MAHDI M. KALAYEH

1987	Born in Duisburg, Germany.
2009	B.S., Tehran Polytechnic, Tehran, Iran.
2010	M.S., Illinois Institute of Technology, Chicago, Illinois.
2011	Visiting Researcher, Tehran Polytechnic, Tehran, Iran.
2012-19	Ph.D., University of Central Florida, Orlando, Florida.

SELECTED AWARDS

2017	CVPR Doctoral Consortium Travel Award
2015	ACM Multimedia Travel Award

TALKS

2018	*Describing Images by Semantic Modeling of Local Represen tations,* **MIT-IBM Watson AI Lab**, Cambridge, MA.
2019	*Training Faster by Separating Modes of Variation in Batch-normalized Models,* **Google AI Research**, Mountain view, CA.
2019	*Training Faster by Separating Modes of Variation in Batch-normalized Models,* International Computer Science Institute (ICSI), **University of California, Berkeley**.

UCF | Center for Research in Computer Vision

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

MAHDI M. KALAYEH

B.S., AMIRKABIR UNIVERSITY OF TECHNOLOGY, 2009
M.S., ILLINOIS INSTITUTE OF TECHNOLOGY, 2010

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

26, March, 2019, 11:00 A.M.
HEC 101

DISSERTATION COMMITTEE

Professor Mubarak Shah, *Chairman*
Professor Gita Sukthankar
Professor Nazanin Rahnavard
Professor Teng Zhang

DISSERTATION RESEARCH IMPACT

An extremely large number of images is being created at this very moment by people like us from all around the world. And this only adds to what has been aggregated and preserved so far. Hence, it is very much expected to ask how are we going to represent, organize and search through our images? The scenario can be as simple as finding a particular photo on our cellphones from a backpacking trip, which we enthusiastically want to show it to a friend. Given the scale of the problem, we have no choice but to design algorithmic techniques which effectively harness the modern computation power. Meanwhile, to shrink the semantic gap and moving towards more seamless and natural human-computer interaction, we should prioritize models that provide a semantic description of images. Such perspective facilitates our understanding of how machines infer visual content and paves the way for explainability of well-performing but complex models.

This dissertation contributes to semantic modeling of images by proposing: (1) an automatic image annotation framework that assigns relevant textual tags (*e.g* "car", "building", "sunny", "old") to images, allowing for natural content-based search through a large and continuously growing collection of images, (2) effective person-related attribute prediction models that robustly detect the appearance of fine-grained visual traits (*e.g* "big lips", "mouth slightly open", "wearing long sleeves") in human images, (3) the first large-scale Selfie dataset, and exploring the limits of current algorithms to automatically analyze the popularity and sentiment of Selfies, which very well reflect the mood, preferences and interests of our society, (4) a simple yet effective approach to accelerate training of deep convolutional neural networks, as the backbone of all today's computer vision frameworks.

SELECTED PUBLICATIONS (h-index: 8, total citation: 300)

1. **Training Faster by Separating Modes of Variation in Batch-normalized Models,** M. M. Kalayeh and M. Shah, *in IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), 2019.*
2. **Human Semantic Parsing for Person Re-identification,** M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, *in IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**),* 2018.
3. **Improving Facial Attribute Prediction using Semantic Segmentation,** M. M. Kalayeh, B. Gong, and M. Shah, *in IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**),* 2017.
4. **How to Take a Good Selfie?,** M. M. Kalayeh, M. Seifu, W. LaLanne, and M. Shah, *in 23rd ACM International Conference on Multimedia (**ACM MM**),* 2015.
5. **Recognition of Complex Events: Exploiting Temporal Dynamics between Underlying Concepts,** S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, and M. Shah, *in IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**),* 2014.
6. **NMF-KNN: Image Annotation using Weighted Multi-view Non-negative Matrix Factorization,** M. M. Kalayeh, H. Idrees, and M. Shah, *in IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**),* 2014.

DISSERTATION

DESCRIBING IMAGES BY SEMANTIC MODELING USING ATTRIBUTES AND TAGS

This dissertation addresses the problem of semantically describing images, a fundamental task that narrows down the semantic gap between visual reasoning of humans and machines. Visual attributes and textual tags allow us to naturally characterize objects (*e.g* "person" and "makeup"), actions (*e.g* "wearing"), and relationships (*e.g* "person wearing makeup") both individually, grounded to the local properties, and in the global context of the entire scene.

Automatic image annotation assigns relevant textual tags to the images. We propose a mathematical framework based on Non-negative Matrix Factorization to perform automatic image annotation such that the proposed technique can seamlessly adapt to the continuous growth of datasets. Our proposed query-specific approach is built on the features of nearest-neighbors and tags. It naturally solves the problem of feature fusion and handles the challenge of rare tags by introducing weight matrices that penalize for incorrect modeling of less frequent tags and images that are associated with them.

Despite their effectiveness, the descriptive power of tags scales linearly with the number of unique training labels. In contrast, attributes being category-agnostic, allow an exponential number of semantic classes to be modeled. Due to above superiority, next, we focus on visual attributes. We hypothesize that integrating pixel-level semantic parsing of the face and human body should improve person-related attribute prediction. In this regard, we propose Semantic Segmentation-based Pooling (SSP) and Gating (SSG). In SSP, the estimated segmentation masks pool the output activations of the last (before classifier) convolutional layer at multiple semantically homogeneous regions, unlike global average pooling that is spatially agnostic. In SSG, we create multiple copies where each preserves the activations within a single semantic region and suppresses otherwise. This mechanism prevents max-pooling from mixing semantically inconsistent regions. SSP and SSG while effective, impose heavy memory utilization. To tackle that, Symbiotic Augmentation (SA) is proposed, where we *learn to generate* only one mask per activation channel.

The massive number of self-portrait images shared on social media is revolutionizing the way people introduce themselves to the world. Due to the Big Data nature of Selfies, it is nearly impossible to analyze them manually. Next, we use both *textual tags* and *visual attributes* to analyze Selfies. We collect the first Selfie dataset with more than 46K images and annotate it with 36 visual attributes covering characteristics such as gender, age, race, and hairstyle. We provide attribute prediction of Selfies, using SIFT and HOG, pre-trained AlexNet on ImageNet, and Adjective Noun Pairs (*e.g "smiling boy"*) of SentiBank. We train l_2-regularized SVR for the log_2-normalized view counts in order to assess the impact of different visual concepts and various Instagram filters on the popularity of Selfies.

Almost all today's deep convolutional neural architectures, including those that we propose in this dissertation, use Batch Normalization (BN), yet the characteristics of BN are not sufficiently studied in the literature. We conclude this dissertation by showing that assuming samples within a mini-batch are from the same probability density function, then BN is identical to the Fisher vector of a Gaussian distribution. That means batch normalizing transform can be explained in terms of kernels that naturally emerge from the probability density function that models the generative process of the underlying data distribution. Specifically, we theoretically demonstrate how BN can be improved by disentangling modes of variation in the underlying distribution of layer outputs. An exten-sive set of experiments confirms that our proposed alternative to BN, Mixture Normaliza-tion, not only effectively accelerates training of different batch-normalized architectures including Inception-V3, DenseNet, and DCGAN, but also achieves better generalization error.