially the constraints for road traffic video detection will be: the capacity to work on camera with wide angle focal length (panoramic focal), automatic calibration, video sensors positioning;
3. The possibility to integrate a video sensor within a lamp pole;
4. With enough processing capacities to be able to detect the situations of interest without the need to send back the images to other location to be analyzed;
5. With low power consumption;
6. At low cost to allow large scale diffusion

# 6. Pre/Post Disaster Assessment and 3D Smart Model

## 6.1 Needs in 3D Information

Cities and municipalities have tried to establish the system that can predict disasters levels and estimate the damages and losses afterwards. Usually evacuation manuals and trainings are offered by the city councils in an effort of minimize the effects, and site-visit is a main approach to calculate the losses from disasters. However, in the rapidly growing urban environment, these approaches will not keep up with the numbers of population or buildings that needs to be contacted. A faster and more efficient methodology is need.

3D building models can be one of the solutions as many studies suggest a feasibility of using 3D models in different disaster scenarios either for analysis[[18]](#footnote-18) or visualization[[19]](#footnote-19). However, the current 3D city models worldwide only provide 3D context to a complete building and as a result do not provide the breakdown of the building to the individual units. This causes a difficulty for emergency planning and response in understanding who and what is contained in that building. Esri Canada has determined a more enhanced data model and platform architecture where all buildings with individual unit ownership (condominiums, stratified properties) are reflected in the 3D model.

The enhanced 3D model, so called 3D smart model, also gives advantages in pre/post disaster assessment, especially when it is combined with other analytical sources such as census, building information, and assessment data. In addition, this will allow to perform other analysis such as view shed, aspect and so on as well as value added services beyond land title and real estate. Since the model represent real world buildings, the model is physically and spatially accurate, allowing for integration of other sophisticated GIS models. It is emphasized that there are apparently many analyses that can be only performed in 3D[[20]](#footnote-20).

## 6.2 3D Smart Model

The pre/post disaster assessment web application provides various data that will be interactively used for visualization and analysis in damage, social, and economic impacts. The availability of 3D building data detailed down to the individual unit level is necessary to support the required analysis.

3D smart model can be generated through data creation and data enhancement part. This may vary depending on input source.

## 3D Data Creation

To create building data in a scale of unit level, a heads up digitizing process is used primarily for converting paper into GIS building data. Afterwards, appropriate attributes such as space type, unit and elevation, etc. will be added. Figure 4 and 5 show the differences between a traditional building model vs. Esri Canada’s smart building model in 2D and 3D.

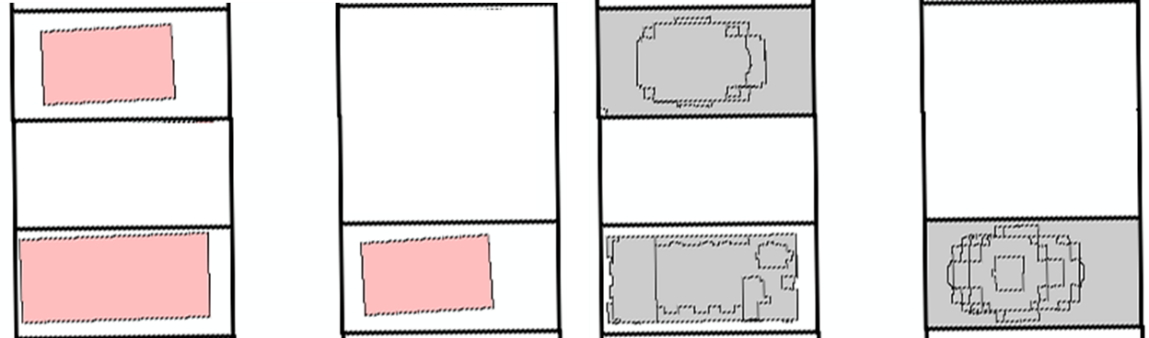


Figure 1 Traditional Vs. Smart-city Building Polygon in 2D

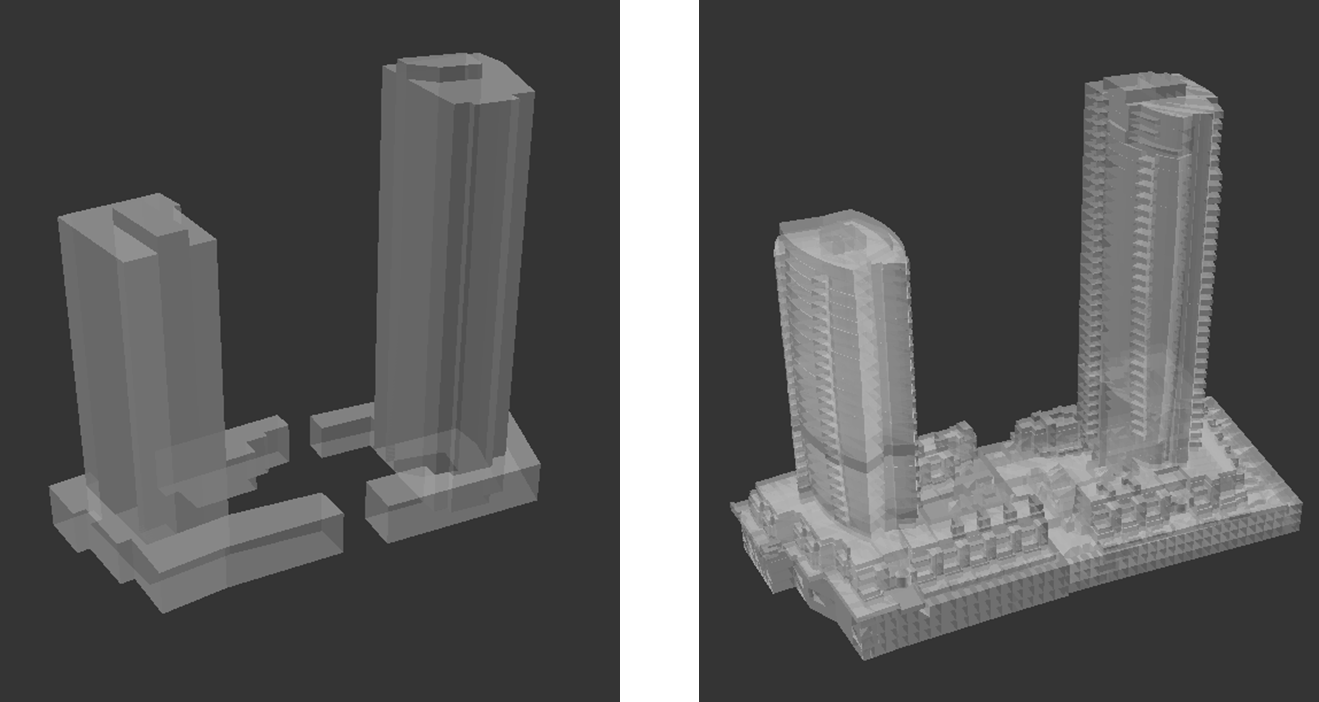


Figure 2 Traditional Vs. Smart-City Building Model in 3D

Once building data is ready in 2D and 3D, both need to be published as a service. While 2D data can be easily created and published, 3D data requires extra processing steps prior to publishing. Particularly for building models, 3D data is created from the existing 2D building data with its vertical (z) information. Once 2D building polygons are extruded based on the z information, a built-in tool in ArcGIS converts this into a 3D GIS object, multipatch. 3D buildings can be simply generated with filled symbology in ArcGIS Pro (Figure 3), or can have realistic textures generated from rule files in CityEngine. With any methods above, 3D data in multipatch will be published as a 3D scene layer (Figure 3).



Figure 3 3D Model Creation Process

## 3D Data Enhancement

Once unit-level building model is created, additional source data can be attached to assist pre/post disaster assessment analysis. To define building control groups that are used to categorize buildings based on their structural characteristics can be manged by adding the building information such as year built, material and the number of floors. This may be applicable in a building level, but also can be applied based on floor or even unit level with the 3D smart model.

For another example, understanding the effect on population is critical in pre/post disaster assessment to reduce injuries and losses. Currently, a block level census information is provided as its minimal size, and this even can be further broken down to unit level by utilizing the number of units in each building in calculation. This type of data enhancement will increase the quality and the accuracy of the assessment by providing more reliable numbers.

## TransLink Data

Not only the data attached to the 3D building model, but also other city objects can be modelled to contribute a complex analysis. Especially in the earthquake scenario, another main concern is transportation structures such as underground railroads or elevated guideways and highways. These massive structures will affect dramatically in case of earthquakes as well as drives the situations even worse when they affect on the surrounding buildings. With the collaboration with the local transit authority, TransLink[[21]](#footnote-21), skytrain[[22]](#footnote-22)data is available to generate 3D models. The data still requires the elevation information to improve 3D analytics along with other layers. Currently, the test model has been generated (Figure 4)

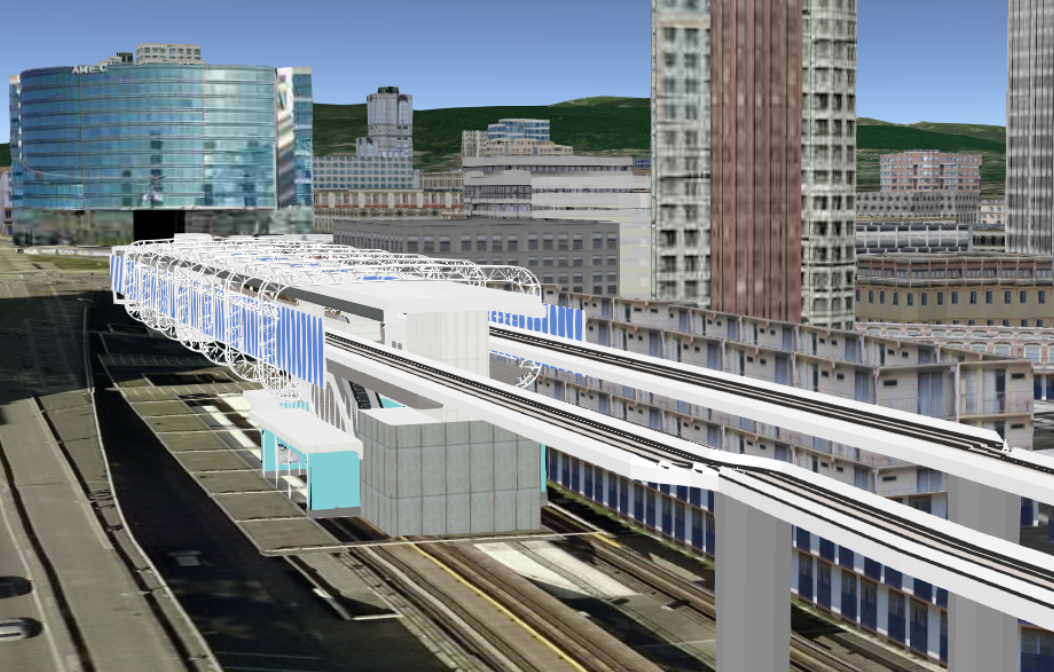


Figure 4 Test SkyTrain Model

## Pre/Post Disaster Assessment and Smart city platforms

Once data is available off-line, this can be published as scene layer to be consumed through a web application. The application contains multiple scene layers published as service to display and analyse the 3D data along with the functionalities defined by the City’s emergency management officers. The application is accessible through regular web browsers with a simple user authentication step which will guarantee rapid propagate use of the application.

## 6.3. Future of 3D with Scene Layer

As mentioned previously, scene layer (I3S) is the source of 3D data in the application. scene layers are cached web layers optimized for displaying a large amount of 2D and 3D features[[23]](#footnote-23), and recently it is approved by the OGC (Open Geospatial Consortium) for use in high performance visualization and spatial analysis. With a collaboration with CycloMedia, this will bring the product to the next step as Scene Layers can support integrated meshes as one of the source: point, point cloud, 3D object and integrated mesh.

CycloMedia’s 3D textured meshes are supported in COLLADA (COLLAborative Design Activity)[[24]](#footnote-24), and this suggests an alternative way to create Scene Layers. With an Esri’s built-in tool called ‘Import 3D Files’, COLLADA data can be converted into Esri’s 3D object model, multipatch. Multipatch is a format that can be published as Scene Layer (I3S), and can be consumed on the web along with other GIS layers (Figure 5).



Figure 5 Test scene wtih Esri Canada's 3D building model and CycloMedia's Mesh data

1. Frost&Sullivan,2014, Safe Cities Technology and Market Trends: Data Analytics and 4G will deliver Operational Efficiencies, Improved Situational Awareness and Better Response [↑](#footnote-ref-1)
2. Frost & Sullivan, 2012, Global Safe Cities market Assessment [↑](#footnote-ref-2)
3. Computers, Environment and Urban Systems,2018, A comparison between the cost effectiveness of CCTV and improved street lighting as a means of crime reduction [↑](#footnote-ref-3)
4. Frost & Sullivan, 2010, Global video surveillance market [↑](#footnote-ref-4)
5. Frost & Sullivan, 2015, Public Safety ICT in the North American Smart City [↑](#footnote-ref-5)
6. <https://www.accenture.com/sg-en/insight-video-analytics-public-safety> [↑](#footnote-ref-6)
7. <https://www.axis.com/en-ae/learning/web-articles/technical-guide-to-network-video/intelligent-video> [↑](#footnote-ref-7)
8. <https://www.cisco.com/c/en/us/products/collateral/physical-security/video-surveillance-manager/white_paper_c89-492776.html> [↑](#footnote-ref-8)
9. <https://www.nist.gov/publications/first-workshop-video-analytics-public-safety> [↑](#footnote-ref-9)
10. [**https://www.ibm.com/nl-en/marketplace/video-analytics-for-security**](https://www.ibm.com/nl-en/marketplace/video-analytics-for-security) [↑](#footnote-ref-10)
11. <https://www.cisco.com/c/en/us/products/physical-security/video-analytics/index.html> [↑](#footnote-ref-11)
12. [**https://www.hitachivantara.com/en-us/products/internet-of-things/video-analytics.html**](https://www.hitachivantara.com/en-us/products/internet-of-things/video-analytics.html) [↑](#footnote-ref-12)
13. [**http://resource.boschsecurity.com/documents/VCA\_Operation\_Manual\_enUS\_23098106251.pdf**](http://resource.boschsecurity.com/documents/VCA_Operation_Manual_enUS_23098106251.pdf) [↑](#footnote-ref-13)
14. [**http://www.rivatech.de/en/**](http://www.rivatech.de/en/) [↑](#footnote-ref-14)
15. [**http://www.brickstream.com/**](http://www.brickstream.com/) [↑](#footnote-ref-15)
16. [**https://www.intelli-vision.com/smart-security/**](https://www.intelli-vision.com/smart-security/) [↑](#footnote-ref-16)
17. [**http://www.puretechsystems.com/video-analytics.html**](http://www.puretechsystems.com/video-analytics.html) [↑](#footnote-ref-17)
18. Biljecki, F. Application of 3D City Models: State of the Art Review, ISPRS Int. J. Geo-Inf. 2015, 4, 2842-2889 [↑](#footnote-ref-18)
19. AMIREBRAHIMI, S. A Planning Decision Support Tool for Assessment and 3D Visualization of Flood Risk to Buildings. https://www.floodplainconference.com/papers2016/Sam%20Amirebrahimi%20Full%20Paper.pdf [↑](#footnote-ref-19)
20. Duggan, N. Why 3D GIS Is the Future. http://www.xyht.com/spatial-itgis/3d-gis-future/ [↑](#footnote-ref-20)
21. The South Coast British Columbia Transportation Authority, operating as TransLink, is the statutory authority[5] responsible for the regional transportation network of Metro Vancouver in British Columbia, Canada, including public transport, major roads and bridges. https://en.wikipedia.org/wiki/TransLink\_(British\_Columbia) [↑](#footnote-ref-21)
22. SkyTrain is the metropolitan rail system of Metro Vancouver, British Columbia, Canada uses fully automated trains on grade-separated tracks running on underground and elevated guideways. https://en.wikipedia.org/wiki/SkyTrain\_(Vancouver) [↑](#footnote-ref-22)
23. ArcGIS Online Help – Scene Layers. https://doc.arcgis.com/en/arcgis-online/reference/scene-layers.htm [↑](#footnote-ref-23)
24. COLLADA: an interchange file format for interactive 3D applications. [↑](#footnote-ref-24)