Optimal Image to Lidar Deep Learning Regression for Height Estimation

Mark Rahmes, Glenn Boudreaux, William Watkins
L3Harris Technologies, Space and Airborne Systems
Melbourne, Florida 32904
Mark.Rahmes@L3Harris.com, Glenn.Boudreaux@L3Harris.com, Bill.Watkins@L3Harris.com

ABSTRACT

We describe a system for estimating pixel heights from a single multispectral RGB image, with or without sensor metadata. System components include an ensemble of convolutional-deconvolutional neural network (CNN) models and an optimization function. The chosen deep learning network model is validated per pixel using high-resolution aerial RGB imagery and lidar datasets. A knowledgebase of historical, time-stamped, multi-modal data for registration and 3D feature classification is provided. Given a large amount of elevation truth data, a model is trained to recognize image features of differing heights using CNN image-to-lidar regression. The models, when applied to an unseen image, estimate a preliminary height per pixel, based on a learned feature set. Multiple models are created and trained end-to-end, and the best model and results are determined. We use linear programming optimization with an ensemble of regression models and semantic segmentation information with a CNN classification model to determine optimized pixel height estimates. Semantic segmentation datasets help classify RGB imagery with feature class labels and refine land use feature classification with CNN classification to improve accuracy. Each land use classified feature can be weighted with a confidence metric that is used to help determine height information. Therefore, we use CNN regression for preliminary height estimation and CNN classification for land use feature classification plus a linear programming reward matrix per pixel to automatically decide optimized height estimation. The rows in the reward matrix contain CNN regression model results from image-to-lidar regression, while columns contain CNN classification model results from RGB imagery. An updated volumetric knowledgebase contains the system output and can be used subsequently for change detection and situational awareness. Both qualitative and quantitative analyses are performed and visualized.

Index Terms: Classification, Cognitive, Decision, Deep Learning, Geospatial Data.

ABOUT THE AUTHORS

Dr. Mark Rahmes has 28 years of experience at L3Harris as an Electrical/Computer Engineer and Senior Research Scientist. He earned his BSEE from The Citadel, MSEE from Duke University and PhD in Operations Research from Florida Tech. He is a retired U.S. Navy Reserve Captain and served 22 years as a Surface Warfare and Engineering Duty Officer. At L3Harris, Dr. Rahmes serves as a Principal Investigator and Chief Engineer on various domain-specific initiatives. He has 56 patents issued and 56 professional publications.

Glenn S. Boudreaux has 27 years of experience in Research and Development for sensor processing, visualization, and artificial intelligence. He earned his BS in Physics from the University of Florida, MSEE from Georgia Tech, and MBA from Florida Tech. He is currently an Advanced Programs Engineer at L3Harris, conducting research in volumetric 3D processing for SAR/EO/IR/LIDAR sensors, multi-mode 3D visualization, and AI-based image processing.

William W. Watkins has 35 years of experience in Research and Development of remote sensing, real-time processing, and autonomous weapons development. He has a BS in Math and Computer Science from Youngstown State University, an MS in Signal Processing from the Florida Institute of Technology, and an MBA from the University of Florida. He is currently employed as an Advanced Programs Engineer and leads many Research and Development efforts at L3Harris, including 3D/4D Volumetric SAR/EO and MultiINT processing.

Optimal Image to Lidar Deep Learning Regression for Height Estimation

Mark Rahmes, Glenn Boudreaux, William Watkins
L3Harris Technologies, Space and Airborne Systems
Melbourne, Florida 32904
Mark.Rahmes@L3Harris.com, Glenn.Boudreaux@L3Harris.com, Bill.Watkins@L3Harris.com

INTRODUCTION

The visualization and simulation communities have shown an interest in classification products for sensor simulation. Visualization and simulation products are created by merging and mosaicking multi-source satellite and aerial imagery of different resolutions on an elevation surface to provide realistic, geo-specific terrain features. These products do, however, require that all image data be orthorectified, seamlessly co-registered, tonally balanced, and feather blended into mosaics from source data of different resolutions [16].

We can determine elevation from a single multispectral image. We improve the estimation of pixel height from various types of images to provide better 2D/3D maps, using images with and without sensor information. Deep learning on geospatial data is performed with a CNN network trained end-to-end. We use image semantic segmentation to classify land-use land-cover (LULC) features. The use of game theoretic decision analysis optimization with an ensemble of models and segmentation information helps determine whether pixel heights are high, medium, or low.

Remote sensing requires that image analysts be able to identify regions in imagery that correspond to an object or material. Automatic extraction of image areas that represent a feature of interest requires two steps: accurate classification of pixels that represent the region, while minimizing misclassified pixels, and vectorization, which extracts a contiguous boundary along each classified region. This boundary, when paired with its geo-location, can be inserted into a feature database independent of the image [3].

The sheer volume of available high-resolution satellite imagery and the increasing rate at which it is acquired present both opportunities and challenges for the simulation and visualization industry. Frequently updating material classification product databases, using high-resolution panchromatic and multispectral imagery, is only feasible if time and labor costs for extracting features, such as pixel labeling, and producing products from the imagery are significantly reduced. Our solution is designed to provide flexible and extensible automated workflows for LULC pixel labeling and material classification. The products of workflows undergo an accelerated review and quality control process for feature extraction accuracy by geospatial analysts [13].

A network can also be trained to predict semantic segmentation maps from depth images [20]. A large body of research in supervised learning deals with analysis of multi-labeled data, where training examples are associated with semantic labels. The concept of learning from multi-label data has attracted significant attention from many researchers, motivated by an increasing number of new applications, such as semantic annotation of images and video [21].

In remote sensing, Digital Terrain Model (DTM) generation is a long-standing problem, involving bare-terrain extraction and surface reconstruction to estimate a DTM from a Digital Surface Model (DSM). Most existing methods have difficulty handling large-scale satellite data of inhomogeneous quality and resolution and often need an expert-driven, manual parameter-tuning process for each geographical type. Feature descriptors based on multiscale morphological analysis can be computed to extract reliable bare-terrain elevations from DSMs [6].

Image-to-height estimation from a single monocular image, using deep learning networks, is a relatively recent research topic. Estimating height in a scene benefits remote sensing tasks, such as feature labeling and change detection, especially when lidar data is not available [15]. We can further advance this technology by adding image semantic segmentation and classification information and game theory optimization from an ensemble of models. Our enhanced solution can then be used as a seed for traditional image parallax height estimation algorithms, thus improving their accuracy. We use this labeled image data to train a CNN U-Net for automatic semantic classification of multispectral imagery. The image features can then be used to help predict elevation height, as shown in Figure 1.

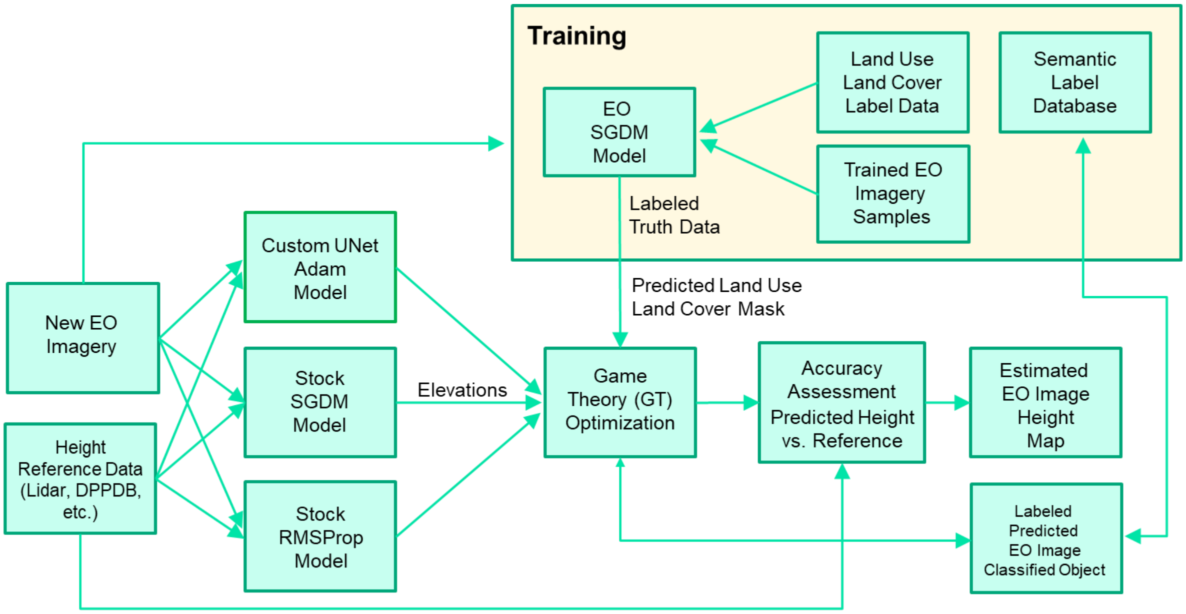

Figure 1. System Overview

DEEP LEARNING

Our Melbourne Florida dataset, shown in Figure 2, contains labeled training, validation, and test sets, with seven object class labels. The lidar data used was downloaded from the International Hurricane Research Center from Florida International University (FIU). We used commercial satellite imagery. Although collection of truth is somewhat subjective, the land cover features were collected via typical specifications. It is important to remember that considerable latitude is allowed when determining what should be collected. Different analysts can and will collect the same area differently, yet both views may be considered acceptable.

Deep convolutional neural networks (CNNs) have recently performed extremely well on different tasks in the domain of computer vision, such as object detection, image classification, image segmentation, and object tracking. The structure of modern deep CNNs has evolved significantly. The renaissance of neural networks has ushered in a new era in which very deep networks have been proposed to carry out various tasks in computer vision. Humans can easily determine approximate height from a single image, based on object recognition and spatial context. [9].

Depth estimation in monocular imagery, which plays a crucial role in understanding 3D scene geometry, is an ill-posed problem. Recent methods have brought about significant improvements by exploring image-level information and hierarchical features from deep CNNs. These methods model depth estimation as a regression problem and train regression networks by minimizing mean squared error, which suffers from slow convergence and unsatisfactory local solutions. Existing depth estimation networks employ repeated spatial pooling operations, resulting in undesirable low-resolution feature maps. To obtain high-resolution depth maps, skip-connections or multilayer deconvolution networks are required, which complicates network training and requires more computations. A multi-scale network structure can be used to avoid unnecessary spatial pooling and capture multi-scale information. [8].

Successful training of deep CNNs often requires many thousands of annotated training samples. Network training strategies rely on the strong use of data augmentation to optimize the efficient use of available annotated samples. Research is currently being conducted to determine whether networks can be trained end-to-end with fewer images and with GPU processing [17].

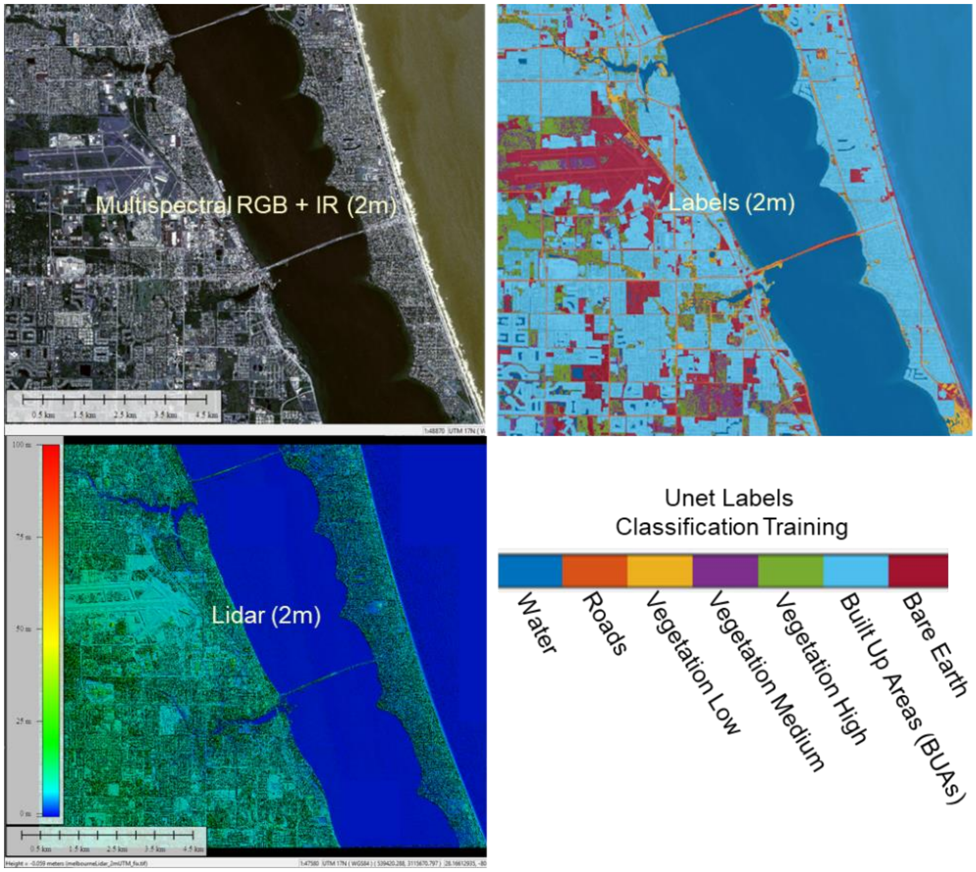

Figure 2. Melbourne Dataset

Learning to predict scene depth from RGB inputs is challenging. Learning for scene depth is provided by monocular videos. Work in unsupervised image-to-depth learning has established strong baselines in this domain. High-quality results can be achieved by using geometric structure in the learning process for modeling, which has been shown to transfer across data domains, e.g., from outdoor to indoor scenes. The approach is of practical relevance, as it allows for transfer across environments by transferring models trained on data collected, for example, for robot navigation in urban scenes to indoor navigation settings [4].

Deep-learning-based approaches are effective for the detection and reconstruction of buildings from single aerial images. An optimized, multi-scale, convolutional-deconvolutional network derives the information needed to reconstruct the 3D shapes of buildings, including height data and linear elements of individual roofs, directly from the RGB image. Networks are composed of two feature-extraction levels to predict the coarse features and then automatically refine them. The predicted features include the normalized digital surface models [1].

Estimating the depth of each pixel in a scene can be done using a single monocular image. Unlike traditional approaches that attempt to map directly from appearance features to depth, semantic segmentation of the scene, using semantic labels, can guide the 3D reconstruction. Knowing the semantic class of a pixel or region allows for easy enforcement of constraints on depth and geometry. In addition, depth can be more readily predicted by measuring the difference in appearance with respect to a given semantic class. The incorporation of semantic features enables better results to be achieved, with simpler models [12].

To automatically extract height information from a multispectral image, we first train a CNN Unet to perform semantic segmentation of a multispectral image with four channels: three color and one near-infrared. This produces pixel-based height maps. The first part of the U in the Unet performs convolutional feature extraction, while the second part of the U performs deconvolutional height estimation [15]. Our network, which we implemented using both Matlab and Python, is shown in Figure 3.

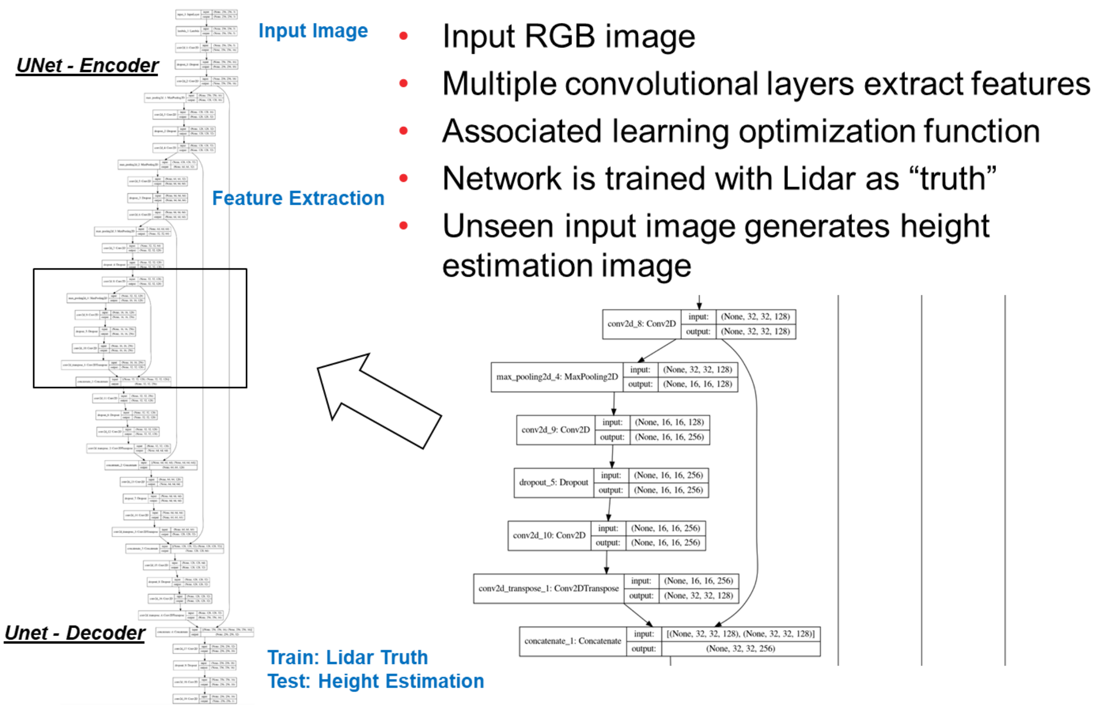

Figure 3. Lidar Deep Learning CNN Model

How well each model works depends on feature properties, quality and quantity of training data, and parameter settings for individual algorithms. Extensive validation of results is needed to properly select the optimal model and model parameters for a given problem. If training data is drawn from a non-linear distribution, it is unlikely that a linear learning method would be a good fit for the data, resulting in a high bias, although this data can be generalized to some extent. If training data is linearly separable, and we use a highly non-linear-based learning algorithm, then it will likely over fit the data, suffer from high variance, and not be able to generalize well with the resulting output. If only minimal training data is available or the data is not adequately representative of the feature space, then accuracy and precision will be negatively affected. We have tested each model on a few different images and geographic areas to understand how well each one might work in practice. Figure 4 shows that the CNN Unet trains to a decreasing Root Mean Squared Error (RMSE) of estimated heights, as compared against lidar truth.

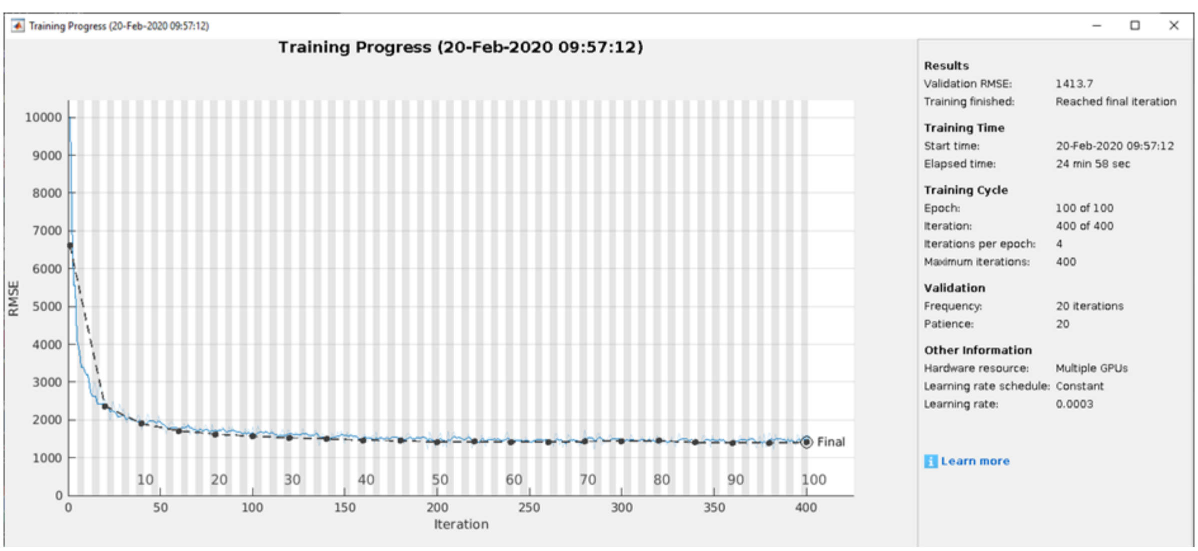

Figure 4. Lidar Feature Training

Optimal decision analysis helps close the gap in terms of the difference between automated feature extraction and feature extraction performed by analysts. To make informed decisions, an expert must reason with multi-dimensional, heterogeneous data and analyze the results. Items in such datasets are typically represented by features. However, as argued in cognitive science, features do not provide an optimal space for human reasoning. In fact, humans tend to organize complex information in terms of prototypes or known cases rather than absolutes. When confronted with unknown data items, humans assess them in terms of similarity to these prototypical elements. Interestingly, an analogous, similarity-to-prototype approach, where prototypes are taken from data, has been successfully applied in machine learning. Combining such a machine learning approach with human prototypical reasoning in a Visual Analytics context requires integration of similarity-based classification with interactive visualizations. To that end, data prototypes should be visually represented such that they trigger direct associations to cases familiar to domain experts. Highly interactive visualizations are used to explore data and classification results. This approach not only supports human reasoning processes but is also suitable for enhancing an understanding of heterogeneous data [14].

A pixel is determined to belong to a classification set when the distance, in feature space, between the pixel's spectral signature and the signature of a representative set of pixels is small. Classification algorithms vary in how the feature vector (and, therefore, feature space) is defined, how the distance metric is defined, how a representative set of pixels or distribution is determined, and in which algorithm to use to identify pixels matches. Nevertheless, they all share the concept of goodness-of-fit, i.e., how well a pixel fits the target spectral distribution, as measured by a per-pixel score. The goal is to accurately identify the boundary of a spatially consistent set of pixels that belong to a region of interest, with the intent being to extract that region as a distinct feature [3].

Semantic segmentation uses a label for each pixel. We can use deep learning to determine a precise measurement of land-use land-cover from high-resolution aerial imagery to differentiate classes with similar visual characteristics. To assign a classification of features over an image, we apply supervised learning to the imagery. Supervised learning creates a classifier model that can infer the classification of a test sample using knowledge acquired from labeled training examples. Figure 5 shows that the CNN network trained with 94% accuracy for our test dataset.

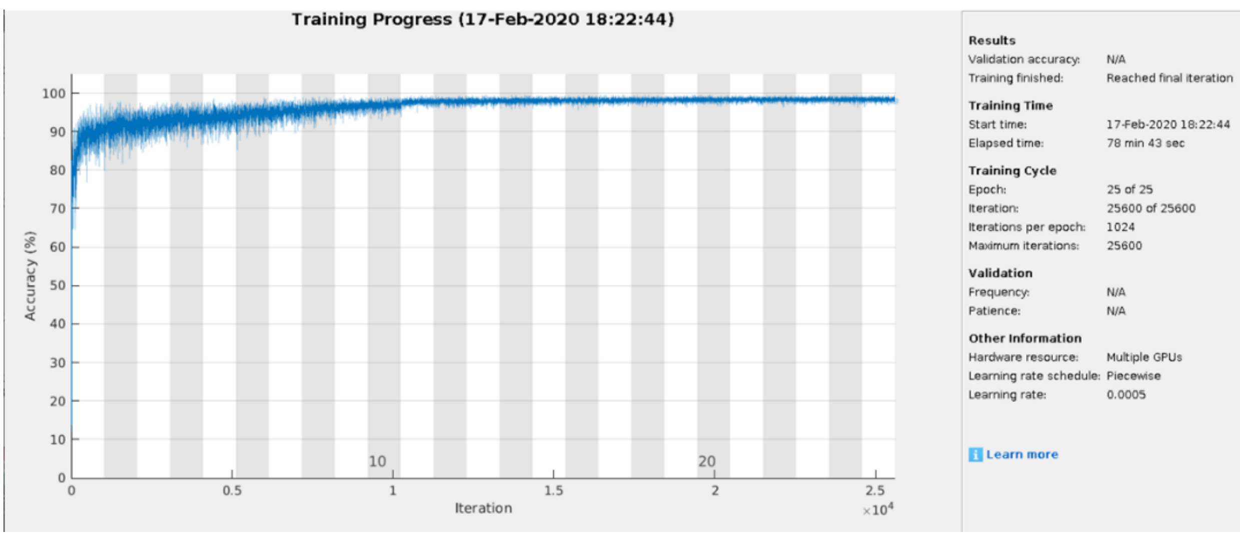

Figure 5. Image Feature Training (94% accuracy)

We used a random patch extraction datastore in Matlab to feed the training data to the network. The datastore extracts multiple corresponding random patches from an image and pixel label datastores. Each minibatch contains 16 patches that are 256x256 pixels in size. We use 25 epochs, with 1000 minibatches per epoch. We use a U-Net structure from Matlab, such that the network can be drawn with a symmetric shape like the letter U. We train the network using stochastic gradient descent method (SGDM) optimization. [11]. Figure 6 shows the results of image feature testing, for which we achieved an accuracy of 92%. We did an 80/20 train/test split of the data. While the test area of Melbourne shows a local area with different geographical and height features, Florida is notoriously flat. The objective of these algorithms is to determine local height variation amongst grass, foliage, small hills and buildings. There is further work to be done to get good accuracy over a wider range of varied topographical characteristics.

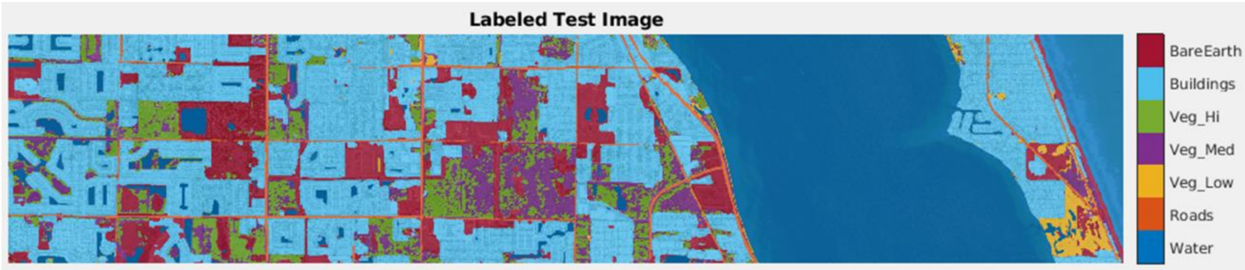

Figure 6. Image Feature Testing (92% accuracy)

OPTIMIZATION

If we can estimate the scene structure, we can better predict the scene heights by knowing the relationships between the features. Estimating height from image features puts a significant burden on the learning algorithm. Using semantic features from the image can unburden the image-to-height learning algorithm [12]. Many image analysis and computer vision problems can be formulated as a scene-labeling problem, in which each site is to be assigned a label from a discrete or continuous label set, with contextual information. An ***n***-person cooperative game yields an efficient deterministic optimization algorithm that exhibits very fast convergence [5]. We use a linear program to optimally guide the height prediction with feature classes from imagery.

We have developed a novel, game-theoretic perspective to solving the problem of supervised classification that takes the best pixel height prediction derived from an ensemble of CNN supervised classifications. This is a game in the sense that pixel data points are “players” that participate in the game to decide their heights by choosing the best network model. The land cover classification labels assist with decision analytics. Within this formulation, we use a weighted reward matrix for consistent labeling of height values with classification factors, resulting in higher accuracy and precision.

We further optimize by performing supervised landmark-based image segmentation that employs game-theoretic concepts [10]. We create a reward matrix with land cover classifications and different model solvers, as shown in Table 1. The reward matrix is constructed from an MxCxN volume, where M is the number of models in the ensemble, C the number of classes, and N the number of surrounding pixels in a neighborhood around the subject pixel height to predict. In our simulation, we used one model for each solver, for a total of three models, i.e., Adam, Stochastic Gradient Descent Method (SGDM), and Root Mean Square Propagation (RMSProp); 7 classes, i.e., water, roads, vegetation low, vegetation medium, vegetation high, built up areas (BUAs), and bare earth; and a 3x3 or 9 neighbors.

Table 1. Game Theory Reward Matrix

Feature Type	**Optimization** **Solver: Adam, SGDM, RMSProp**	**Window size: 3x3**
Water	min	min
Roads	min	min
Veg_low	min	mean
Veg_med	mean	mean
Veg_hi	max	max
BUA	max	mean
Bare_earth	min	mean

An A matrix is then constructed and solved with linear programming, which is useful for solving game theory problems and finding optimal strategies. We use an interior-point algorithm, the primal-dual method, which must be feasible for convergence. We choose the best machine learning model per pixel. The primal standard form, which is used to calculate optimal tasks and characteristics [23], is shown in Equation 1. The **x**'s are the decision variables; **A**'s are the coefficients in the reward matrix; **b**'s are coefficients which satisfy the constraints; and **f** is a linear objective function of constants.

$$\begin{gathered} maximize\ \boldsymbol{f}(\boldsymbol{x})\ subject\ to \\ \boldsymbol{Ax} \leq \boldsymbol{b} \\ \boldsymbol{x} \geq \boldsymbol{0} \end{gathered} \tag{1}$$

SIMULATION

Figure 7 shows our simulation. Once the data and pretrained model are loaded, processing can begin. The heights of each pixel from the multispectral image are calculated with an ensemble of several solvers: Adam, SGDM, and RMSProp. The optimal choice is determined using a game theoretic algorithm with a segmented image land-use land-cover classification. The accuracy for each solver is calculated for both the current image tile and a cumulative value. The Game Theoretic (GT) solution is shown to have a significantly better accuracy compared with any of the ensemble image-to-height network models.

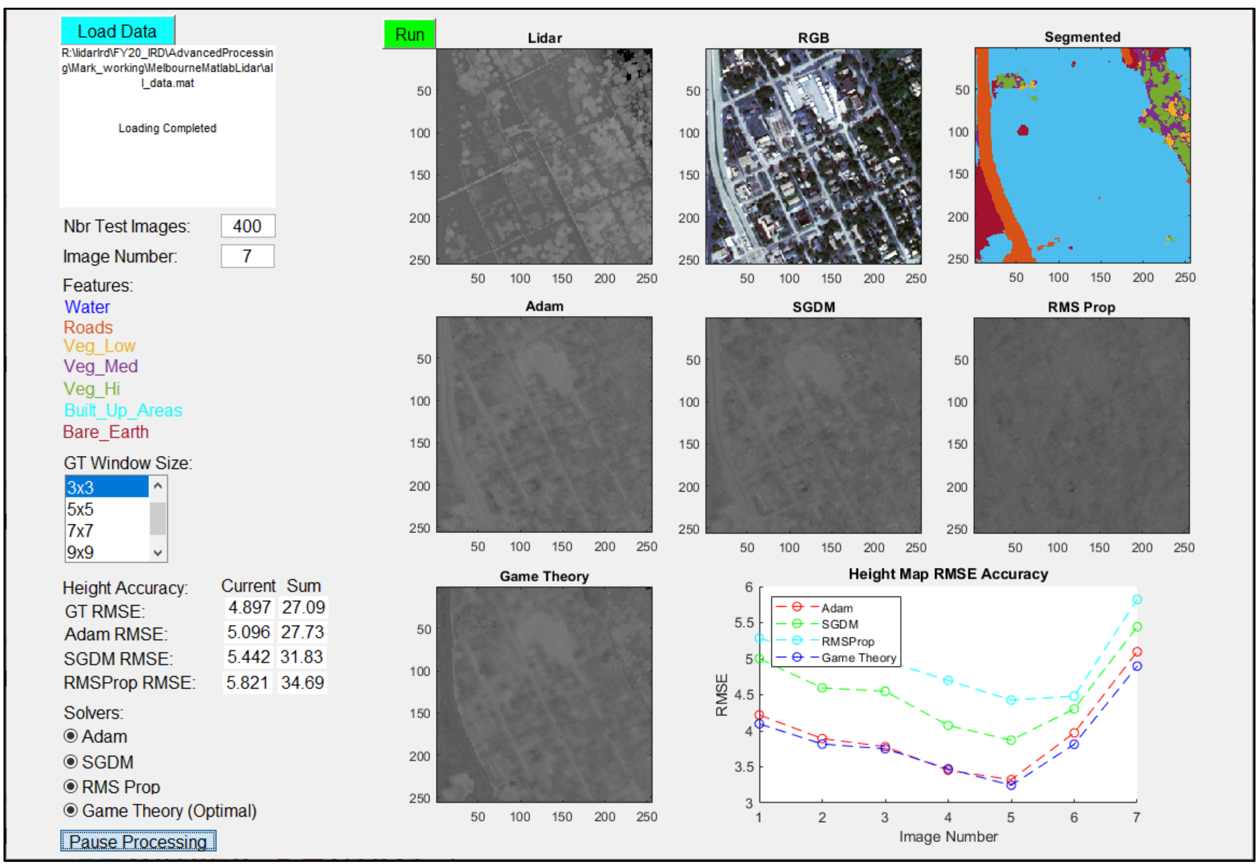

Figure 7. Simulation

There is a need for detailed surface representations so that a feasible platform can be provided for detailed simulation of urban modeling. First, a digital surface model is generated based on aerial image stereo pairs, using a matching method. Methods to generate the 3D city structures have been investigated and possible solutions tested. Features and 3D models extracted from these data can provide benefits in various GIS applications, for which the building is necessary. For example, 3D surface objects extracted from aerial photographs can represent a significant layer of GIS databases for the simulation of natural disasters, telecommunications planning (i.e., positioning of antennas), 3D land-use zoning, and allowed building volumes, usage, and density. They are the main tools that help define the image of a city and bring into focus, for instance, the model of best practice for rehabilitation and conservation [7].

It is a time-consuming process to generate cost coefficients for defining a 3D cost cube, using image-matching operators based on stereo-geographic image data. To save time and computational overhead, we initialize the adjustment of the cost coefficients of the 3D cost cube based on the geographic feature data to generate an adjusted 3D cost cube for a best-cost surface. We can use the game-theoretic height prediction as an initialization seed value to enhance DSM height extraction, using a cost cube algorithm. Processing time is milliseconds for initial height map estimation from aerial imagery, using a trained model. Our use of this initial height map speeds up processing time and improves DSM accuracy. We also use predicted LULC features to determine the search range. Here, we add value by refining the area to search along each sensor ray. This not only allows for faster processing but also for a better starting point for improved height extraction accuracy. One product is shown in Figure 8.

DSM extraction is the most complex and time-consuming part of the process. A great deal of effort goes into making the process efficient and accurate. Of high importance is the fact that high correlations can occur at multiple voxel locations. For example, the corner of one building could correlate very well with a corner on another building. Much effort and ingenuity go into algorithms and logic to sort out these ambiguities. The processing time is dependent upon the computer hardware used.

Figure 8. Extracted Digital Surface Model over Rochester, NY.

High-resolution imagery from today's commercial satellites and airborne systems, with accompanying metadata, can be used to make very accurate, detailed, high-resolution reflective surface DSMs. The surface models accurately depict large features, such as mountain ridges, valleys and drainage patterns, as well as small features, such as roads, trails, buildings, houses and bridges. These models can be used in a wide range of applications, such as oil and gas industry seismic planning, well site planning, pipeline routing, drainage analysis and emergency response planning.

Our reflective surface DSMs are produced from commercially available, high-resolution stereo imagery. The images are taken from overhead, using commercial satellites or airborne sensors. Our DSM extraction algorithm has been designed to handle multiple stereo pairs of the same scene. The major benefit of this feature is that scene portions that are occluded in one stereo pair may be clear in the other. This feature is usually impractical with satellite sensors but is very feasible with airborne cameras. The collection plan for airborne sensors is more flexible and typically includes considerable redundant stereo coverage over the scene.

Monitoring of changes in topographic urban geospatial databases is one of the main requirements of urban planners, urban decision-makers and managers. An automatic change detection (ACD) process uses two types of datasets. In the first type, aerial and satellite images are used as data sources to generate DSMs and extract textural and spectral information. Aerial imagery, because of its geometric stability, provides metric information, while satellite imagery, because of the abundance of spectral information, can be used to generate spectral data. The second dataset type is comprised of topographic urban geospatial databases. These datasets provide reference information that supplements the more recent information and changes provided by the aerial and the satellite imagery. The change detection process includes object identification, object extraction, object recognition, and change detection phases. [18].

Digital spatial data can be vulnerable to strong underlying temporal changes. Typically, to ensure that the data remains current, these changes must be updated by manually checking the data for correctness and superimposing any changes on updated orthophotos. Typical update cycles for large datasets are on the order of several years. Currently, there are two reasons that shorter update cycles are not practical. First, manual inspection of the data is very costly and time-consuming, and, second, aerial photographs for large areas are often not available in the time intervals required. A significant new development is emerging, however, in the area of data availability. New satellite systems provide up-to-date high-resolution orthophotos in short time periods and at high quality [22].

Photogrammetry and remote sensing have proven their efficiency for spatial data collection. Skilled operators routinely perform interactive mapping at digital workstations. Many national GIS databases have been acquired and supported, and considerable production effort is still devoted to them. In the field of image analysis, it has become evident that algorithms for scene interpretation and 3D reconstruction of topographic objects, which rely on a single data source, do not function efficiently. Research in two areas, however, does hold some promise. First, multiple, largely complementary, sensor datasets, such as range data from laser scanners, synthetic aperture radar (SAR), and panchromatic or multi-/hyper-spectral aerial images, have helped achieve robustness and better performance in image analysis. Second, GIS databases, e.g., layers from topographic maps, can be considered virtual sensor data, with geometric information and explicit semantics. In this case, image analysis seeks to supplement missing information, e.g., the extraction of the third dimension for 2D databases. A related goal that many expect will become more important in the future is the revision and update of existing GIS databases [2].

KNOWLEDGEBASE

We improve the accuracy associated with creating a geospatial model, using available data from multiple sources. Change detection (understanding changes) and resulting track extraction (understanding activities) is an important part of many Intelligence Community and commercial GIS-related applications. Given the recent explosion in available imagery data and the increasing number of areas-of-interest throughout the world, there is an increasing trend toward rapid, automated change detection algorithms. To ensure effective use of these imagery databases, care must be taken to verify that the new imagery matches the existing imagery in terms of coverage, field-of-view, spectral content, and, most notably, sensor location and viewpoint. In addition, the need exists to reliably monitor change over time to determine the route of objects (movers), using persistent change detection to derive tracks from multi-int, multi-modal data, if the collection cadences are adequate to determine activity, e.g., multiple frames per hour. This is problematic in that it is often time-consuming, difficult or even impossible to obtain, process and correlate imagery from multi-modal sources to generate persistent change detections and track extractions. The challenges include image-to-image registration; multi-modal image-to-image co-registration; and image-to-ground multi-modal registration. As a result, large amounts of collected multi-modal imagery go underutilized in terms of the potential for change detection and track extractions given lost opportunities for detailed analyses of change over time.

Generation and maintenance of a Virtual Persistent Data Volume enables the creation of 2D, 3D, and 4D change detection products. It also enables the separation of the virtual products' background and foreground, which allows for derivation of virtual track data (activity). Change detection involves the combined processing of elevation model differences (3D), multi-modal imagery content (2D), and voxel-level historical volumetric attribution. An automated method compares a collected image to a reference (source) image extracted from a pre-existing 3D scene (site model, lidar model, high-res DEM, etc.) through a synthetic camera created and placed in the scene such that it matches the collected image sensor's location and parameterization (e.g., field-of-view, hyperspectral vs. monochromatic, etc.). Furthermore, relevant known and stored historical "real-world" phenomenology, such as atmospheric and time-of-day effects, overall ground lighting/reflectivity properties (e.g., soil/vegetation/water), etc., can be simulated in the scene before the reference image is extracted for enhanced change detection performance. An automated method to co-register multi-int data enables the generation of predictive and forensic products that creates a Virtual Persistent Data Volume from any input source.

An important application is the use of single-channel SAR data with Moving Reference Processing (MRP) to focus and geolocate moving targets. Moving targets within a standard SAR image scene are defocused, displaced, or completely missing in the final image. The SAR-MRP method focuses and geolocates moving targets by reprocessing the SAR data to focus on the movers rather than the stationary clutter. SAR change detection is used so that target detection and focusing is performed more robustly [19]. Figure 9 shows the knowledgebase concept overview.

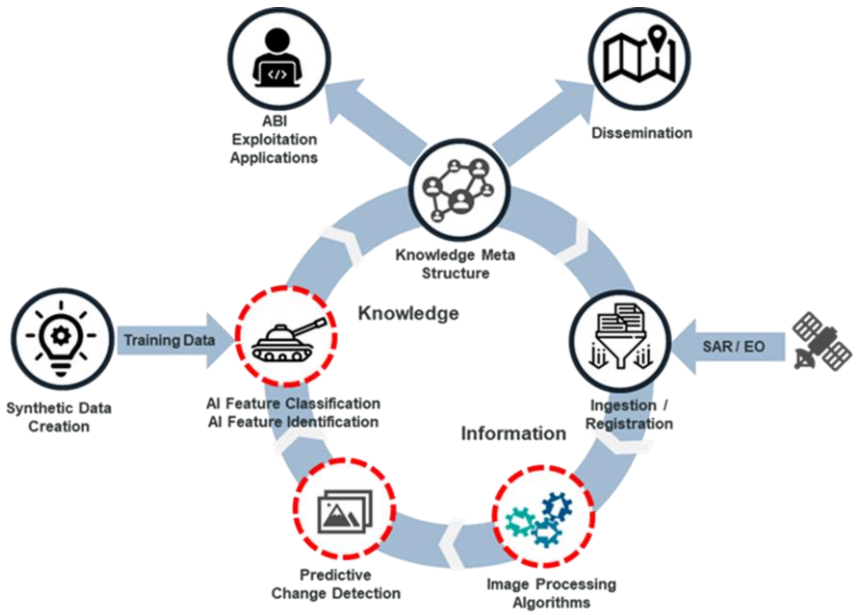

Figure 9. Knowledgebase

CONCLUSION

We described a system for estimating pixel heights from a single multispectral RGB image, with or without sensor metadata. System components included an ensemble of convolutional-deconvolutional neural network (CNN) models and an optimization function. The chosen deep learning network model was validated per pixel using high-resolution aerial RGB imagery and lidar datasets.

A data knowledgebase provided historic, time-stamped, multi-modal data for registration and 3D feature classification. Given a large amount of height truth data, a model was trained to recognize image features of differing heights, using CNN image-to-lidar regression. The models, when applied to an unseen image, estimated a preliminary height per pixel, based on a learned feature set. Multiple models were created and trained end-to-end and the best model and results were determined.

We used linear programming optimization with an ensemble of regression models and semantic segmentation information with a weighted classification model to decide optimized pixel height estimates. Semantic segmentation datasets help classify RGB imagery with feature class labels and refine land use feature classification with CNN classification to improve accuracy. We weighted each land use classified feature with a confidence metric that we used to help determine height information.

We used CNN regression for preliminary height estimation and CNN classification for land use feature classification plus a linear programming reward matrix per pixel to automatically decide optimized height estimation. An updated volumetric knowledgebase contains the system output and can be used subsequently for change detection and situational awareness. Both qualitative and quantitative analyses were performed and visualized.

REFERENCES

[1] Alidoost, F., Arefi, H., & Tombari, F. (2019). 2D Image-To-3D Model: Knowledge-Based 3D Building Reconstruction (3DBR) Using Single Aerial Images and Convolutional Neural Networks (CNNs). *Remote Sensing*, *11*(19), 2219.

[2] Baltsavias, E. P., & Hahn, M. (2000). Integrating Spatial Information and Image Analysis - One Plus One Makes Ten. *IAPRS*, *33*.

[3] Bandukwala, F. (2011, June). Extracting spatially and spectrally coherent regions from multispectral images. In *CVPR 2011 WORKSHOPS* (pp. 82-87). IEEE.

[4] Casser, V., Pirk, S., Mahjourian, R., & Angelova, A. (2019, July). Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 8001-8008).

[5] Guo, G., Yu, S., & Ma, S. (1998, October). An image labeling algorithm based on cooperative game theory.

[6] Duan, L., Desbrun, M., Giraud, A., Trastour, F., & Laurore, L. (2019). Large-scale DTM Generation from Satellite Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[7] El Garouani, A., Alobeid, A., & El Garouani, S. (2014). Digital surface model based on aerial image stereo pairs for 3D building. *International Journal of Sustainable Built Environment*, *3* (1), 119-126.

[8] Fu, H., Gong, M., Wang, C., Batmanghelich, K., & Tao, D. (2018). Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2002-2011).

[9] Hanif, M. S., & Bilal, M. (2020). Competitive residual neural network for image classification. *ICT Express*, *6*(1), 28-37.

[10] Ibragimov, B., Likar, B., & Pernus, F. (2012). A game-theoretic framework for landmark-based image segmentation. *IEEE Transactions on Medical Imaging*, *31*(9), 1761-1776.

[11] Kemker, R., Salvaggio, C., & Kanan, C. (2017). High-resolution multispectral dataset for semantic segmentation. *arXiv preprint arXiv:1703.01918*.

[12] Liu, B., Gould, S., & Koller, D. (2010, June). Single image depth estimation from predicted semantic labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1253-1260). IEEE.

[13] McGonagle, M., Brown, P., Tipping, A., Faulkner, T., Rahmes, M., Yates, J.H., "Cost Effective Evaluation of Change Detection", ASPRS 2012 Annual Conference, Tampa, Florida, October 31, 2012.

[14] Migut, M. A., van Gemert, J. C., & Worring, M. (2011, October). Interactive decision-making using dissimilarity to visually represented prototypes. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 141-149). IEEE.

[15] Mou, L., & Zhu, X. X. (2018). IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. *arXiv preprint arXiv:1802.10249*.

[16] Rahmes, M., Akbari, M., McGonagle, M., "Optimizing Supervised Learning for Pixel Labeling and Classification", ITSEC, Dec 2014.

[17] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234-241). Springer, Cham.

[18] Samadzadegan, F., Abbaspour, R. A., & Hahn, M. Automatic Change Detection of Geospatial Databases Based on a Decision-Level Fusion Technique.

[19] Scarborough, S., Lemanski, C., Nichols, H., Owirka, G., Minardi, M., and Hale, T. "SAR Change Detection MTI." In Algorithms for Synthetic Aperture Radar Imagery XIII, vol. 6237, p. 62370V. International Society for Optics and Photonics, 2006.

[20] Tian, Y., Krishnan, D., & Isola, P. (2019). Contrastive Representation Distillation. *arXiv preprint arXiv:1910.10699*.

[21] Tsoumakas, G., Katakis, I., & Vlahavas, I. (2009). Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook* (pp. 667-685). Springer, Boston, MA.

[22] Walter, V. (1999). Automated GIS Data Collection and Update.

[23] Winston, W., & Goldberg, J. B. (2003). Operations Research Applications and Algorithms 4th. *Belmont, CA: Brooks/Cole*.