SANDIA REPORT
SAND2020-3703
Printed MMMM 2020

Sandia National Laboratories

Multimodal Data Fusion via Entropy Minimization

Lisa M. Linville, Joshua J. Michalenko, Dylan Z. Anderson

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185
Livermore, California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods

ABSTRACT

The use of gradient-based data-driven models to solve a range of real-world remote sensing problems can in practice be limited by the uniformity of available data. Use of data from disparate sensor types, resolutions, and qualities typically requires compromises based on assumptions that are made prior to model training and may not necessarily be optimal given over-arching objectives. For example, while deep neural networks (NNs) are state-of-the-art in a variety of target detection problems, training them typically requires either limiting the training data to a subset over which uniformity can be enforced or training independent models which subsequently require additional score fusion. The method we introduce here seeks to leverage the benefits of both approaches by allowing correlated inputs from different data sources to co-influence preferred model solutions, while maintaining flexibility over missing and mismatching data. In this work we propose a new data fusion technique for gradient updated models based on entropy minimization and experimentally validate it on a hyperspectral target detection dataset. We demonstrate superior performance compared to currently available techniques using a range of realistic data scenarios, where available data has limited spacial overlap and resolution.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Multi-sensor data fusion promises improvements in accuracy, persistence, and timeliness in remote sensing data exploitation. In the context of Visible through Near Infrared (VNIR) and Short Wave Infrared (SWIR) hyperspectral imagery, fusion promises enhanced target detection models that can leverage the unique phenomenology provided in each spectral range [3]. However, practical solutions for fusion remain elusive; in practice, VNIR-SWIR fusion is plagued by different collection footprints, spatial resolutions, and collection times. Contemporary data fusion pipelines typically rely on one of two different approaches, with each approach representing a different set of significant compromises for overall fusion performance [4].

In the first approach, sensor data are concatenated at the pixel level to construct a single multi-modal model (for example, see [5]). We denote this approach as Fused Model Input (FMI). FMI can yield a particularly powerful fused model as co-information between modes are used during model construction (for example, traits of the atmospheric profile apparent in the SWIR but not the VNIR). In practice, data available for FMI model fitting is limited as the concatenation process necessitates that both VNIR and SWIR be available to form model input. In the second approach, outputs from independent per-modality models are combined using some statistical criteria (for example, mean of model outputs or a more sophisticated technique such as described in [9, 10, 6, 1]). We denote this as Fused Model Output (FMO). FMO provides great flexibility and can significantly improve performance over using a single modality in isolation. Since models are created on a per-mode basis, all available imagery can be used for model fitting. However, FMO leaves crucial complementary information between modes unexploited during model fitting, thereby limiting overall fusion performance.

Both FMO and FMI represent significant compromises in overall fusion performance. In this paper, we develop entropy minimization (EMIN), which seeks to leverage the benefits of FMI, namely allowing correlated inputs from different data sources to co-influence preferred model solutions, while maintaining the flexibility afforded through FMO to handle missing or misaligned data to enhance overall fusion performance. EMIN is formulated as an additional penalty loss term that links observations containing multiple modalities, allowing information sharing between models when both VNIR and SWIR are available. This approach is reasonably flexible and can be integrated into gradient-based learning schemes such as deep neural networks. In Section 2, we describe the mathematical underpinnings of EMIN, in Section 3 we demonstrate the efficacy of EMIN over FMO and FMI in a VNIR-SWIR fusion benchmark, and we conclude in Section 4.

2. ENTROPY MINIMIZATION

We now detail a formulation of entropy minimzation as a useful penalty term for multimodal data fusion. Let $\mathbf{x} = (x^1, x^2, \ldots, x^M)$ be a sequence of observations denoting $m \in [M]$ separate modalities describing the input $\mathbf{x}$. In FMO, M independent models are trained to produce $p(y|x^m)$ which are subsequently combined into a single distribution $p(y|\mathbf{x})$ using various fusion techniques. Since $p(y|x^m)$ are independent, features learned from one model can not be used by another to aid in learning or inference. This results in poor performance when x^m are correlated,

which is true for most real world datasets. EMIN addresses this by explicitly linking parameter updates of model m to parameters of $\{\bar{m}\}$. One method to link these models during training is to add a penalty term which is a function of the M models. Such a function should describe a desirable property of the collective model outputs. For most data fusion problems we desire the following characteristics:

1. *Model agreement*: $\text{argmax}_i(p(y_i|x^m)) = c \quad \forall \quad m$ which subsequently makes score fusion trivial (majority vote).

2. *Confident predictions*: Output distributions are close to dirac delta function.

These characteristics are exemplified in the entropy function. Entropy [8], denoted as $H(\mathbf{X})$, is a measure of the uncertainty of a random variable $\mathbf{X}$ and is defined as:

$$H(\mathbf{X}) = -\sum_{i \in \mathscr{C}} p_i \log(p_i) \tag{1}$$

Where p_i is the probability $\mathbf{X}$ takes on the i^{th} outcome in the set of possible outcomes $\mathscr{C}$. $H(\mathbf{X})$ is maximized when $\mathbf{X}$ is distributed uniformly (maximum uncertainly) and is minimized when $p(\mathbf{X})$ takes on the dirac delta function (no uncertainty). To enforce the desired criteria, let

$$p(\mathbf{X}) = \frac{\sum_{m \in [M]} p(y|x^m)}{M} \tag{2}$$

be the empirical mean of all model outputs, and compute $H(\mathbf{X})$ as described in Eq 1. $H(\mathbf{X})$ is minimized (at a value of 0) when all models output the same delta function, encouraging both model agreement and confident predictions. In practice, the maximum value $H(\mathbf{X})$ takes on varies as a function of $|\mathscr{C}|$, therefore it can be useful to normalize entropy by $\log(|\mathscr{C}|)$ thereby creating a loss term that is bounded and behaves consistently regardless of the cardinality of model outcomes.

Training of the M models is performed similar to most traditional gradient based update schemes. The multimodal loss function takes the form:

$$\mathscr{L}(y, \hat{y}) = \left(\sum_{m \in [M]} \text{CE}(y, \hat{y_m}) \right) + \gamma H(\mathbf{X}) \tag{3}$$

Where γ is a tunable hyperparameter and CE stands for the traditional cross entropy loss for classification problems. We note the inclusion of cross entropy in $\mathscr{L}(y, \hat{y})$ is specific to data fusion of multiple classifiers, which may not be the case in other data fusion problems such as regression. Our contribution mainly lays in the $\gamma H(\mathbf{X})$ term of the loss. With the addition of $H(\mathbf{X})$, parameter updates of model m rely on the current parameters of the other models, effectively allowing for correlated inputs to be utilized during training and a large advantage of our framework over previous methods. In practice and during our hyperparameter search, we find that a constant value of γ works well, but still underperforms compared to when we ramp γ, increasing the value and subsequently the co-information sharing between models over training epochs.

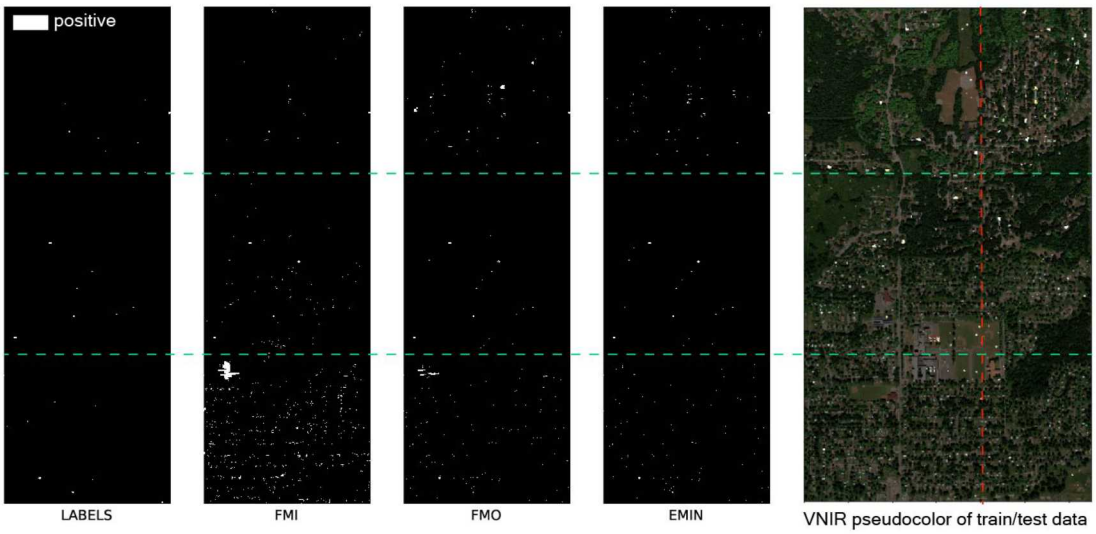

Figure 3-1 True target labels (left), method detection scores (middle three) and pseudo data overview (right). Green dashed lines indicate regions of missing VNIR/SWIR and red dashed line indicates train/test split.

3. EXPERIMENTS

We experimentally validate the efficacy of our proposed entropy minimization method by utilizing the hyperspectral target detection dataset described in [2]. The scene is built atop DIRSIG Megascene [7], which covers over half a square mile designed to represent an area of Northeast Rochester, NY. We utilize the mid-latitude summer atmosphere and 1200 render from [2], and use the inserted "green_paint_1" disks as targets to detect. The dataset consists of a single AVIRIS-like sensor spanning 0.4 to 2.5 μm with 10 nm band spacing. To construct a VNIR-SWIR fusion exemplar problem, we separate this data artificially into two "sensors", with VNIR composed of bands such that $\lambda_{vnir} \leq 0.9\mu m$ and SWIR composed of bands such that $\lambda_{swir} > 0.9\mu m$. This problem setup represents an ideal and unrealistic scenario for FMI: VNIR and SWIR are perfectly aligned and overlapping at the pixel level. To construct more realistic fusion scenarios we test performance over varying spatial overlaps and spectral resolutions. For overlap scenarios we drop the bottom (top) fraction of the VNIR (SWIR) data. This results in the top of the scene as VNIR only, the middle fraction as both VNIR and SWIR (overlap region), and the bottom as SWIR only. We test overlap fractions of 10%, 33%, 50%, and 90%. While simplistic, this simulates a common real problem in practical fusion scenarios in which sensor collection footprints are only partially overlapping. For multi-resolution scenarios we degrade the SWIR pixel resolution by Gaussian pyramids at 2.5x and 5.0625x downsampling. Given the more expensive and sophisticated lithographic processes needed for SWIR detectors, it is common for SWIR resolution to be much coarser than VNIR. We utilize the leftmost 50% of data to train, followed by 10% to validate, and the rightmost 40% of the 50/50 overlap scenario of Megascene to test on. See Fig. 3-1 for visualizations of the overlap fraction of 33%, or 1/3 for each region.

To focus on the effects of different fusion approaches, a relatively simple neural network is

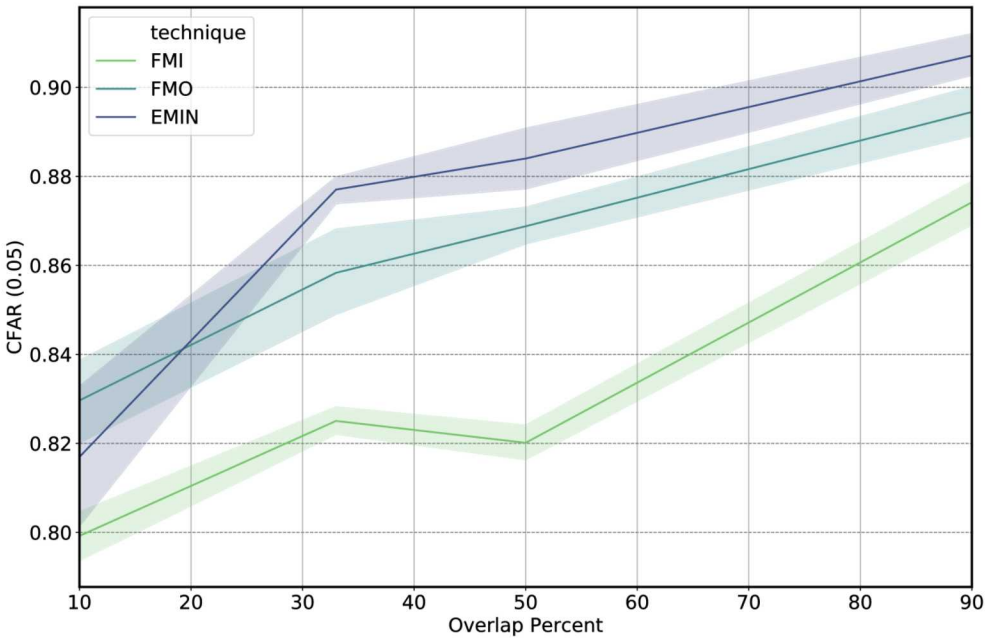

Figure 3-2 CFAR at threshold 0.05 for baseline techniques FMI and FMO compared to EMIN for 1-1 pixel resolution between VNIR and SWIR in different overlap scenarios.

utilized as a testbed for FMI, FMO, and EMIN. The model consists of 3 fully connected layers with 50, 25, and 10 hidden units and rectified linear unit (ReLU) activations between layers. Extensive hyperparameter search is performed over all major model parameters, with the addition of γ in the EMIN model to ensure well tuned baselines. Several statistical techniques for combining VNIR and SWIR model outputs under FMO and EMIN were compared, and a simple average was selected as providing good performance and simple implementation. Following [2], we utilize the probability of detection at 5% constant false alarm rate (CFAR) to evaluate and compare techniques.

By varying the amount of overlap between VNIR and SWIR we simulate realistic differences in collection footprints. The FMI baseline is expected to perform well when significant fractions of the collection footprints overlap (when the overlap fractions are large) whereas FMO is expected to perform well for individual data modalities even with limited collection uniformity. Fig. 3-2 highlights CFAR performance for the three techniques in different overlap scenarios when the pixel resolution for each modality is equivalent (resolution = 1). Both the FMI and FMO baselines increase with increasing overlap, with no saturation in FMO performance. EMIN outperforms both baselines when the overlap is above 20% increasing the average CFAR from .870 to .885. The lack of saturation for FMO may be a limitation of the dataset; the data available for model training may be insufficient to constrain the upper boundaries on the performance for each technique. Error bands represent the 95% confidence interval of CFAR for the top 10 models seen during hyperparameter search.

The resolution studies provide evidence that the performance increase for EMIN in Fig. 3-2 are conservative estimates. Fig. 3-3 by comparison shows the effects of varying SWIR resolution on

overall test performance. While the benefits of EMIN are modest compared to FMO and significant compared to FMI with equivalent pixel resolution between modalities (Fig. 3-2 and Fig. 3-3 far left), EMIN significantly outperforms both baselines as the SWIR resolution decreases (Fig. 3-3, center and right). Decreasing SWIR resolution (2.25x decimation) decreases overall test performance on the FMO baseline compared to no pixel degradation while maintaining EMIN performance despite the decrease in SWIR resolution. By comparison, significant pixel level degradation of SWIR resolution (5.0625x decimation) significantly decreases model performance for FMO and EMIN, while maintaining a large spread (4.1% difference in CFAR) between EMIN and FMO. FMI in comparison outperforms the FMO baseline likely because the context available through shared modalities during learning is more beneficial than the fusion performance from individual models in FMO for this data scenario. Overall, the more SWIR resolution is degraded the larger the average test performance increase we observe using EMIN compared to baseline models.

A different view of model performance can be acquired by partitioning the test data by modality. The bar chart in the top of Fig. 3-4 represents the overall test performance at degraded SWIR resolution in a 50% overlap data regime. The overall performance averages all sections of the test partition which includes 50% for which both SWIR and VNIR modalities are available, and the remaining 50% with either VNIR or SWIR only data (revisit Fig. 3-1 for data partitioning). As observed in Fig. 3-2, EMIN outperforms FMO and FMI baselines on the average overall test set. Likewise, EMIN outperforms baselines when only a single modality is available (Fig. 3-4; bottom). It remains an open question whether the exceptional VNIR performance (bottom left) in this data regime obfuscates the need for SWIR data during inference. Regardless, training with SWIR and VNIR using EMIN when overlap exists continues to enhance predictive performance even when only the VNIR modality is available.

The above discussion offers different views of data performance under different collect footprints and sensor resolutions scenarios. The complete set of data scenarios we test are reported in Table. 3-1 and although non-exhaustive provide a strong foundation for further exploration of data fusion through EMIN.

4. CONCLUSIONS

We outline a new method for multimodal data fusion called entropy minimization (EMIN). EMIN is formulated as a conjugate loss term sensitive to the expectation of decision agreement across disparate data sources. We test our new method on a synthetic hyperspectral target detection problem, and observe significant performance gains compared to baseline fusion techniques. In the overlap scenarios, FMI is data limited (can only utilize the overlapping data for training) and achieves a mean CFAR that significantly underperforms FMO and EMIN, but increases with increasing overlap fraction. EMIN offers the largest performance benefits under non-uniform data scenarios. We frame our analysis of EMIN in the context of VNIR-SWIR hyperspectral image fusion for target detection. However, the methodology and formulation of EMIN is not specific to this scenario, and should be readily applicable to other fusion scenarios. The next steps for this method are to explore the limit of data scenarios over which EMIN continues to outperform

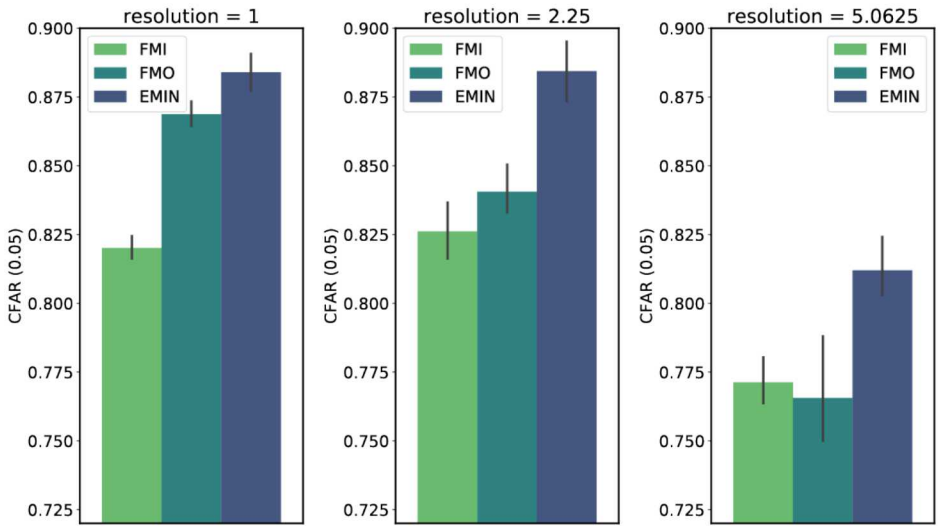

Figure 3-3 CFAR at threshold 0.05 for perfect pixel resolution for each modality (resolution = 1) and for degraded SWIR at 2.25 and 5.0625 at a 50% overlap fraction.

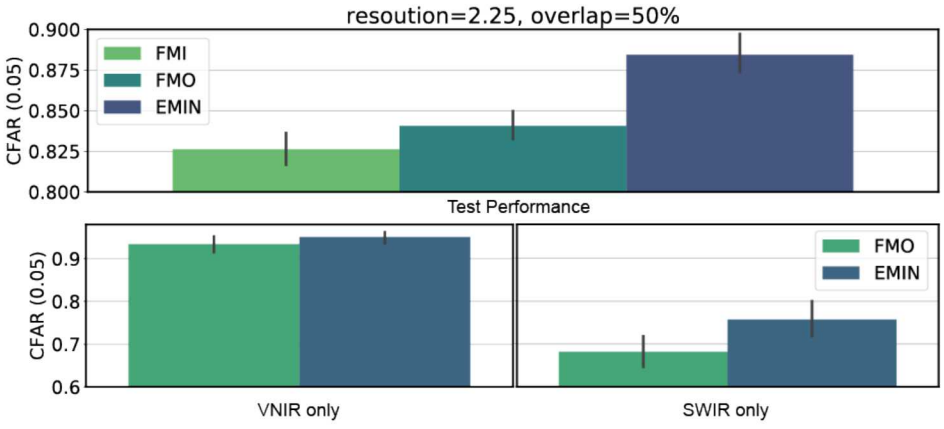

Figure 3-4 CFAR at threshold 0.05 for SWIR and VNIR only sections of test data.

Table 3-1 CFAR scores for all overlap and resolution data scenarios.

Modality	Overlap	Resolution	CFAR (0.05)
EMIN	10	1	0.816165
	33	1	0.877179
	50	1	0.884311
		2.25	0.87401
		5.0625	0.809033
	90	1	0.904913
FMI	10	1	0.798732
	33	1	0.824089
	50	1	0.820919
		2.25	0.816957
		5.0625	0.767829
	90	1	0.875594
FMO	10	1	0.829146
	33	1	0.854992
	50	1	0.868463
		2.25	0.836767
		5.0625	0.757528
	90	1	0.897781

baseline models, and demonstrate the efficacy of EMIN on target tasks over a wider variety of fusion applications and scenarios.

REFERENCES

[1] Dale N Anderson, Deborah K Fagan, Mark A Tinker, Gordon D Kraft, and Kevin D Hutchenson. A mathematical statistics formulation of the teleseismic explosion identification problem with multiple discriminants. *Bulletin of the Seismological Society of America*, 97(5):1730–1741, 2007.

[2] Dylan Z Anderson, Joshua D Zollweg, and Braden J Smith. Paired neural networks for hyperspectral target detection. In *Applications of Machine Learning*, volume 11139, page 111390J. International Society for Optics and Photonics, 2019.

[3] José M Bioucas-Dias, Antonio Plaza, Gustavo Camps-Valls, Paul Scheunders, Nasser Nasrabadi, and Jocelyn Chanussot. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and remote sensing magazine*, 1(2):6–36, 2013.

[4] Federico Castanedo. A review of data fusion techniques. *The Scientific World Journal*, 2013, 2013.

[5] Yushi Chen, Hanlu Jiang, Chunyang Li, Xiuping Jia, and Pedram Ghamisi. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10):6232–6251, 2016.

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[7] Emmett J Ientilucci and Scott D Brown. Advances in wide-area hyperspectral image simulation. In *Targets and Backgrounds IX: Characterization and Representation*, volume 5075, pages 110–121. International Society for Optics and Photonics, 2003.

[8] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[9] Katherine M. Simonson. Probabilistic fusion of atr results. Technical report, Sandia National Laboratories (SNL-NM), Albuquerque, NM, 1998.

[10] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

DISTRIBUTION

Email—Internal (encrypt for OUO)

Name	Org.	Sandia Email Address
Technical Library	01177	libref@sandia.gov

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.