ROUTING AND ACTION

MEMORANDUM

ROUTING

TO:(1) (,)

Report is available for review

(2) Proposal Files Report No.: Proposal Number:

DESCRIPTION OF MATERIAL

CONTRACT OR GRANT NUMBER:

INSTITUTION:

PRINCIPAL INVESTIGATOR:

TYPE REPORT:

DATE RECEIVED:

PERIOD COVERED: through

TITLE:

ACTION TAKEN BY DIVISION

(x) Report has been reviewed for technical sufficiency and satisfactory.

Based on my technical review, I have identified no OPSEC or Technology Protection concerns that need to be addressed regarding this report.

(x) Perfomance of the research effort was accomplished in a satisfactory manner and all other technical requirements have been fulfilled.

(x) Based upon my knowledge of the research project, I agree with the patent information disclosed.

Approved by on

ARO FORM 36-E

REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
01-09-2023	Final Report	29-May-2018 - 24-May-2020

4. TITLE AND SUBTITLE
Final Report: Instrumentation for Multimodal Data Analysis for Representation Learning and Visual Recognition

5a. CONTRACT NUMBER
W911NF-18-1-0220

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611103

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Johns Hopkins University
3400 North Charles Street
Malone 146
Baltimore, MD 21218 -2608

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
72161-MI-RIP.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Rene Vidal
UU	UU	UU	UU		19b. TELEPHONE NUMBER 410-369-8115

RPPR Final Report
as of 08-Sep-2023

Agency Code: 21XD

Proposal Number: 72161MIRIP **Agreement Number: W911NF-18-1-0220**
INVESTIGATOR(S):

Name: Haider Ali
Email: hali@jhu.edu
Phone Number: 4105163286
Principal: N

Name: Rene Vidal
Email: vidalr@seas.upenn.edu
Phone Number: 4103698115
Principal: Y

Organization: **Johns Hopkins University**
Address: 3400 North Charles Street, Baltimore, MD 212182608
Country: USA

DUNS Number: 001910777 EIN: 520595110
Report Date: 24-Aug-2020 Date Received: 01-Sep-2023
Final Report for Period Beginning 29-May-2018 and Ending 24-May-2020
Title: Instrumentation for Multimodal Data Analysis for Representation Learning and Visual Recognition
Begin Performance Period: 29-May-2018 **End Performance Period:** 24-May-2020
Report Term: 0-Other

Submitted By: Rene Vidal Email: vidalr@seas.upenn.edu
Phone: (410) 369-8115

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 0 **STEM Participants:**

Major Goals: See attached

Accomplishments: See attached

Training Opportunities: See attached

Results Dissemination: See attached

Honors and Awards: See attached

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Rene Vidal
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

RPPR Final Report
as of 08-Sep-2023

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Rene Vidal
Signature Date: 9/1/23 1:21AM

Final Report for MURI Award W911NF1910354

Instrumentation for Multimodal Data Analysis for Representation Learning and Visual Recognition

5/29/2018-6/24/2020

Principal Investigator (PI): Prof. René Vidal, Mathematical Institute for Data Science and Department of Biomedical Engineering, Johns Hopkins University. Email: rvidal@jhu.edu, Phone: (410) 516-7306.

MAJOR GOALS

The goal of this project was to setup a shared GPU computing system for the implementation and evaluation of novel algorithms for (1) characterizing semantic information content in multimodal data, such as text, images, and videos, and (2) recognizing actions in video data.

ACCOMPLISHMENTS UNDER GOALS

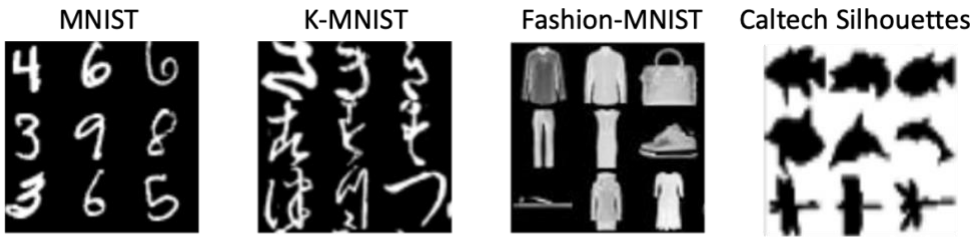
Evaluation of Algorithms for Characterizing Semantic Information Content

In work originally submitted to ICLR2020 and later published in TPAMI 2022 [1], we developed a novel framework for characterizing the semantic information content in multimodal data. In the proposed framework, we defined a notion of semantic entropy as the minimum expected number of semantic queries (from a user-specified query set) that need to be asked about data in order to solve a task. We implemented this framework using a greedy algorithm called information pursuit (IP), which selects one query at a time in order of information gain. The implementation of the framework involves learning a complex probabilistic generative model that relates data, queries, answers, and predictions, and using this model to determine the most informative queries.

We evaluated our proposed measure of semantic entropy and the IP strategy for various binary image classification tasks of increasing complexity. Rather than simply training a classifier (e.g. an SVM or a neural network), our goal is to quantify the complexity of binary image classification on various datasets by employing the proposed approach. To that end, we define the query set Q as the set of all possible 3 x 3 patch locations in a 28 x 28 image. Each patch query could be seen as a semantic question regarding pixel intensities in a particular region of the image. A natural measure of the semantic complexity for these tasks is the number of patch queries required to predict the class label. Given an image x, the IP Encoder chooses a sequence of queries/patches such that the decoder can accurately predict the corresponding class label y by observing x only through these patches.

In our experiments we considered four datasets; MNIST, Fashion-MNIST, K-MNIST and Caltech Silhouettes, where we first binarized the images as a pre-processing step. The primary motivation behind working solely with binary images was to make these tasks comparable since the patch queries are of the same complexity. Figure 1 illustrates sample images from each of the datasets considered and the corresponding semantic complexity as determined by the proposed theory. We note that our results are consistent with the intuition that semantically complex datasets require a higher number of queries to be resolved (predict label y from data x) and are correspondingly harder to solve via other approaches (such as neural networks).

A byproduct of the above framework is that one can also compute the most informative parts of an image for image classification. Figure 2 elucidates how IP is used to unveil most informative regions of a given image. Figure 2a shows the input image x and the sequence of patch queries the encoder asks about x. Figure 2b shows all the answers to the queries the decoder takes as input to predict the class label. Figure 2c displays how the posterior over the class labels changes as more and more evidence (answers to patch queries) is curated by the encoder.



	MNIST	K-MNIST	Fashion-MNIST	Caltech Silhouettes
$SE_Q(X;Y)$ (approx.)	11.54	27.39	38.73	70.27
CNN Test Accuracy	99.3	95.1	86.96	65.15

Figure 1. *Semantic entropy for image classification tasks for several binary image datasets. The results conform with intuition of more complex datasets having higher semantic entropy. For instance, Caltech Silhouettes, a dataset of binarized images of 101 classes from the Caltech dataset is obviously semantically more complex than handwritten digits in the MNIST dataset. A dedicated CNN trained on each of these datasets has a test accuracy that correlates negatively with the semantic entropy that further validates our hypothesis that tasks with higher semantic entropy are harder to learn. Since IP was used to compute $SE_Q(X; Y)$ it is only an approximation.*

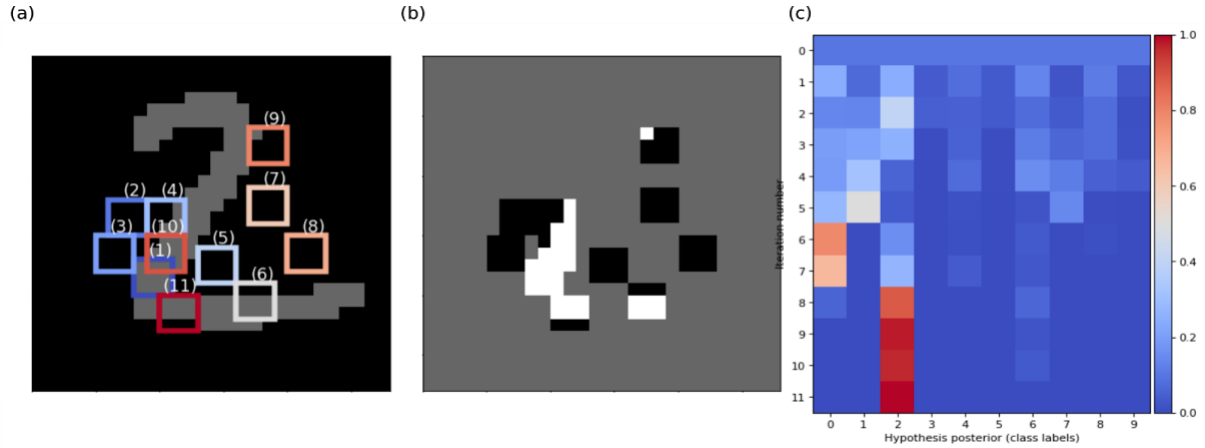


Figure 2. *Using IP to quantify semantic information content for the task of classifying images of handwritten digits.* **(a)** *The square boxes denote the patch queries asked by the encoder upon receiving image x as input. These boxes are numbered according to the order in which they were queried during IP.* **(b)** *The binary-valued patches indicate "answers" to the sequence of questions, where white means "yes" and black means "no". The gray background illustrates the parts of the image that were not looked at by the decoder.* **(c)** *Each entry (i,j) of the matrix represents the posterior probability of the j^{th} class given the first i queries (patches). Notice that as the IP iterations proceed, the class posterior becomes more peaked and after 11 queries IP determines that the class label is "two".*

Evaluation of Algorithms for Recognizing Activities in Long Extended Videos

Events in natural videos typically arise from spatio-temporal interactions between actors and objects and involve multiple co-occurring activities and object classes. To capture this rich visual and semantic context, in [2] we proposed Visual-Symbolic Spatio-Temporal Message Passing Neural Networks (VS-ST-MPNN), a new framework for representation learning on visual-symbolic graphs. Figure 3 shows

our model uses (1) an attributed spatio-temporal visual graph whose nodes correspond to actors and objects and whose edges encode different types of interactions, and (2) a symbolic graph that models semantic relationships. VS-ST-MPNN further uses a graph neural network to refine the representations of actors, objects and their interactions on the resulting hybrid graph. Our model goes beyond current approaches that assume nodes and edges are of the same type, operate on graphs with fixed edge weights and do not use a symbolic graph. In particular, VS-ST-MPNN: a) has specialized attention-based message functions for different node and edge types; b) uses visual edge features; c) integrates visual evidence with label relationships; and d) performs global reasoning in the semantic space.

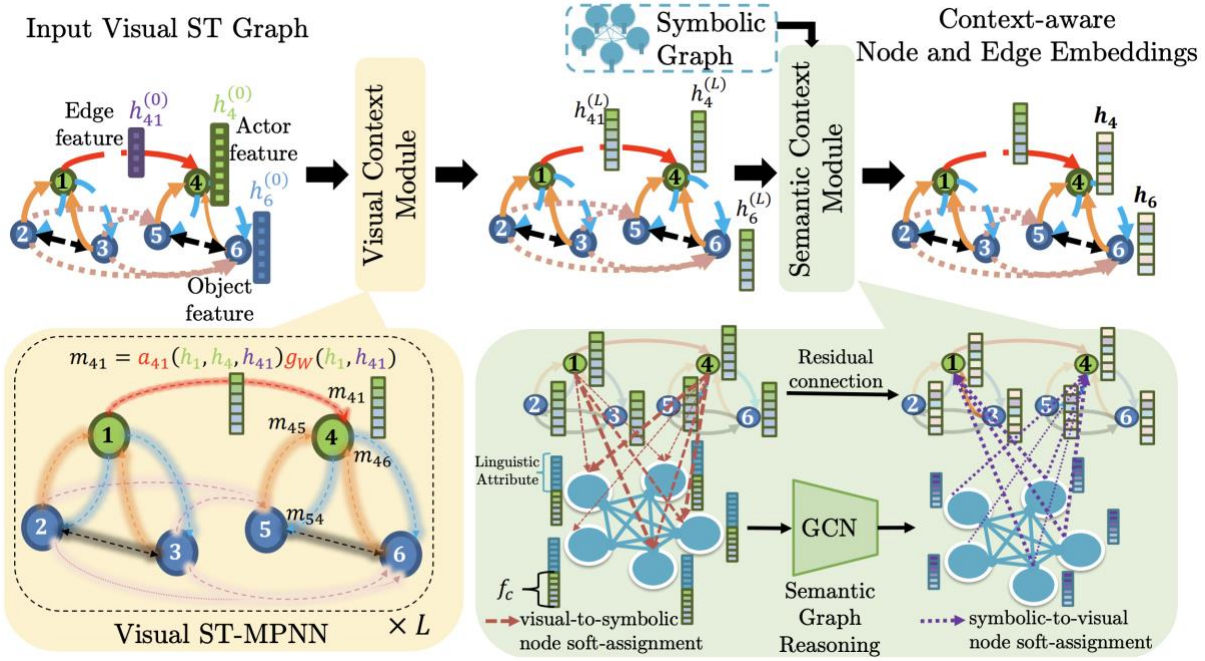


Figure 3. *Overview of our VS-ST-MPNN model that performs representation learning on a hybrid visual-symbolic graph. Given an input video that is represented as a visual st-graph, with nodes corresponding to detected actors and objects and edges capturing latent interactions, our framework has two modules that integrate context in the local representations of its nodes and edges: (a) a Visual Context Module that updates the visual graph via specialized neighborhood aggregation functions and (b) a Semantic Context Module that integrates visual evidence with semantic knowledge encoded in an external symbolic graph and learns global semantic interaction-aware features.*

Table 1. *Results on CAD-120 [27] for sub-activity and object affordance detection, measured via F1-score. Our results are averaged over five random runs, with the standard deviation reported in parentheses.*

Method	Detection F1-score (%)	
	Sub-activity	Object affordance
ATCRF [27]	80.4	81.5
S-RNN [21]	83.2	88.7
S-RNN [21] (multitask)	82.4	**91.1**
GPNN [47]	88.9	88.8
STGCN [12]	88.5	-
VS-ST-MPNN (ours)	**90.4** (±0.8)	*89.2* (±0.3)
only visual graph (ours)	89.6 (±1.1)	88.6 (±0.6)

Table 1 compares the sub-activity and affordance detection performance of our method with prior work. Our method obtains state-of-the-art results for sub-activity detection, with an average performance of 90.4% and a best of 91.3%, and the second best result on affordance detection (89.2%) - being only second to the S-RNN (multi-task). The S-RNN was trained on the joint task of detection and anticipation and we outperform it by 8% in the sub-activity classification task. Even without using the symbolic graph, our method improves upon recent GNNs, which were applied on the same attributed visual st-graph, validating our novel layer propagation rules.

In work published at WACV 2022 [2], we extended our VS-ST-MPNN model to spatio-temporal action localization in extended videos. As shown in Figure 4, the proposed framework has two main components: (a) actor-centric tubelet generation and (b) temporal multi-label action recognition per tubelet. The first component handles the spatial localization of activities. It generates spatio-temporal tubelets of interest, which are associated with a single primary actor (person or vehicle) and capture all the relevant spatio-temporal visual context (scene cues, interacting objects, etc.). Our tubelet generator utilizes actor detections obtained with an off-the-shelf object detector, such as Faster R-CNN, which are then tracked over time using an off-the-shelf tracker, such as SORT, to generate actor tracklets. The bounding boxes of each actor tracklet are extended in space to capture relevant visual context, as defined by hand-crafted rules. For instance, the visual context of a person includes all other close-by people and static vehicles. The second component predicts the activities performed by an actor over time based on local motion cues (optical flow) and spatio-temporal actor-object interactions. Such interactions are modeled with a visual spatio-temporal graph, whose nodes correspond to detected actors and objects. Node attributes are local appearance/motion features extracted from a spatio-temporal CNN, such as an I3D network, while edge attributes capture geometric relations. Our novel VS-ST-MPNN model performs representation learning on that graph with the goal of refining local actor node and edge attributes using visual contextual cues. These refined features are then fed to action classifiers that yield activity scores for each time step. The time series of scores are converted to action detections with start/end frames and confidence scores using a peak-finding approach. Finally, activity detections from all tubelets are aggregated to generate the output set of activity detections for the input video.

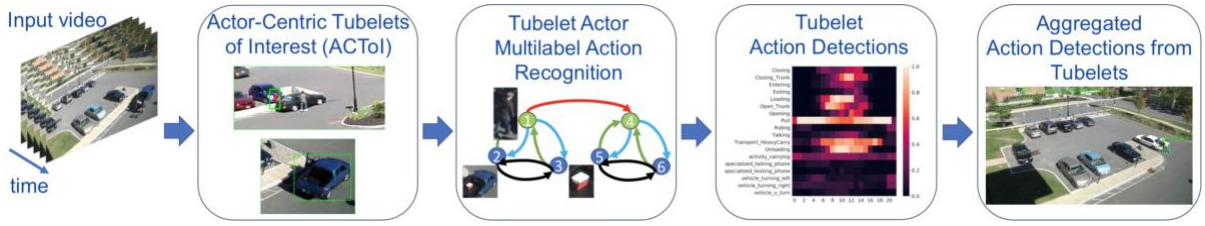


Figure 4. *Overview of proposed framework for spatio- temporal action localization in extended videos.*

Table 2. *Temporal action detection performance on two MEVA validation sets measured using the ActEV scorer evaluation tool (ActEV SDL V1).*

Metric	UCF val	Kitware val
nAUDC@0.2tfa	53.3%	50.6%
pmiss@0.04tfa	59.0%	53.7%

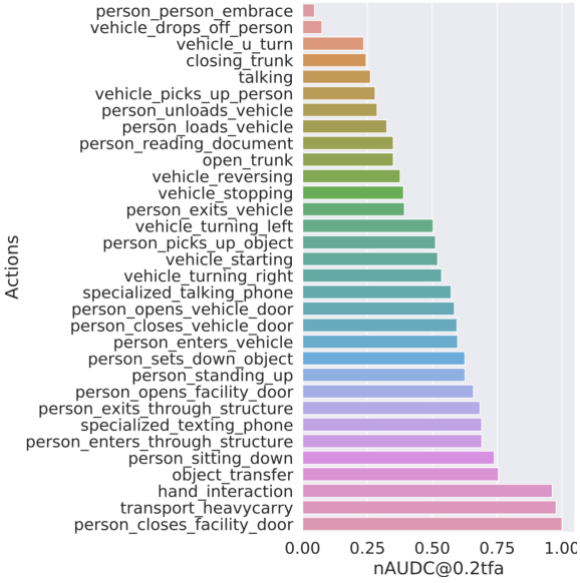


Figure 5. *Activity detection performance on MEVA Kit- ware validation set*

We train our system using 251 videos from the MEVA dataset and evaluate on two custom validation sets: UCF validation set with 68 videos and Kitware validation set with 123 videos. Each ground truth activity instance in the training set is annotated by human annotators with its start/end frame and a sparse sequence of actor bounding boxes. Preliminary results are summarized in Table 2. These results are obtained by using the I3D network as a feature extractor, which was trained on the VIRAT training set, and with the ST-MPNN graph neural network trained only on MEVA. Our approach achieves a competitive nAUDC@0.2tfa = 51%, despite being trained with only 251 videos and using a backbone I3D trained on VIRAT. Figure 5 analyzes the performance per activity class on the Kitware validation set. Possible directions for future work towards improving the action detection performance on MEVA are: (a) finetuning an I3D network on MEVA data, (b) using more training data, and (c) improving tubelet recall by improving actor detection and tracking.

TRAINING OPPORTUNITIES AND COLLABORATIONS

This project has trained 1 postdoc, 2 PhD students and 1 MSc student at the intersection of information theory, statistics, machine learning, optimization, and computer vision. The trainees have been exposed to a number of topics through various collaborations both within each group, as well as across different groups. For example, a collaboration between Bober's, Kittler's and Vidal's group led to exploiting the work of (França, Robinson and Vidal, 2018b) on accelerated optimization algorithms for constrained problems to accelerate algorithms for visual object tracking (Xu et al., 2020). The proposed tracking approach is the winner of the multimodal video tracking (visible and infrared) challenge VOT 2020 (ECCV2020).

RESULTS DISSEMINATION

Prof. Vidal organized SIAM Mathematics of Data Science 2020 Mini-Symposium "Advances in Subspace Learning and Clustering Mini-Symposium." See https://www.minds.jhu.edu/siam-mds20-ms13-advances-in-subspace-learning-and-clustering-mini-symposium/ for details.

Prof. Vidal gave several presentations, including plenary lectures at CAMSAP 2019, Medical Imaging meets NeurIPS workshop 2019, Math+X Symposium 2020, Deep Image Analysis Workshop at ISBI 2020, and invited talks at Mathematical and Computational Research in Data Science 2020.

HONORS AND AWARDS

René Vidal was Elected Fellow of the American Institute for Medical and Biological Engineering (AIMBE), 2020

PUBLICATIONS

[1] A. Chattopadhyay, S. Slocum, B. Haeffele, R. Vidal, and D. Geman. Interpretable by Design: Learning Predictors by Composing Interpretable Queries. IEEE Transactions in Pattern Analysis and Machine Intelligence, 2022.

[2] E. Mavroudi, B. Haro, and R.Vidal. Representation learning on visual-symbolic graphs for video understanding. European Conference on Computer Vision, 2020.

[3] E Mavroudi, P Bindal, R Vidal. Actor-Centric Tubelets for Real-Time Activity Detection in Extended Videos. IEEE/CVF Winter Conference on Applications of Computer Vision, 2022.