

MULTI-MODAL SENSOR REGISTRATION FOR VEHICLE PERCEPTION VIA DEEP NEURAL NETWORKS

Michael Giering, Kishore Reddy, Vivek Venugopalan

Decision Support & Machine Intelligence Group

United Technologies Research Center

E. Hartford, CT 06060, USA

Email: {gierinmj, reddykk, venugov}@utrc.utc.com

ABSTRACT

The ability to simultaneously leverage multiple modes of sensor information is critical for perception of an automated vehicle’s physical surroundings. Spatio-temporal alignment of registration of the incoming information is often a prerequisite to analyzing the fused data. The persistence and reliability of multi-modal registration is therefore the key to the stability of decision support systems ingesting the fused information. LiDAR-video systems like on those many driverless cars are a common example of where keeping the LiDAR and video channels registered to common physical features is important. We develop a deep learning method that takes multiple channels of heterogeneous data, to detect the misalignment of the LiDAR-video inputs. A number of variations were tested on the Ford LiDAR-video driving test data set and will be discussed. To the best of our knowledge the use of multi-modal deep convolutional neural networks for dynamic real-time LiDAR-video registration has not been presented.

1 MOTIVATION

Navigation and situational awareness of optionally manned vehicles requires the integration of multiple sensing modalities such as Light Detection and Ranging (LiDAR) and video, but could just as easily be extended to other modalities including Radio Detection And Ranging (RADAR), Short-Wavelength Infrared (SWIR) and Global Positioning System (GPS). Spatio-temporal registration of information from multi-modal sensors is technically challenging in its own right. For many tasks such as pedestrian and object detection tasks that make use of multiple sensors, decision support methods rest on the assumption of proper registration. Most approaches Bodensteiner & Arens (2012) in LiDAR-video for instance, build separate vision and LiDAR feature extraction methods and identify common anchor points in both. Alternatively, by generating a single feature set on LiDAR, Video and optical flow, it enables the system to capture mutual information among modalities more efficiently. The ability to dynamically register information from the available data channels for perception related tasks can alleviate the need for anchor points *between* sensor modalities. We see auto-registration as a prerequisite need for operating on multi-modal information with confidence.

Deep neural networks (DNN) lend themselves in a seamless manner for data fusion on time series data. It has been shown [Ngiam et al. (2011)] for some challenges in which the modalities share significant mutual information, the features generated on the fused information can provide insight that neither input alone can. In effect the ML version of, “the whole is greater than the sum of its parts”.

Autonomous navigation places significant constraints on the speed of perception algorithms and their ability to drive decision making in real-time. Though computationally intensive to train, our implemented DCNN run easily within our real-time frame rates of 8 fps and could accommodate more standard rates of 30 fps. With most research in deep neural networks focused on algorithmic improvements and novel applications, a significant benefit to applied researchers is sometimes under appreciated. The automated feature generation of DNNs enables us to create multi-modal systems with far less overhead. The need for domain experts and hand-crafted feature design are lessened,

allowing more rapid prototyping and testing. The generalization of auto-registration across multiple assets is clearly a path to be explored.

In this paper, the main contributions are: (i) formulation of an image registration problem as a fusion of modalities from different sensors, namely LIDAR (L), video (Grayscale or R,G,B) and optical flow (U,V); (ii) performance evaluation of deep convolutional neural networks (DCNN) with various input parameters, such as kernel filter size and different combinations of input channels (R,G,B,Gr,L,U,V); (iii) fusion of patch-level and image-level predictions to generate alignment at the frame-level. The experiments were conducted using a publicly available dataset from FORD and the University of Michigan [Pandey et al. (2011)]. The DCNN implementation was executed on an NVIDIA Tesla K40 GPU with 2880 cores and compute power of 5 TFLOPS (single precision). The paper is organized into the following sections: Section 1 describes the introduction and motivation for this work; Section 2 provides a survey of the related work; the problem formulation along with the dataset description and the preprocessing is explained in Section 3; Section 4 gives the details of the DCNN setup for the different experiments; Section 5 describes the experiments and the post-processing steps for visualizing the qualitative results; finally Section 6 summarizes the paper and concludes with future research thrusts.

2 PREVIOUS WORK

A great amount has been published on various multi-modal fusion methods [Ross & Jain (2003), Gregor & LeCun (2011), Wu et al. (2004), Snoek et al. (2006)]. The most common approaches taken generate features of interest in each modality separately and create a decision support mechanism that aggregates features across modalities. If spatial alignment is required across modalities, as it is for LiDAR-video such filter methods Thrun (2011) are required to ensure proper inter-modal registration. These filter methods for leveraging 3D LiDAR and 2D images are often geometric in nature and make use of projections between the different data spaces.

Automatic registration of 2D video and 3D LiDAR has been a widely researched topic for over a decade [Wang & Neumann (2009) Kim et al. (2014) Mastin et al. (2009) Bodensteiner & Arens (2012)]. Its application in real-time autonomous navigation makes it a challenging problem. Majority of the 2D-3D registration algorithms are based on feature matching. Geometric features like corners and edges are extracted from detected vanishing points [Liu & Stamos (2007), Ding et al. (2008)], line segments [Frueh et al. (2004), Stamos et al. (2008)], and shadows [Troccoli & Allen (2004)]. Feature based approaches generally rely on dense 3D point cloud and additional knowledge of relative sun position and GPS/inertial navigation system (INS). Another approach used for video and LiDAR auto-registration is to reconstruct 3D point cloud from video sequences using structure from motion (SFM) and performing 3D-3D registration [Zhao et al. (2004) Liu et al. (2006)]. 3D-3D registration is more difficult and computationally expensive compared to 2D-3D registration.

The use of deep neural networks to analyze multi-modal sensor inputs has increased sharply in just the last few years, including audio-video Ngiam et al. (2011) Kim et al. (2013), image/text Srivastava & Salakhutdinov (2012), image/depth Lenz et al. (2013) and LiDAR-video To the best of our knowledge the use of multi-modal deep neural networks for dynamic LiDAR-video registration has not been presented.

A common challenge for data fusion methods is deciding at what level features from the differing sensor streams should be brought together. The deep neural network (DNN) approach most similar to the more traditional data fusion methods is to train DNNs independently on sensor modalities and then use the high-level outputs of those networks as inputs to a subsequent aggregator, which could also be a DNN. This is analogous to the earlier example of learning 3D/2D features and the process of identifying common geometric features.

It is possible however to apply DNNs with a more agnostic view enabling a unified set of features to be learned across multi-modal data. In these cases the input channels aren't differentiated. Un-supervised methods including deep Boltzman machines and deep auto-encoders for learning such joint representations have been successful.

Deep convolutional neural networks (DCNN) enable a similar agnostic approach to input channels. A significant difference is that target data is required to train them as classifiers. This is the approach chosen by us for automating the registration of LiDAR-video and optical-flow, in which we are

combining 1D/3D/2D data representations respectively to learn a unified model across as many as 6D.

3 PROBLEM STATEMENT

Being able to detect and correct the misalignment (registration, calibration) among sensors of the same or different kinds, is critical for decision support systems operating on their fused information streams. For our work DCNNs were implemented for the detection of small spatial misalignments in LiDAR and Video frames. The methodology is directly applicable to temporal registration as well. LiDAR-video data collected from a driverless car was chosen for the multi-modal fusion test case. LiDAR-video is a common combination for providing perception capabilities to many types of ground and airborne platforms including driverless cars Thrun (2011).

3.1 FORD LiDAR-VIDEO DATASET AND EXPERIMENTAL SETUP



Figure 1: Left: The modified Ford F-250 pickup truck. Right: Sample image from front facing camera and green dots indicate the region of LiDAR data.

The FORD LiDAR-video dataset Pandey et al. (2011) is collected by an autonomous Ford F-250 vehicle integrated with the following perception and navigation sensors as shown in Figure 1:

- Velodyne HDL-64E LiDAR with two blocks of lasers spinning at 10 Hz and a maximum range of 120m.
- Point Grey Ladybug3 omni-directional camera system with six 2-Mega-pixel cameras collecting video data at 8fps with 1600×1600 resolution.
- Two Riegl LMS-Q120 LIDAR sensors installed in the front of the vehicle generating range and intensity data when the laser sweeps its 80° field of view (FOV).
- Applanix POS-LV420 INS with Trimble GPS system providing the 6 degrees of freedom (DOF) estimates at 100 Hz.
- Xsens MTi-G sensor consisting of accelerometer, gyroscope, magnetometer, integrated GPS receiver, static pressure sensor and temperature sensor. It measures the GPS coordinates of the vehicle and also provides the 3D velocity and 3D rate of turn.

This dataset is generated by the vehicle while driving in and around the Ford research campus and downtown Michigan. The data includes feature rich downtown areas as well as featureless empty parking lots. As shown in Figure 2, we divided the data set into training and testing sections A to B and C to D respectively. They were chosen in a manner that minimizes the likelihood of contamination between training and testing. Because of this, the direction of the light source is never the same in the testing and training sets.

3.2 OPTICAL FLOW

In the area of navigation of mobile robots, optical flow has been widely used to estimate egomotion [Prazdny (1980)], depth maps [Shaharay & Brown (1988)], reconstruct dynamic 3D scene depth

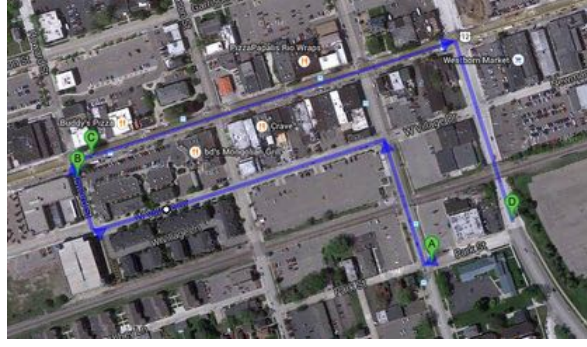


Figure 2: Training (A to B) and testing (C to D) tracks in the downtown Dearborn Michigan.



Figure 3: Optical flow: Hue indicates orientation and saturation indicates magnitude

[Yang et al. (2012)], and segment moving objects [Chien et al. (2002)]. Optical flow provides information of the scene dynamics and is expressed as an estimate of velocity at each pixel from two consecutive frames, denoted by \vec{u} and \vec{v} . The motion field from these two frames is measured by the motion of the pixel brightness pattern, where the changes in image brightness is due to the camera or object motion. Liu (2009) describes an algorithm for computing optical flow from images, which is used during the preprocessing step. Figure 3 shows an example of the optical flow computed using two consecutive frames from the Ford LiDAR-video dataset. By including optical flow as input channels, we imbue the DCNN with information on the dynamics observed across time steps.

3.3 PREPROCESSING

At each video frame timestep, the inputs to our model consist of C channels of data with C ranging from 3-6 channels. Channels consist of grayscale Gr or (R,G,B) information from the video, horizontal and vertical components of optical flow (U,V) and depth information L from LiDAR. The data from each modality is reshaped to a fixed size of 800×256 values, which are partitioned into $p \times p$ patches at a prescribed stride. Each patch $p \times p$ is stacked across C channels, effectively generating a vector of C dimensions. The different preprocessing parameters are denoted by patch size p , stride s and the number of input channels C .

Preprocessing is repeated N times, where N is the number of offset classes. For each offset class, the video (R,G,B) and optical flow (U,V) channels are kept static and the depth (L) channel from the LiDAR is moved by the offset simulating a misalignment between the video and the LiDAR sensors. In order to accurately detect the misalignment in the LiDAR and Video sensor data, a threshold is set to limit the information available in each channel. The LiDAR data has regions of sparsity and hence the LiDAR patches with a variance $(\sigma^2 < 15\%)$ are dropped from the final dataset. This leads to the elimination of the majority of foreground patches in the data set, reducing the size of the training and testing set by approximately 80%. Figure 4(a) shows a $N = 9$ class elliptically distributed set of offsets and Figure 4(b) shows a $p \times p$ patch stacked across all the different C channels.

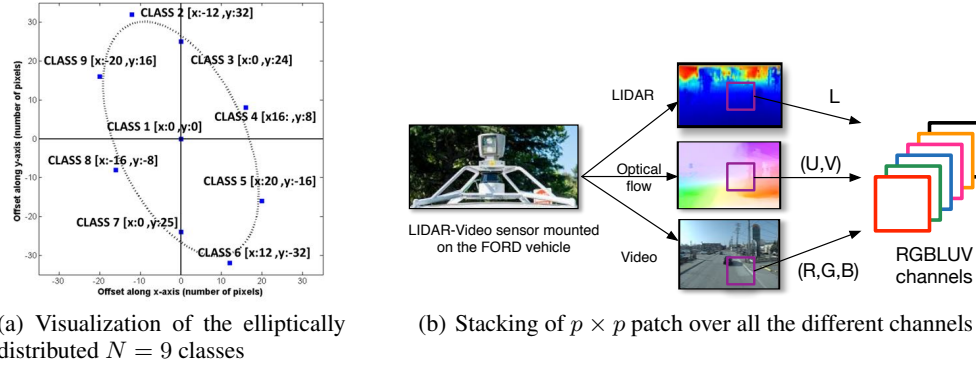
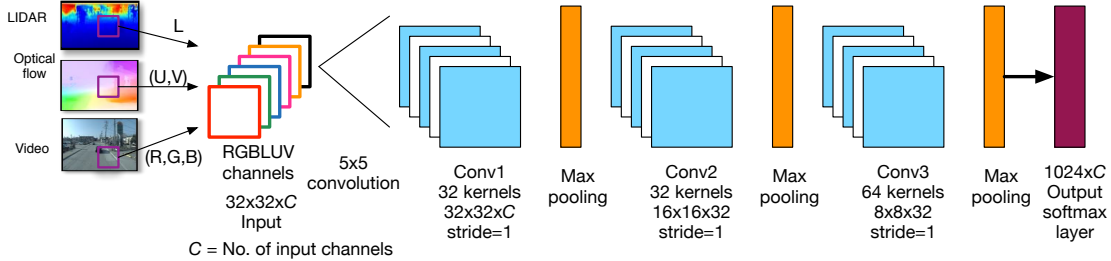


Figure 4: Preprocessing steps

4 MODEL DESCRIPTION

Our models for auto-registration are DCNNs trained to classify the current misalignment of the LiDAR-video data streams into one of a predefined set of offsets. DCNNs are probably the most successful deep learning model to date on fielded applications. The fact that the algorithm shares weights in the training phase, results in fewer model parameters and more efficient training. DCNNs are particularly useful for problems in which local structure is important, such as object recognition in images and temporal information for voice recognition. The alternating steps of convolution and pooling generates features at multiple scales which in turn imbues DCNN's with scale invariant characteristics.

Figure 5: Experimental setup of the LiDAR-video DCNN with 5×5 convolution

The model shown in Figure 5 consists of 3 pairs of convolution-pooling layers, that estimates the offset between the LiDAR-video inputs at each time step. For each patch within a timestep, there are O variants with the LiDAR-video-optical flow inputs offset by the predetermined amounts. The CNN outputs to a softmax layer, thereby providing an offset classification value for each patch of the frame. As described in Section 3.3, 32×32 patches were stacked across the different channels and provided as the input to the DCNN. All the 6 channels *RGBLUV* were used for the majority of the experiments, whereas only 4 channels were required for the *RGBL* and the *GrLUV* experiments. The first convolutional layer uses 32 filters (or kernels) of size $5 \times 5 \times C$ with a stride of 1 pixel and padding of 2 pixels on the edges. The following pooling layer generates the input data (of size $16 \times 16 \times 32$) for the second convolutional layer. This layer uses 32 filters of size $5 \times 5 \times 32$ with a stride of 1 pixel and padding of 2 pixels on the edges. A second pooling layer, similar to the first one is used to generate input with size $8 \times 8 \times 32$ for the third convolutional layer that uses 64 filters of size $5 \times 5 \times 32$ with the stride and padding same as previous convolutional layer. The third pooling layer with similar configuration as the two previous pooling layers connects to an output softmax layer with labels corresponding to the $N = 9$ classes. The DCNN described above was trained using stochastic gradient descent with a mini-batch size of 100 epochs.

The NVIDIA Kepler series K40 GPUs [NVIDIA Inc. (2012)] are very FLOPS/Watt efficient and are being used to drive real-time image processing capabilities [Venugopal & Kannan (2013)]. These GPUs consist of 2880 cores with 12 GB of on-board device memory (RAM). Deep Learning appli-

cations have been targeted on GPUs previously in [Krizhevsky et al. (2012)] and these implementations are both compute and memory bound. Stacking of channels results in a vector of $32 \times 32 \times C$, which is suitable for the Single Instruction Multiple Datapath (SIMD) architecture of the GPUs. At the same time, the training batch size caches in the GPU memory, so the utilization of the K40 GPU's memory is very high. This also results in our experiments to run successfully on a single GPU instead of partitioning the different layers over multiple GPUs.

5 EXPERIMENTS

5.1 DATASET USING ELLIPTICALLY DISTRIBUTED OFFSETS

In our experiments, elliptically distributed set of 9 offsets of the LiDAR-video data were considered. The LiDAR data is displaced along an ellipse with a major axis of 32 pixels and a minor axis of 16 rotated clockwise from x-axis by 45 degrees as shown in Figure 4(a). Separate training and testing sets were generated from two different tracks as shown in Figure 2 for all the 9 offsets of LiDAR data. Training and testing tracks have never seen regions and also have different lighting conditions. Our preprocessing step described in Section 3.3 results in 223371 and 126513 patches for testing and training extracted from 469 and 224 images respectively.

In the testing phase, for each frame a simple voting scheme is used to aggregate the patch level offset predictions to a single frame level prediction. A sample histogram of the patch level predictions is shown in Figure 6. We color each patch of the frame with a color corresponding to the predicted class as shown in Figure 6.

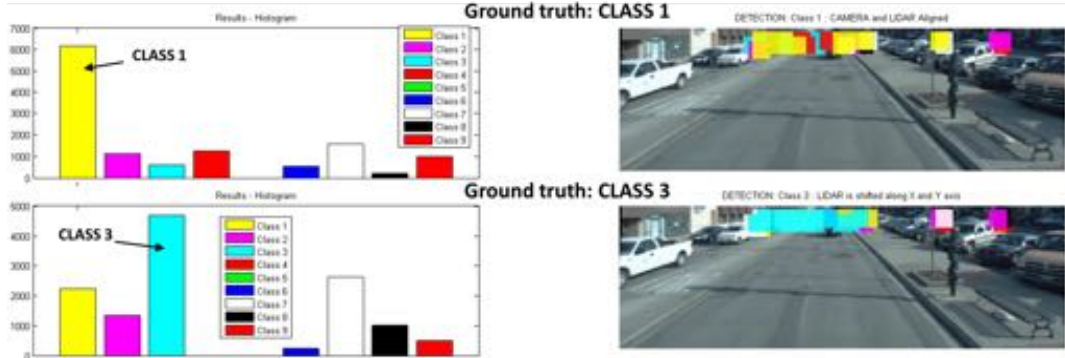


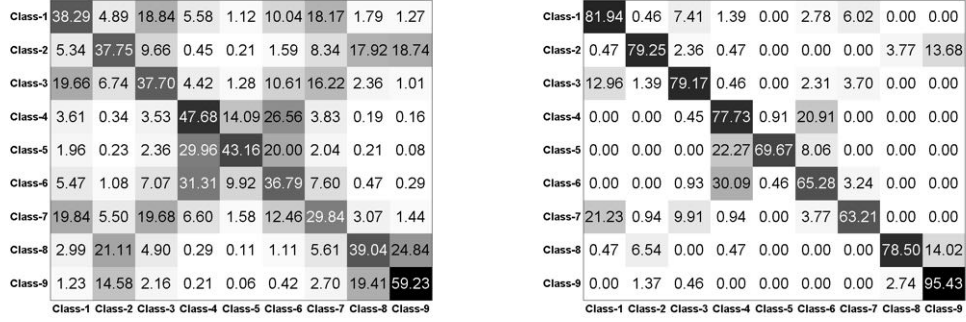
Figure 6: Left: Histogram of the classes detected by all patches. Right: Location of patches color coded with the predicted class

5.2 EXPERIMENTAL RESULTS

Table 1 lists the inputs and CNN parameters explored ranked in the order of increasing accuracy. We averaged the values across the diagonal of the confusion matrix to determine the image level and patch level accuracy. Patch level accuracy is the individual performance of all the 32×32 patches from the testing images. Classification of patches belonging to a single time-step are voted to predict the shift for image level accuracy. In Table 1, the first 3 columns show the results for different number of filter combinations in the convolutional layers with fixed number of filters and input channels *RGLUV*. We observed that the image and patch level accuracy decreased with the increase in the number of filters. For experiments shown in columns 4 and 5, the filter size was increased, with the number of filters constant at (32, 32, 64). We observed that for the 6 channels *RGLUV*, filter size of 9 gave the best image level accuracy of 63.03%. Column 6 shows the results of our experiment after dropping the optical flow *UV* channels. The image and patch level accuracy decreased for this case, indicating that optical flow contributed significantly towards image registration. The remaining experiments utilized the Grayscale information *Gr* instead of *RGB* and produced the best results with 76.69% and 41.05% image and patch level accuracy respectively.

Table 1: Results of different combination of input channels C and CNN parameters

Channels	RGBLUV			RGBL		GrLUV	
Filter size	5			7	9	5	9
# of filters	(32,32,32)	(32,32,64)	(64,64,64)	(32,32,64)			
Image Level accuracy	61.75	61.06	60.09	61.79	63.03	60.66	68.03
Patch Level accuracy	38.74	38.57	38.49	38.03	39.00	39.28	40.96



(a) Patch level confusion matrix (41.05% accuracy) (b) Image level confusion matrix (76.69% accuracy)

Figure 7: Confusion Matrix for elliptically distributed $N = 9$ classes using Greyscale, Optical Flow and LiDAR channels with a filter size of 9

6 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a deep learning method to do LiDAR-Video registration. We demonstrated the effect of filter size, number of filters and different channels. We also showed the advantage of optical flow and grey scale. The next step in taking this work forward is to complete our development of a deep auto-registration method for ground and aerial platforms requiring no apriori calibration ground truth. Our aerial applications in particular present noisier data with an increased number of degrees of freedom. The extension of these methods to simultaneously register information across multiple platforms and larger numbers of modalities will provide interesting challenges that we look forward to working on.

REFERENCES

- Bodensteiner, Christoph and Arens, Michael. Real-time 2D Video 3D LiDAR Registration. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 2206–2209. IEEE, 2012.
- Chien, Shao-Yi, Ma, Shyh-Yih, and Chen, Liang-Gee. Efficient moving object segmentation algorithm using background registration technique. *Circuits and Systems for Video Technology, IEEE Transactions on*, 12(7):577–586, Jul 2002. ISSN 1051-8215. doi: 10.1109/TCSVT.2002.800516.
- Ding, Min, Lyngbaek, K., and Zakhori, A. Automatic registration of aerial imagery with untextured 3d lidar models. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, June 2008. doi: 10.1109/CVPR.2008.4587661.
- Frueh, C., Sammon, R., and Zakhori, A. Automated texture mapping of 3d city models with oblique aerial imagery. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pp. 396–403, Sept 2004. doi: 10.1109/TDPVT.2004.1335266.
- Gregor, Karol and LeCun, Yann. Learning Representations By Maximizing Compression. *arXiv preprint arXiv:1108.1169*, 2011.

- Kim, Hyojin, Correa, Carlos D, and Max, Nelson. Automatic registration of lidar and optical imagery using depth map stereo. In *Computational Photography (ICCP), 2014 IEEE International Conference on*, pp. 1–8. IEEE, 2014.
- Kim, Yelin, Lee, Honglak, and Provost, Emily Mower. Deep Learning for Robust Feature Generation in Audiovisual Emotion Recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 3687–3691. IEEE, 2013.
- Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet Classification With Deep Convolutional Neural Networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- Lenz, Ian, Lee, Honglak, and Saxena, Ashutosh. Deep Learning for Detecting Robotic Grasps. *arXiv preprint arXiv:1301.3592*, 2013.
- Liu, Ce. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2009. AAI0822221.
- Liu, Lingyun and Stamos, I. A systematic approach for 2d-image to 3d-range registration in urban environments. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, Oct 2007. doi: 10.1109/ICCV.2007.4409215.
- Liu, Lingyun, Stamos, I., Yu, Gene, Wolberg, G., and Zokai, S. Multiview geometry for texture mapping 2d images onto 3d range data. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pp. 2293–2300, June 2006. doi: 10.1109/CVPR.2006.204.
- Mastin, Andrew, Kepner, Jeremy, and Fisher, J. Automatic Registration of LIDAR and Optical Images of Urban Scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2639–2646. IEEE, 2009.
- Ngiam, Jiquan, Khosla, Aditya, Kim, Mingyu, Nam, Juhan, Lee, Honglak, and Ng, Andrew Y. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 689–696, 2011.
- NVIDIA Inc. NVIDIA’s Next Generation CUDA Compute Architecture: Kepler TM GK110. Whitepaper, May 2012.
- Pandey, Gaurav, McBride, James R, and Eustice, Ryan M. Ford Campus Vision And Lidar Data Set. *The International Journal of Robotics Research*, 30(13):1543–1552, 2011.
- Prazdny, K. Egomotion and relative depth map from optical flow. *Biological Cybernetics*, 36(2): 87–102, 1980. ISSN 0340-1200. doi: 10.1007/BF00361077. URL <http://dx.doi.org/10.1007/BF00361077>.
- Ross, Arun and Jain, Anil. Information Fusion In Biometrics. *Pattern recognition letters*, 24(13): 2115–2125, 2003.
- Shahraray, B. and Brown, M.K. Robust depth estimation from optical flow. In *Computer Vision., Second International Conference on*, pp. 641–650, Dec 1988. doi: 10.1109/CCV.1988.590045.
- Snoek, Cees GM, Worring, Marcel, Van Gemert, Jan C, Geusebroek, Jan-Mark, and Smeulders, Arnold WM. The Challenge Problem For Automated Detection Of 101 Semantic Concepts In Multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pp. 421–430. ACM, 2006.
- Srivastava, Nitish and Salakhutdinov, Ruslan. Multimodal Learning With Deep Boltzmann Machines. In *Advances in neural information processing systems*, pp. 2222–2230, 2012.
- Stamos, Ioannis, Liu, Lingyun, Chen, Chao, Wolberg, George, Yu, Gene, and Zokai, Siavash. Integrating automated range registration with multiview geometry for the photorealistic modeling of large-scale scenes. *International Journal of Computer Vision*, 78(2-3):237–260, 2008. ISSN 0920-5691. doi: 10.1007/s11263-007-0089-1. URL <http://dx.doi.org/10.1007/s11263-007-0089-1>.

- Thrun, Sebastian. Google's driverless car. *Ted Talk, Ed*, 2011.
- Troccoli, Alejandro J. and Allen, Peter K. A shadow based method for image to model registration. In *IEEE Workshop on Image and Video Registration, Conf. on Comp. Vision and*, 2004.
- Venugopal, Vivek and Kannan, Suresh. Accelerating real-time lidar data processing using gpus. In *Circuits and Systems (MWSCAS), 2013 IEEE 56th International Midwest Symposium on*, pp. 1168–1171, August 2013. doi: 10.1109/MWSCAS.2013.6674861.
- Wang, Lu and Neumann, Ulrich. A Robust Approach For Automatic Registration Of Aerial Images With Untextured Aerial LIDAR Data. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2623–2630. IEEE, 2009.
- Wu, Yi, Chang, Edward Y, Chang, Kevin Chen-Chuan, and Smith, John R. Optimal Multimodal Fusion For Multimedia Data Analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 572–579. ACM, 2004.
- Yang, Y., Liu, Q., Ji, R., and Gao, Y. Dynamic 3D Scene Depth Reconstruction via Optical Flow Field Rectification. *PLoS ONE*, 7:47041, November 2012. doi: 10.1371/journal.pone.0047041.
- Zhao, W., Nister, D., and Hsu, S. Alignment of continuous video onto 3d point clouds. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pp. II–II, June 2004. doi: 10.1109/CVPR.2004.1315269.