Learning Compositional Sparse Bimodal Models

Suren Kumar, *Member, IEEE*, Vikas Dhiman, Parker A Koch, and Jason J. Corso, *Senior Member, IEEE*

Abstract—Various perceptual domains have underlying compositional semantics that are rarely captured in current models. We suspect this is because directly learning the compositional structure has evaded these models. Yet, the compositional structure of a given domain can be grounded in a separate domain thereby simplifying its learning. To that end, we propose a new approach to modeling bimodal perceptual domains that explicitly relates distinct projections across each modality and then jointly learns a bimodal sparse representation. The resulting model enables compositionality across these distinct projections and hence can generalize to unobserved percepts spanned by this compositional basis. For example, our model can be trained on *red triangles* and *blue squares*; yet, implicitly will also have learned *red squares* and *blue triangles*. The structure of the projections and hence the compositional basis is learned automatically; no assumption is made on the ordering of the compositional elements in either modality. Although our modeling paradigm is general, we explicitly focus on a tabletop building-blocks setting. To test our model, we have acquired a new bimodal dataset comprising images and spoken utterances of colored shapes (blocks) in the tabletop setting. Our experiments demonstrate the benefits of explicitly leveraging compositionality in both quantitative and human evaluation studies.

Index Terms—Multimodal learning, compositional learning, symbol grounding, artificial intelligence, tabletop robotics, human-robot interaction

1 Introduction

Consider a robot that can manipulate small building-blocks in a tabletop workspace, as in Fig. 1. Task this robot with following vocal human utterances that guide the construction of non-trivial building-block structures, such as *place an orange rectangle on top of the blue tower to the right of the green tower*. This experimental setup, although contrived, is non-trivial even for state-of-the-art frameworks in perception, artificial intelligence and robotics. The robot must be able to interpret the spoken language (audio perception); segment individual structures, *orange rectangle*, (visual perception); it must be able to reason about collections of structures, *blue tower*, (physical modeling); it must be able to relate such collections, *to the right of*, (linguistics); and it must be able to execute the action, *place*, (manipulation). These challenging points are underscored by the frequency of tabletop manipulation as the experimental paradigm in many recent papers in our community, e.g., [1], [2].

To achieve success in this experimental setup, the robot would need to satisfy Jackendoff's Cognitive Constraint [3], or at least a robotic interpretation thereof. Namely, there must exist a certain representation that relates percepts to language and language to percepts because otherwise the robotic system would neither be able to understand its visual-linguistic percepts nor execute its tasks. This and similar cognitive-semantic theories have led to symbol-grounding [4] and language-grounding [5], [6].

Most existing work in symbol- and language-grounding has emphasized tying visual evidence to known language [7], [8], learning language models in the context of navigation [6] and manipulation tasks [9], [10] for a fixed set of perceptual phenomena, and even language generation from images and video [11], [12], [13]. Considering a pre-existing language or fixed set of percepts limits the generality of these prior works. Recently, one method has enabled joint perceptual-lingual adaptation to novel input [2], but it does not explicitly capture the compositional nature of language, leading to scalability challenges.

Indeed, the majority of works in language grounding do not exploit the compositional nature of language despite the potential in doing so [14]. One major limitation of the paired (non-compositional) representation is the resulting overwhelming learning problem. Take, the example of *orange rectangle* and *green tower* from earlier. The adjectives *orange* and *green* are invariant to the objects *rectangle* and *tower*. Compositional representations exploit this invariance whereas paired ones have combinatorial growth in the size of the learning problem.

In this paper, we exploit the compositional nature of language to represent bimodal visual-audial percepts describing tabletop scenes similar to those in our example (Fig. 1). However, we do not directly learn the compositional structure of these percepts—attempts at doing so have met with limited success in the literature, given the challenge of the structure inference problem [15], [16]. Instead, we ground the bimodal representation in a language-based compositional model. We fix a two-part structure wherein groupings of visual features are mapped to audio segments.

The specific mapping is not hand-tuned. Instead, it is automatically learned from the data, and all elements of the

• *The authors are with Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109. E-mail: {surenkum, dhiman, pakoch, jjcorso}@umich.edu.*

Manuscript received 15 Mar. 2016; revised 9 Feb. 2017; accepted 3 Apr. 2017. Date of publication 11 Apr. 2017; date of current version 10 Apr. 2018. Recommended for acceptance by T. Darell, C. Lampert, N. Sebe, Y. Wu, and Y. Yan. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TPAMI.2017.2693987

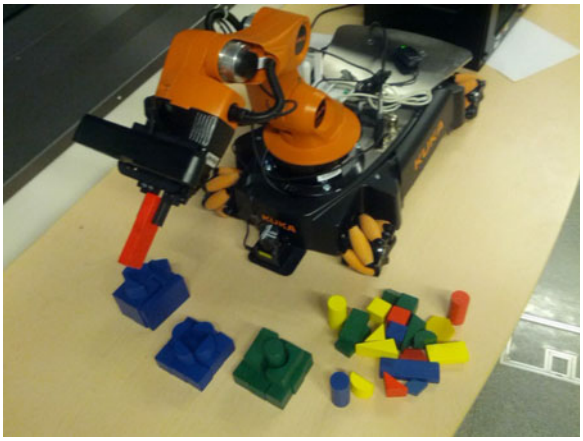


Fig. 1. Our overarching goal is to improve human-robot/robot-robot interaction across sensing modalities while aiming at generalization ability. Multimodal compositional models are important for effective interactions.

compositional model are learned jointly. This two-part compositional structure can take the form of adjective-noun, e.g., *orange rectangle*, or even noun-adjective; the method is agnostic to the specific form of the structure. An earlier version of our work required the specific form of the structure to be consistent across the entire dataset [17]; however, in this paper, we relax this requirement and propose a more general approach that allows for this structure to vary per sample. The structure is hence fully induced by the data itself (what is inherent in the spoken language). We also add a probabilistic interpretation of the proposed model and prove its convergence.

The specific representation we use is a sparse representation as it allows interpretability because the signal is represented by few bases while minimizing a goodness of fit measure. There is increasing physiological evidence that humans use sparse coding in representation of various sensory inputs [18], [19], [20]. The need for sparse coding is supported by the hypothesis of using least energy in neuron's excitation to represent input sensory data. Furthermore, evidence suggests that multi-modal sensory data is projected together on a common basis [21], like we do in our compositional model.

We have implemented our compositional sparse model learning for bimodal percepts in a tabletop experiment setting with real data. We observe a strong ability to learn the model from fully observed examples, i.e., the training set consists of all classes to be tested. More significantly, we observe a similarly strong ability to generalize to unseen but partially observed examples, i.e., the testing set contains classes for whom only partial features are seen in the training set. For example, we train on *blue square* and *red triangle* and we test on *blue triangle* and *red square*. Furthermore, this generalization ability is not observed in the state-of-the-art joint baseline model we compare against.

Contributions. The underlying problem being addressed in the current paper is to discover the underlying semantics given N data samples of the form $\{x_i, y_i\}, i \in [1, N]$, where x_i and y_i is the data from two different sensing modalities such as audial and vision. Furthermore, we want to be able to obtain a reversible mapping between individual sensing modalities, i.e., learning an invertible mapping $x_i = \phi(y_i)$. This work can be extended to multiple modalities but for the clarity of exposition we will focus on only two different modalities. There are two major contribution of our work.

First, our work enables representing the relationship between various modalities that belong to different feature spaces. This is similar to low rank structure based multi-task learning (MTL) where the multiple task are the representation of semantics in various modalities given the data of the form $\{\{x_i, c_i\}, \{y_i, c_i\}\}, i \in [1, N], c_i \in \mathcal{C}$, where $\mathcal{C}$ is a set of semantic labels [22]. However, our method neither needs explicit semantic labels nor requires the two different modalities $\{x_i, y_i\}$ to lie in same feature space [23], making it significantly more general. Previous work has explored the application of sparse multitask learning method to multimodal data where different modalities lie in different feature space for classification [24], [25]. To the best of our knowledge, we are not aware of any generative MTL method to approximate an invertible mapping between feature spaces.

We compare performance to multi-task learning on the current problem (learning an invertible mapping ϕ), by regressing over each dimension of y as an individual task. We show that MTL approximates the distribution $P(y|x)$ where as the proposed model approximates the joint distribution $P(x, y)$ allowing our model to learn an invertible mapping between x and y ($P(y|x)$ and $P(x|y)$). We demonstrate that our method improves retrieval accuracy (proxy to test for generative mapping) over the MTL approach. Furthermore, we use the independence between parts of feature space to better approximate the joint distribution $P(x, y)$ by using the compositionality in language. Please note that we use the semantic labels c_i only as a proxy to evaluate the performance of mapping. The invertible mapping between various modalities is essential to achieve the generalization performance demonstrated in Section 4.3.

The second major contribution of this paper is to use compositionality to reduce the number of class labels. Let us assume that our semantic space consists of two different semantic entities, $\mathcal{C} = \mathcal{N} \times \mathcal{S}$. The total number of tasks to solve will be the outer product of these two different semantic spaces ($|\mathcal{C}| = |\mathcal{N}||\mathcal{S}|$). However, our compositional framework can exploit the compositionality and convert the problem to only requiring $|\mathcal{N}| + |\mathcal{S}|$ number of tasks.

This rest of this paper is organized as follows. Section 2 introduces both the paired and compositional sparse models in context of using two different modalities. We then describe the feature representation of both modalities, vision and audio, in Section 3. Section 4 demonstrates the experimentation validation of our models along with visualization to provide more insight into the structure being represented by our models. We also show an extension of our model to perform visually grounded translation between two different languages mediated by the proposed sparse learning models. Finally, we conclude and provide some directions for future research in Section 5.

2 Model

We begin the description of the theoretical model with a brief background in sparse coding. Then, we introduce the MTL approach for the current problem. Next, we describe the bimodal paired sparse model, which learns a sparse basis jointly over specific visual and audial modalities. This

paired sparse model is not new [26], [27], [28]; it is the fabric of our compositional sparse model, which we describe in section. The novel compositional sparse model jointly learns a mapping between certain feature/segment subsets, which then comprise the compositional parts to our model, and the paired sparse model for each of them. We provide a consistent probabilistic interpretation of all of these approaches and prove the convergence of the compositional sparse learning method.

2.1 Dictionary Learning

Dictionary learning is an important signal representation method that reconstructs a signal using a sparse linear combination of bases, which constitute a dictionary. Given N data samples from a m-dimensional real space, $x_i, x_i \in \mathcal{R}^m$, $i \in [1, N]$, the dictionary learning step involves finding an overcomplete set of basis vectors $U \in \mathcal{R}^{m \times q}$, where the number of basis vector is much larger than the dimension of the space $q >> m$. Sparse coding uses the learned dictionary to estimate the coefficient vector α_i in order to reconstruct the input signal,

$$x_i = U\alpha_i + \epsilon, \tag{1}$$

as a linear combination of the dictionary basis/atoms with an additive noise. Assuming the additive noise ϵ to be Gaussian with zero mean and σ^2 variance, we can write the data likelihood as

$$P(x_i|U, \alpha_i) \propto \exp\left(\frac{-1}{2\sigma^2}||x_i - U\alpha_i||_2^2\right). \tag{2}$$

However, the sparse coding step is an underdetermined linear system. As a consequence, sparsity on the reconstruction vector is used in order to obtain a computationally efficient and representationally meaningful solution. This is imposed via a zero mean Laplace prior on the reconstruction vector as $P(\alpha_i) \propto \exp(-\lambda||\alpha_i||_1)$. Given the probabilistic model in Equation (2) along with the Laplace prior, the posterior probability for α can be written as

$$P(\alpha_i|x_i, U) = P(x_i|U, \alpha_i)P(\alpha_i). \tag{3}$$

Minimizing the negative log posterior and subsuming the $\frac{1}{2\sigma^2}$ within the λ parameter, we can state the sparse coding step as

$$\arg\min_{\alpha_i} ||x_i - U\alpha_i||_2^2 + \lambda||\alpha_i||_1. \tag{4}$$

Since, we may not have access to the dictionary U, it is usually learned jointly along with the sparse coding step [29]. The general dictionary learning optimization problem can be stated as

$$\begin{aligned} &\arg\min_{U,\{\alpha\}} \sum_{i=1}^{N} ||x_i - U\alpha_i||_2^2 \\ &\text{s.t.} \quad ||\alpha_i||_p \leq \lambda, \forall \ \ i \in [1, N]. \end{aligned} \tag{5}$$

The sparsity inducing L_p norms are used as a constraint on the reconstruction vector. Dictionary learning is NP-hard and imposing sparsity constraints, such as the L_0 norm on α_i, further complicates the optimization landscape. Hence, as a workaround, L_1 norm (Laplace prior on reconstruction vector) is imposed [30]. Various open source libraries, such as sparse modeling software (SPAMS) have been proposed to solve both the dictionary learning and sparse coding problem [31]. To summarize, dictionary learning and sparse coding is an unsupervised model to approximate the distribution $P(x)$ by using a sparsity-inducing prior on the reconstruction vector.

2.2 Multi-Task Learning

Multi-task learning seeks to improve the performance of a machine learning problem by using inductive bias derived from other related problems [32]. Inductive bias is derived by learning various associated problems with the main machine learning task via a shared representation. Consider, m different tasks with training data of the form $(\mathbf{x_1}, \mathbf{a_1})$, $(\mathbf{x_2}, \mathbf{a_2}), \ldots, (\mathbf{x_m}, \mathbf{a_m})$, where the $(\mathbf{x_j}, \mathbf{a_j})$ represents the collection of training samples with each sample consisting of a d dimensional feature, $x_i \in \mathbb{R}^d$ and a corresponding real target, $a_i \in \mathbb{R}$ for the task j. The learning consists of determining a weight matrix $W = [w_1, w_2, \ldots, w_m], W \in \mathbb{R}^{d \times m}$, whose each column serves as a linear predictor $f_j(x_i) = w_j^T x_i$ for the task j. The overall optimization problem can be written as

$$\min_W \mathcal{L}(W) + \Omega(W), \tag{6}$$

where $\mathcal{L}(W)$ is the error metric between the linear predictor and targets of each task summed over the entire training data. $\Omega(W)$ represents the prior knowledge over the tasks and serves as a regularization. With a uniform prior over $\Omega(W)$, the optimization problem 6 decomposes to m different single task learning.

Argyriou et al. [33] use a $l_{2,1}$ matrix norm ($\Omega(W) = ||W||_{2,1}^2$) which corresponds to selecting a few rows (feature set) out of d rows for all the tasks. The $l_{2,1}$ matrix norm corresponds to a Laplace prior over the rows of the matrix. Assuming a Gaussian error on each task predictor, it can be shown that multi-task learning problem corresponds to modeling the distribution $P(\mathbf{A}|x)$ by using a sparsity inducing prior over the weight matrix W [34]. $\mathbf{A} \in \mathbb{R}$ is the space of all the tasks. Recent efforts by Yan et al. [35] have started using dictionary learning in a multi-task learning framework by additionally learning a shared sub-space across all the tasks. For the experiments in the current paper, we use the publicly available MALSAR [36] package.

For the comparison in the current paper, we treat the individual dimension of a modality y in learning to map a modality x to y ($\phi(x) = y$) as the individual task. Hence $\mathbf{A} \equiv y$ and the MTL problem learns the conditional distribution $P(\mathbf{A}|x)$. However, MTL can only learn a forward mapping and as a consequence the inverse mapping ϕ^{-1} is not available. Furthermore, MTL ignores the distribution of feature space x and requires uniform feature dimension for all the m different tasks. We overcome these limitations by developing a unsupervised paired sparse model to learn invertible mappings while operating over tasks with different dimensions.

2.3 Paired Sparse Model

Paired sparse modeling is driven by two findings from neurobiology [18], [19], [20], [21]: i) sparsity in representations

and ii) various modality inputs are directly related. We hence use paired dictionary learning in which individual sensory data is represented by a sparse basis and the resulting representation shares coefficients across those bases. We are inspired by the success of paired dictionary learning in visualizing images from features [28], cross-style image synthesis, image super-resolution [26], [27] and beyond.

We adapt paired dictionary learning to our problem by learning over-complete dictionaries for sparse bases in both the visual and audial domain while using the same coefficients across domain-bases. Following similar notation to [28], let x_i, y_i represent visual and audio features for the ith sample. These are related by the function mapping, $x_i = \phi(y_i)$. We seek to estimate forward (ϕ) and inverse (ϕ^{-1}) mappings while representing the audio and visual features with over-complete dictionaries (bases) U and V, respectively, coupled by a common sparse coefficient vector α

$$x_i = U\alpha_i \quad \text{and} \quad y_i = V\alpha_i. \tag{7}$$

Sparsity in the coefficient vector is enforced by an l_1 metric [37] as $||\alpha||_1 \leq \lambda$. This ensures that only few bases are actively used for representing a particular input. For a given training dataset of size N, the over-complete dictionaries U and V, and the sparse coefficient vectors $\{\alpha\}_i$ are jointly estimated by minimizing the l_2 norm of the reconstruction error in both bases

$$\begin{aligned} &\arg\min_{U,V,\alpha} \sum_{i=1}^{N} \left(\|x_i - U\alpha_i\|_2^2 + \|y_i - V\alpha_i\|_2^2\right) \\ &\text{s.t.} \quad \|\alpha_i\|_1 \leq \lambda \; \forall i, \|U\|_2 \leq 1, \|V\|_2 \leq 1. \end{aligned} \tag{8}$$

Note that the bases of the over-complete dictionaries are further constrained to belong to a convex set such that individual bases have l_2 norm less than or equal to unity.

The inverse mapping ϕ^{-1} for a novel sample is found by first projecting y on the learned dictionary V and then using the obtained coefficients α^* to compute $x = U\alpha^*$. The process of finding these coefficients involves the following optimization problem

$$\alpha^* = \arg\min_{\alpha} \|V\alpha - y\|_2^2 \text{ s.t } \|\alpha\|_1 \leq \lambda. \tag{9}$$

Similarly one can obtain the forward mapping by first projecting x on learned dictionary U to obtain α^* and then using the learned dictionary V to obtain y. Thus, the estimation of forward and inverse mappings gives one the ability to go from audial features to visual features and vice versa. In contrast to MTL dictionary learning [35], the paired dictionary learning does not requires various features to share a common subspace.

Paired dictionary model requires the common sparse representation α_i to reconstruct both x_i and y_i. The error objective in Equation (8) represents the dictionary learning over the concatenated spaces,

$$\overline{x} = \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad \text{and} \quad \overline{U} = \begin{bmatrix} U \\ V \end{bmatrix}. \tag{10}$$

Following the analysis in Section 2.1, we are modeling the joint distribution $P(x, y)$ by assuming a Laplace prior on the shared sparse representation vector α.

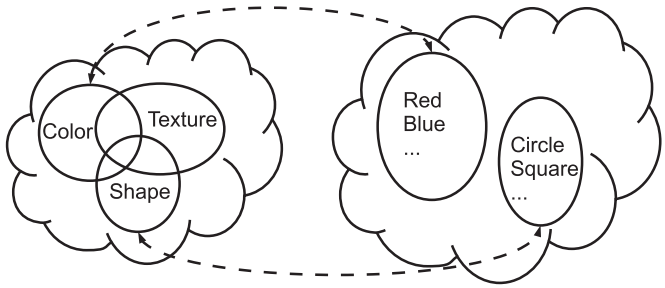


Fig. 2. Mapping the physical concepts from visual domain such as color, texture and shape to the spoken language domain.

We use the online dictionary learning method which alternates between the dictionary update and sparse coding steps [38]. The sparse coding step is performed by using least angle regression (LARS) [39] and the dictionary update is performed via updating each dictionary element to minimize the reconstruction error. We use the open source sparse modeling software [31] for solving all the optimization problems in this paper.

2.4 Compositional Sparse Model

The paired model can link the two domains, but it can not exploit the compositionality inherently present in the language. Consider, again, the utterance *red square*. The part *red* describes the color of the object and the part *square* describes the shape of the object. The two parts are captured by distinctive and co-invariant visual features. We can hence explicitly map individual percepts between domains. Fig. 2 illustrates the kind of mappings we expect to obtain between physically grounded concepts from the visual and audial (spoken human language) domain. We refer to physically observable properties, such as color, texture and shape of an object as "concepts".

Observable high-level visual characteristics are referred to as "attributes" in the computer vision literature [40], [41], [42]. Object attributes describe the semantic content present in an image. Consider, for example, an image of an elephant and a lamb. The object-level attributes "is big", "is gray" can be applied to the elephant object while the attributes "is small, "is white" can be applied to the lamb object. Clearly, then, the various attributes are correlated. While "is white" and "is gray" refer to color, "is small" and "is big" refer to size. Each attribute group [41] denotes a concrete physically observable property. In this paper, we refer to these groups as "concepts". Furthermore, attributes apart from simply being binary (presence/absence) or categorical variables (colors), can also be relative. For example, the lamb object for the concept "size" can be assigned the relative "is smaller than" attribute, instead of "is small" [40], but we do not address these in this paper.

Begin by considering n concepts, e.g., shape, from visual domain $\mathcal{V}$, each of which corresponds to one of n concepts from the audial domain $\mathcal{A}$. This correspondence can be represented as a permutation $\pi \in S_n$, where S_n is the symmetric group, i.e., the set of all bijective mappings from $\{1 \ldots n\}$ to itself. We define π such that if visual concept $\mathcal{V}_j$ corresponds to audial concept $\mathcal{A}_k$, $\pi(j) = k$. This implies that each visual concept is linked to one and only one audial concept, and thus, we assume we will never see audio descriptions with duplicated concepts such as *red blue* or *hexagon circle*. Please

note that $\mathcal{V}_j$ and $\mathcal{A}_k$ refer to a subspace in the joint visual and audio feature space corresponding to jth visual and kth audio concept respectively.

We seek to learn not only the correspondence π between visual and audial features, but also the mappings between them, $\phi_k(x_i^k) = y_i^{\pi(k)}$, via dictionary learning. For a set of N visual samples $\{\mathbf{x_i}\}_{i=1}^N$ and audial samples $\{\mathbf{y_i}\}_{i=1}^N$, each consisting of n concepts $\mathbf{x_i} = [x_i^1, x_i^2, \ldots x_i^n]$ and $\mathbf{y_i} = [y_i^1, y_i^2, \ldots y_i^n]$, our goal is to find dictionaries $\mathbf{U} = \{U^k\}_{k=1}^n$ and $\mathbf{V} = \{V^k\}_{k=1}^n$ by minimizing

$$\begin{aligned} &\arg\min_{\pi,\mathbf{U},\mathbf{V},\alpha} \sum_{i=1}^{N}\sum_{k=1}^{n}(\|x_i^k - U^k\boldsymbol{\alpha}_i^k\|_2^2 + \|y_i^{\pi(k)} - V^k\boldsymbol{\alpha}_i^k\|_2^2) \\ &\text{s.t.} \quad \pi \in S_n, \|\boldsymbol{\alpha}_i^k\|_1 \leq \lambda^k, \|U^k\|_2 \leq 1, \|V^k\|_2 \leq 1 \; \forall\{i,k\}. \end{aligned} \tag{11}$$

The inverse mapping $\mathbf{y} \mapsto \mathbf{x}$ is obtained by first projecting $\mathbf{y}$ on the learned basis

$$\boldsymbol{\alpha}^* = \arg\min_{\alpha^k} \sum_{k=1}^{n} \|V^k\boldsymbol{\alpha}^k - y^{\pi(k)}\|_2^2 \text{ s.t. } ||\boldsymbol{\alpha}^k||_1 \leq \lambda^k, \tag{12}$$

and then applying the shared coefficients to the visual dictionary: $\mathbf{x} = \mathbf{U}\alpha^*$. Note that when $n = 1$, the compositional model reduces to the paired model described earlier in Section 2.3.

Assume we have a permutation π (we show how to learn it in the next section) that aligns the audial and visual concepts. Finding the dictionaries $\mathbf{U}$ and $\mathbf{V}$ is straightforward. Notice in Equation (11) that once we know the permutation π, each of the n concepts is independent of the others, allowing us to find U^k, V^k, and $\boldsymbol{\alpha}^k$ individually for each k. By stacking the squared penalty terms for a single k in Equation (11), we obtain a traditional dictionary learning minimization problem

$$\begin{aligned} &\arg\min_{U^k,V^k,\boldsymbol{\alpha}^k} \sum_{i=1}^{N} \left\| \begin{bmatrix} x_i^k \\ y_i^{\pi(k)} \end{bmatrix} - \begin{bmatrix} U^k \\ V^k \end{bmatrix} \boldsymbol{\alpha}_i^k \right\|_2^2 \\ &\text{s.t.} \quad \|\boldsymbol{\alpha}_i^k\|_1 \leq \lambda^k, \|U^k\|_2 \leq 1, \|V^k\|_2 \leq 1 \; \forall k. \end{aligned} \tag{13}$$

We further use this finding towards the convergence proof in Section 2.4.3. We use the same techniques to perform this minimization as with the paired dictionary model.

2.4.1 Estimating Permutation π

Observe that the constraint on π is combinatorial, and for n concepts, there are $n!$ possible choices for π. A brute-force search for the optimal permutation would thus involve solving a dictionary learning optimization $n!$ times. For our example, where $n = 2$, this is feasible, but a more general and scalable approach is needed. We estimate the mapping π separately before the dictionary learning, based on the intuition that distance in visual and in audial feature representations of the same physically-grounded element should co-vary.

Because the features belong to different vector spaces, correlation coefficients cannot be computed directly. Instead, we form a cost matrix C such that $C_{j,k}$ is the cost of assigning visual concept $\mathcal{V}_j$ to audial concept $\mathcal{A}_k$, i.e., $\pi(j) = k$, and use the Hungarian algorithm [43] to find the minimum-cost assignment. To determine $C_{j,k}$, we cluster visual domain $\mathcal{V}_j$ and audial domain $\mathcal{A}_k$ (using assumed prior knowledge of the number of observed classes within $\mathcal{V}_j$) and find their V-measure [44]. V-measure is a harmonic mean of homogeneity and completeness of clustering methods. The homogeneity criterion measures the entropy of class distribution in individual clusters (each cluster should only have a single class) and the completeness criterion measures the entropy of cluster distribution assignment for each class (each class should be assigned to a single cluster). Because a V-measure of 1 represents high cluster similarity, while a V-measure of 0 represents low similarity, we set the cost $C_{j,k}$ to 1 minus the V-measure, so that desirable assignments have low cost.

2.4.2 Variable Word Order

Having a single permutation π for an entire set of data requires that the audial concepts appear in the same order for every sample. This is problematic, however, since there is not always a well-defined grammatical order for words in a description. If we were to use the concepts of color, shape, and size, *large red rectangle* and *red large rectangle*, while both grammatically valid, would require different choices of π, and thus could not exist in the same data set.

This motivates a generalization of the above compositional sparse learning framework. Instead of a single permutation π, we estimate a separate permutation $\{\pi_i\}_{i=1}^N$ for each sample, representing the mapping between the (known) visual concepts and the audial concepts as they appear in sample i. Formally, $\pi_i(j) = k$ if visual concept $\mathcal{V}_j$ corresponds to y_i^k, the kth audial concept of sample i. Thus, we now must minimize

$$\begin{aligned} &\arg\min_{\pi_i,U^k,V^k,\boldsymbol{\alpha}^k} \sum_{i=1}^{N}\sum_{k=1}^{n}(\|x_i^k - U^k\boldsymbol{\alpha}_i^k\|_2^2 + \|y_i^{\pi_i(k)} - V^k\boldsymbol{\alpha}_i^k\|_2^2) \\ &\text{s.t.} \quad \pi_i \in S_n, \|\boldsymbol{\alpha}_i^k\|_1 \leq \lambda^k, \|U^k\|_2 \leq 1, \|V^k\|_2 \leq 1 \; \forall\{i,k\}, \end{aligned} \tag{14}$$

with the inverse mapping $\mathbf{y} \mapsto \mathbf{x}$ being similarly modified.

We proceed as before, estimating the permutations $\{\pi_i\}_{i=1}^N$ before performing the dictionary learning step. Using the same intuition that distance in visual and audial representations of a common physical element will co-vary, we conjecture that samples near to x_i^k in $\mathcal{V}_k$ should all have one audial feature that is near to $y_i^{\pi_i(k)}$ in $\mathcal{A}_{\pi_i(k)}$. Furthermore, for $\pi_i(j) \neq k$, the audio samples for those same neighbors will not, on average, be near y_i^j in $\mathcal{A}_j$.

This puts us in a position to perform an assignment similar to the one above, but this time, separately for each sample. We again form a cost matrix C, where $C_{j,k}$ encodes the cost of assigning visual concept $\mathcal{V}_j$ to y_i^k. The cost $C_{j,k}$ is found as the average distance between y_i^k and the nearest audial feature of each sample that neighbors x_i^j in $\mathcal{V}_j$

$$C_{j,k} = \frac{1}{|\mathcal{N}_i^j|} \sum_{s:s\in\mathcal{N}_i^j} \min_t \|y_i^k - y_s^t\|_2, \tag{15}$$

where $\mathcal{N}_i^j$ is the set of samples that neighbor x_i^j in $\mathcal{V}_j$.

Ideally, the neighbor set $\mathcal{N}$ would capture all samples grounded in the same physical element as x_i^j. This is trivial if we, again, assume prior knowledge of the number of

observed classes in each domain: One can cluster the data and let $\mathcal{N}$ to be the entire cluster to which x_i^j belongs. However, we found that using a K-nearest-neighbors approach worked equally well for a wide range of values of K, and freed the model of the unnecessary assumption.

It is worth noting that this formulation allows us to handle test data that is missing words. For a sample with the kth word missing, we can simply set the kth column of C to zeros, and the missing word will be assigned whichever visual concept is not assigned to the other words. With more than one missing word, the remaining concepts will be distributed arbitrarily. Our dictionary learning model cannot handle such data in the training process, however, and the information available from missing concepts is not explored in detail in this paper.

2.4.3 Convergence Proof

Theorem 1 (Convergence of Compositional Sparse Model). *Given the correct permutation π, the compositional sparse model uniformly converges.*

Proof. Assuming the correct alignment to drop the superscript notation and using the independence of concepts, i.e., x^i is independent of x^j and same for y^i, we can write

$$\begin{aligned} P(x,y) &\equiv P(x^1, x^2, \ldots x^n, y^1, y^2, \ldots y^n) \\ &= P(x^1, y^1) P(x^2, y^2) \ldots P(x^n, y^n). \end{aligned} \tag{16}$$

Hence, data likelihood is factorized as

$$P(x, y|\mathbf{U}, \mathbf{V}, \alpha) = \prod_k P(x^k, y^k|U^k, V^k, \alpha^k). \tag{17}$$

Taking the negative log-likelihood

$$-\log P(x, y|\mathbf{U}, \mathbf{V}, \alpha) = \sum_k -\log P(x^k, y^k|U^k, V^k, \alpha^k),$$

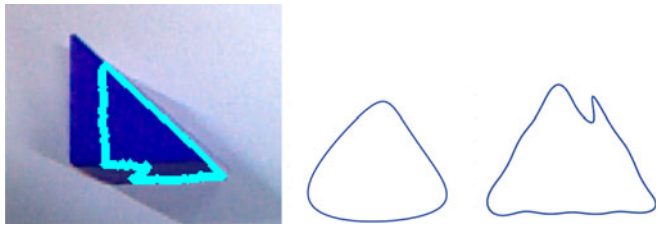
and following the derivation from Section 2.3, each term on the right is dictionary learning term of the combined space (x^k, y^k). Hence, the overall objective in Equation (11) under the Laplace prior on each α^k and zero mean additive Gaussian noise in signal approximation corresponds to dictionary learning in the joint space (x, y).

Based on the main result from Gribonval et al. [45], the solution of Equation (11) would converge with "overwhelming" majority to the true solution if the samples (x_i, y_i) are drawn independently. □

Compositional sparse dictionary learning problem exploits the independence of concepts in various domains to model the joint distribution $P(x, y)$ as captured in paired dictionary model. Minimizing the negative log-likelihood of $P(x, y)$ directly corresponds to the sum of negative log-likelihoods of each concept $P(x^i, y^i)$ which in turn is the objective from paired-dictionary models.

3 Features

In this paper, we restrict our study to the concepts of color and shape, without loss of generality. We extract color and shape features from the visual domain and segment the audio into two parts (ideally, words) representing individual concepts. The emphasis of this work is on the paired and compositional sparse modeling rather than the features, so we resort to capable yet simple features that allow for manual interpretation.



(a) Segmented Image (b) 2 Harmonics (c) 10 Harmonics

Fig. 3. Fourier representation of a triangular shape with 2 and 10 Fourier harmonics.

Please note that the proposed models are independent of underlying feature space representation. The only requirement for the current work is the independence of feature representation for different concepts.

3.1 Visual

Similar to Matuszek et al. [2], we first segment the image (in HSV space). Since the shapes used in current work consist of basic colors, they are relatively easily segmented from the background using saturation. To represent color, we describe each segment by its mean RGB values. To represent the shape of each segment we opt for a global shape descriptor based on Fourier analysis of closed contours [46].

Fourier features represent a closed contour by decomposing the contours over spectral frequency. Lower frequencies capture the mean of shape while higher frequencies account for subtle variations in the closed contours. The visual system of humans is found to have capabilities to form two- and three-dimensional percepts using only one-dimensional contour information [47].

We extract contours of the segmented/foreground object and use chain codes [48] to simplify analytical extraction of elliptic Fourier features [46]. After removing the constant Fourier component, we introduce rotational invariance by rotating of Fourier elliptical loci with respect to the major axes of the first harmonic [46]. Fig. 3 shows the Fourier feature representation of a contour of a segmented triangular shape. It can be seen that the shape is represented as a triangle even with 2 harmonics given the imperfect segmentation. Note, the representation is invariant to position, rotation and scale and hence the figure shows the triangle in a standardized coordinate frame.

3.2 Audio

We use Mel Frequency Cepstral Coefficients (MFCC) [49], which are widely used in the audio literature to represent audio signals. MFCC features are obtained by dividing the audio signal into small temporal frames and extracting cepstral features for each frame. This feature models important human perception characteristics by ignoring phase information and modeling frequency on a "Mel" scale [49]. Since the audio files are of different time lengths, we select the 15 frames with the highest sum of squared coefficients, intuitively capturing the highest-energy temporal sections

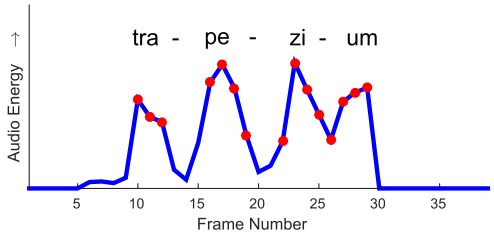


Fig. 4. Approximate energy per frame for an audio recording of the word *trapezium*. Note the four peaks corresponding to the four syllables in the word. Red markers indicate 15 frames selected for use in the audio feature vector representing the sample.

of the signal. Fig. 4 shows the sum of squared coefficients for each frame in an audio sample for the word *trapezium*, with marked points denoting the 15 frames selected. The MFCCs for these frames are then concatenated into a feature vector in their original temporal order.

4 Experiments and Results

We perform rigorous qualitative and quantitative evaluation to test generalization and reproduction abilities of the paired sparse and compositional sparse models. Quantitative performance is estimated to assess reproduction ability of the algorithm by performing 3-fold cross-validation. Qualitative performance is evaluated to infer the generalization capabilities of the proposed compositional sparse model and compare its performance with the non-compositional paired sparse model. For the purpose of presenting results, we only consider mapping from audio to visual in order to depict results in the paper. However, with the model both audio to visual and visual to audio representations can be derived.

We extract 195-dimensional audio features (13 coefficients each from the selected 15 audio frames), 20 Fourier harmonics, 3-dimensional color features and fix $\lambda = 0.15$ for all of the experiments involving paired and compositional sparse models. For MTL, we chose the norm penalty parameter $\lambda = 1.0$ (default in MALSAR package) as it seemed to generate best results on our dataset after trying out few other values.

Dataset. We acquired a new dataset of shapes and colors with 156 different examples (Table 1) of images showing a shape captured from a camera in various rotation and translation on the tabletop. We generated machine audio[1] that describes the color and shape of the captured image (e.g., *red rectangle*) with random speeds. We also produced segmented audio by generating machine audio separately for color and shape of the referred image to be used with the compositional model.

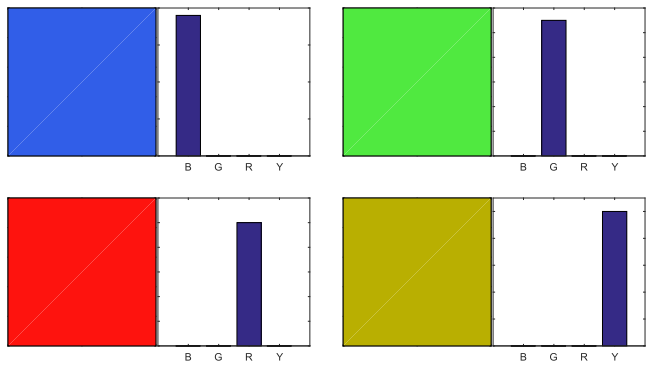
Visualization. To generate a visualization (audial-to-visual generation), we use the inverse mapping ϕ^{-1} and concept assignment π to generate visual features from audial features. The generated visual feature consists of Fourier features and mean RGB intensity values. Since Fourier features are rotation and translation invariant, a close representation of the original image can not be generated. Instead, we reconstruct the contour using these features and fill the contour with predicted RGB values. Figs. 7 and 12 contain examples generated by this visualization method.

1. We use the `espeak` package for generating machine audio.

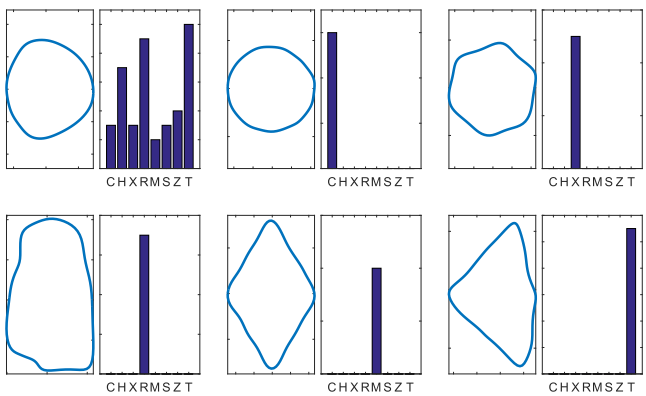
TABLE 1
Shape and Color Exemplars in the Dataset

Shape\Color	Blue	Green	Red	Yellow	Total
Circle	6	6	2	6	20
HalfCircle	6	4	4	4	18
Rectangle	6	6	6	2	20
Rhombus	10	0	0	0	10
Square	10	10	10	10	40
Triangle	8	6	8	6	28
Trapezium	0	0	10	0	10
Hexagon	0	0	0	10	10
Total	46	32	40	38	

Additionally, visualizing individual dictionary elements is often useful to verify and inspect the bimodal information being coded into each concept's dictionary. Fig. 5 shows six elements of each dictionary, visualized as a patch of the dictionary element's RGB hue (for the color dictionary) or the reconstructed Fourier contour (for the shape dictionary), alongside a histogram of the ground-truth labels of the samples that make use of that element in their sparse representation.



(a) Color dictionary elements (Histogram key: *Blue*, *Green*, *Red*, *Yellow*)



(b) Shape dictionary elements (Histogram key: *Circle*, *Halfcircle*, *heXagon*, *Rectangle*, *rhoMbus*, *Square*, *trapeZium*, *Triangle*)

Fig. 5. Visualization of elements from both dictionaries. Each element is a concatenation of a visual feature and an audial feature. The visual features (the RGB color and the Fourier contour coefficients) are plotted for each sample element above, alongside a histogram indicating the number of original data samples that make use of the element in their sparse representations. The histogram is a visual substitute for the audial feature, which is difficult to visualize.

TABLE 2
V-Measure Distance Matrix Between the Feature Representation

	a_1	a_2
v_1	0.1	1
v_2	0.4	0.1

v_1 and v_2 represent RGB and Fourier descriptor features, respectively, a_1 and a_2 represent the feature extracted from first and second audio segment.

4.1 Concept Assignment Evaluation

The success of the compositional sparse model depends largely on the correct estimation of the correspondence between the visual and audial concepts. In our first model, this amounts to estimating the single permutation π correctly. Table 2 shows the V-measure scores used to construct the cost matrix C, computed using the entire dataset. RGB and shape concepts are denoted by v_1 and v_2, respectively. Audio concepts a_1 and a_2 represent the first and second words in each sample, corresponding in this instance to color descriptors and shape descriptors, respectively. These V-measure scores result in the correct alignment of $v_1 = \pi(a_2)$ and $v_2 = \pi(a_1)$, which was subsequently used in our reproduction and generalization experiments.

In the second model, we must estimate the word order permutation π_i for each sample. To examine the accuracy of this estimation, we swap the word order of randomly chosen samples and record the proportion of samples for which our estimation procedure correctly guesses the word order. As noted before, this can be done assuming prior knowledge of the class distribution within each concept, and using this prior afforded 100 percent alignment accuracy. However, we prefer to avoid this assumption, and instead select a constant number of neighbors for use in the alignment procedure. The word alignment accuracy for a wide range of values for K is shown in Fig. 6. Note that there are many K values for which all samples are aligned correctly. When this happens, the subsequent dictionary learning process proceeds exactly as before, yielding the same results.

The alignment procedure depends on K being large enough to capture variety in the feature spaces. For example, when aligning the utterance "blue square", we gain no information about the word "blue" if most of the neighbors in color space also happen to be squares; thus, small values of K yield poor alignment. Conversely, if we choose too many neighbors, our information about the word "blue" may be polluted with information from red, yellow, and green samples; thus, large values of K yield similarly poor results.

4.2 Reproduction Evaluation

For reproduction, we seek to evaluate the performance of a robot for a theoretical command, e.g., *pick a 'red rectangle' from a box full of all the shapes*, which is a subset of the broader picture described in the Introduction (Section 1). We perform a 3-fold cross-validation study to assess this retrieval performance by dividing the dataset into 3 parts, using 2 parts for training and remaining part for testing (and then permuting the sets). We test retrieval performance for different concepts (color and shape) separately for paired sparse learning and compositional sparse learning. A color or shape is determined

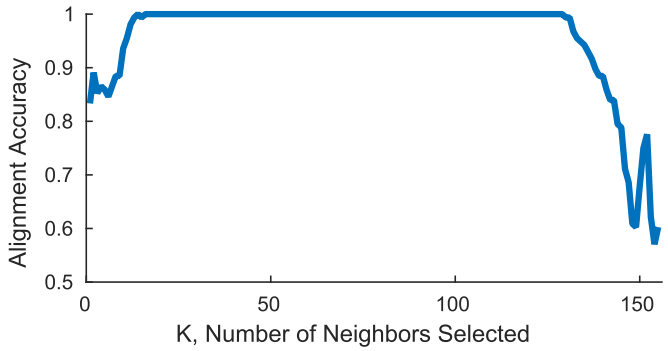


Fig. 6. Word alignment accuracy over our dataset ($N = 156$) versus K, the number of neighbors selected for use in the alignment procedure described in Section 2.3.1. Note that accuracy is 1.0 from $K = 16$ to $K = 129$.

to be correctly understood by the robot if the said color or shape is present in top k retrieved examples. Retrieval is performed by first extracting the audial feature from the audial stream, using the learned dictionaries and permutation π to extract visual features and then picking the closest object from all the training examples.

The closeness of a visual object to generated visual feature is measured by a distance metric in the visual feature space. We compare the feature vectors to extract k nearest neighbors using an l_1 distance, in the appropriate concept feature subspace. For evaluation, we set the parameter k to be 5, which means that if there is a match in the top 5 nearest neighbors, the retrieval is deemed to be successful. Fig. 7 shows the reproduction performance for an audial utterance *blue halfcircle*. It is observed that while the compositional model gets both the color and shape correct, the paired model fails in reproducing the correct shape.

Fig. 8 compares the quantitative retrieval performance for the compositional, paired and MTL models. The MTL model uses the convex multi-task learning approach that selects both task-specific and common-across-tasks features [33]. It is observed that both of the proposed models perform better than MTL: The overall accuracy for convex MTL is 100 percent for colors and 42.95 percent for MTL; for the paired model, 99.36 percent for colors and 51.92 percent for shapes; for the compositional model, 100 percent for colors

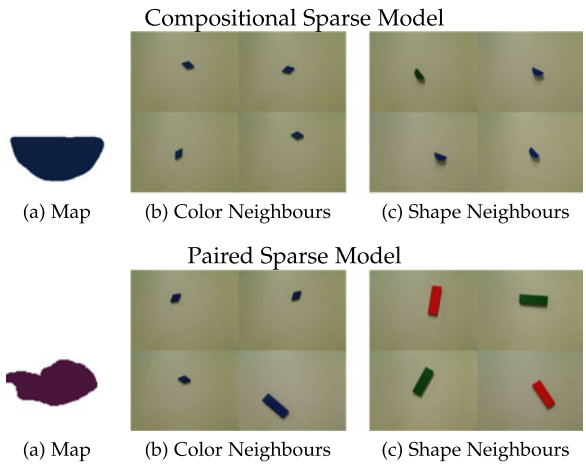


Fig. 7. For the audial utterance *blue halfcircle*, (a) generated image by mapping from audial to visual domain. (b), (c) retrieval of color and shape neighbors by both models.

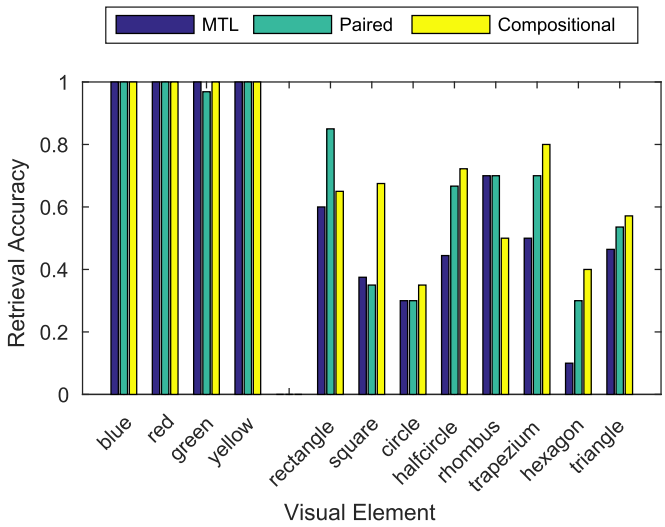


Fig. 8. Comparison of correct retrievals by three different algorithms, multi-task learning (MTL), paired, and compositional. The dictionary learning methods use a dictionary size of 25 elements.

and 61.54 percent for shapes. We tried other models of MTL including Lasso regression on the weight matrix but the convex MTL approach gave better results. Clearly, both the proposed models are capable in certain combinations of color and shape. However, while the overall performance on color is saturated for all the models, this is not the case for shape. In fact, the paired model slightly outperforms for rectangle and shape whereas the compositional model outperforms, sometimes significantly, for square, circle, halfcircle, trapezium, hexagon, and triangle. We suspect the dominant performance of the compositional on many of these shapes is due to the fact that it learns a separate shape dictionary whereas the paired model learns a combined one (recall Fig. 5). Furthermore, both paired and compositional sparse models are better then MTL methods because they model the target distribution apart from the input feature distribution.

Effect of Dictionary Size. The comparable results described above arise because both models are appropriate for the task of retrieving data belonging to composite classes seen in training. As discussed in the Introduction, however, the compositional model learns the shape and color separately, while the paired model effectively learns each composite class on its own, intuitively necessitating a larger dictionary to achieve the explanatory power observed in the compositional model. For example, for the shape dictionary of size 50 learned by the compositional model, there are multiple dictionary elements capturing only noise; three of these are visualized in Fig. 9. We demonstrate the disparity in model size in Fig. 10 by plotting the overall shape and color retrieval accuracy values for

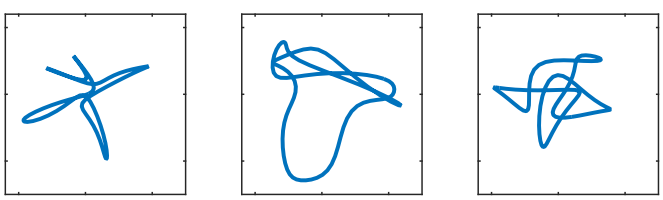


Fig. 9. Noise elements of the compositional model's shape dictionary, for dictionary size of 50.

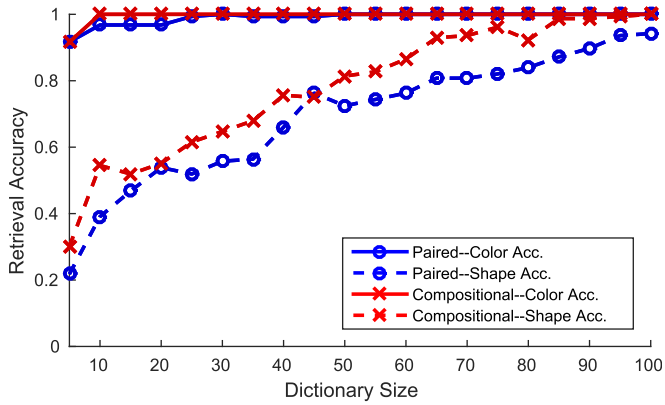


Fig. 10. Overall color and shape retrieval accuracy for both the paired and compositional models, by the number of dictionary elements used in learning.

both models over a range of dictionary sizes. As expected, the paired model lags behind the compositional model for both color and shape accuracy scores.

Effect of Feature Space. The underlying feature space for various concepts has a significant impact on the retrieval performance. To concretely analyze the effect of feature space, we chose the shape concept and extracted Histogram of Oriented Gradients (HOG) features [50] on a cropped image guided by the contours computed for the Fourier features. The contours are used to determine a bounding box of size 160×120 around the object in the image which is then resized to half the size. The resized image is used to compute HOG features using OpenCV implementation with default parameters. This yields a feature vector of size 14,000 which is unusable with the size of our dataset. To reduce the dimensions of the feature vector, we perform Principal Components Analysis (PCA) on our dataset to reduce the dimensionality to 40 which is same as the Fourier feature vector. Fig. 11 demonstrates that the performance of compositional model increases from an average accuracy of 61.54 to 91.37 percent by using HOG features.

However, it is hard to reconstruct the original image using HOG features because a significant amount of information is lost in the feature extraction process. Indeed,

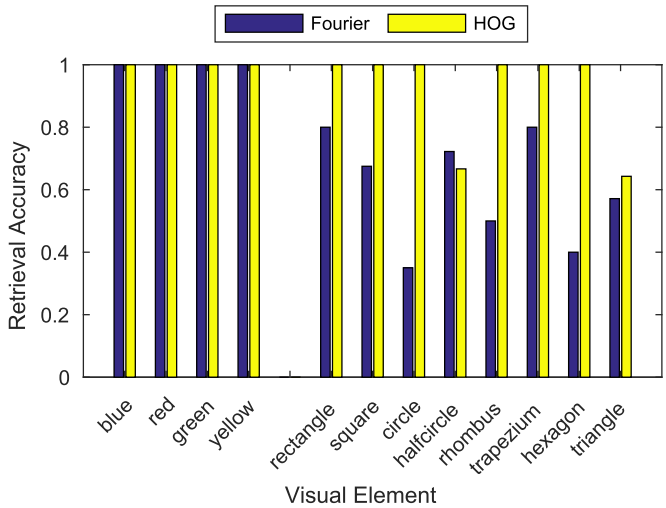


Fig. 11. Effect of underlying feature space on retrieval accuracy for shape concept using Fourier features and HOG features.

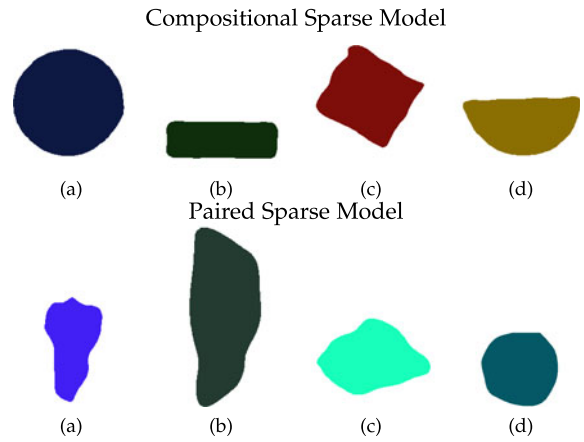


Fig. 12. Generalization performance result depiction for audial utterances (a) *blue circle* (b) *green rectangle* (c) *red square* (d) *yellow halfcircle*.

image reconstruction using visual features alone is an active research area [51]. For the rest of the paper, we only work with Fourier features because it is easy to reconstruct the original image and the compositional model is independent of the underlying feature space.

4.3 Generalization Evaluation

We test compositional sparse and paired sparse models with respect to their generalization capabilities on novel samples. Here, we test generalization across color and shape. Generalization is evaluated by generating images of a particular color and shape whose training examples have been removed from the dataset. For a good generalization performance, the model must generate implicit meaning of utterances such as *green* and *triangle*.

Fig. 12 shows the pictorial results from various audio utterances from compositional sparse and paired sparse models. For the audio utterance *blue circle*, both models get the right color but compositional model achieves better shape generation which is the case for utterance *green rectangle* as well. For the audial utterance *red square* compositional model achieves both shape and color while the paired model is not able to represent color. From these examples, it is clear that compositional model can handle generalization both across shape and color much better compared to the paired model. The paired sparse model—as reflected in these results—is incompetent for this task because it does not distinguish between individual percepts.

For qualitative evaluation of generalization capabilities, we use evaluation by human subjects. For this, we generate two sets of images, one from the compositional model and one from the paired model. Each set of these images is then presented to human subjects through a web-based user interface, and the humans are asked to "Describe the color and shape of the above image" while being presented with the color and shape options along with "None of these" from the training data. Note that in this experiment the human subject is not shown any samples from the training data. Hence, we call these experiment *compositional unbiased* and *paired unbiased* depending on the generating model.

In another set of experiments we *bias* the human subject by showing them an example image of the color and shape for which the image has been generated. The subject is expected to answer in "Yes" or "No" to the question: "Is the color (shape) of the above image same as the example image?" Whenever the subject says "Yes", we take the response as the expected color/shape; for "No" we assume "None of these" option.

Fig. 13 shows the human qualitative performance metrics for this test. It is observed that color generalizes almost perfectly using our proposed compositional sparse model while

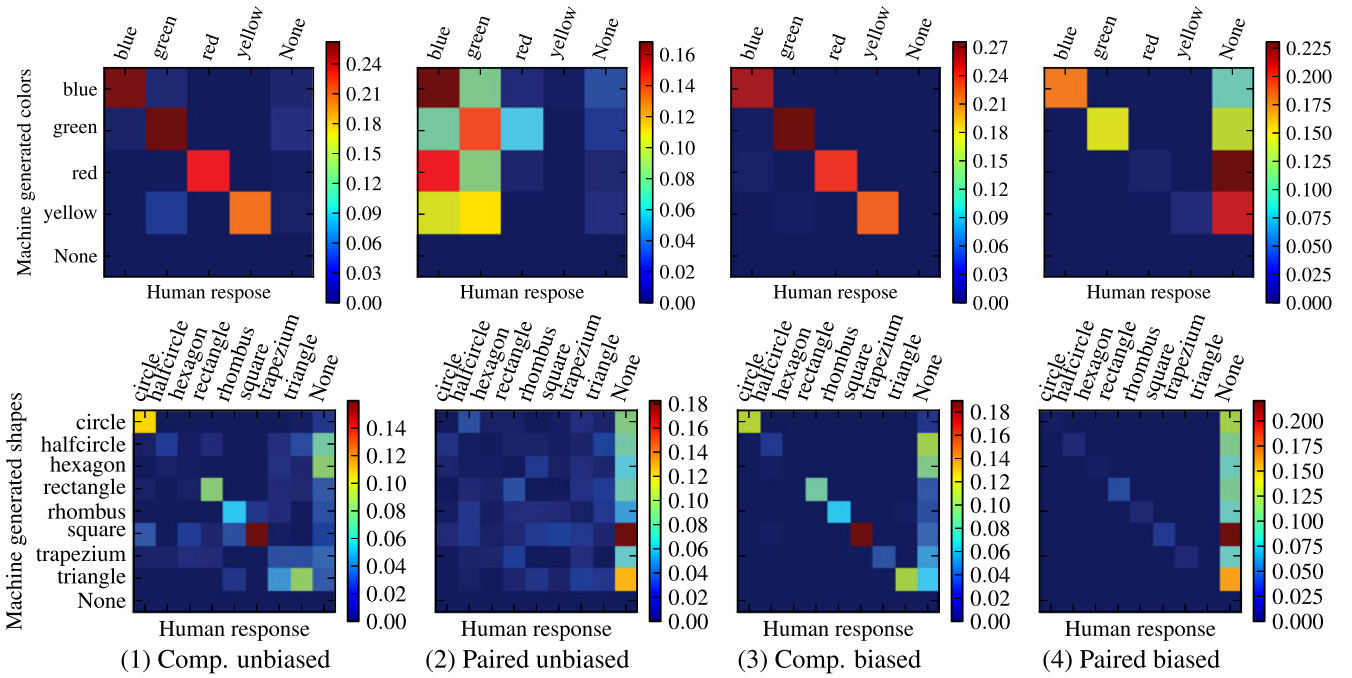


Fig. 13. Confusion matrices for generalization experiments evaluated by human subjects. Rows are for different features: Colors and shapes. Columns from left to right are four different experiments (1) images generated by compositional model are evaluated by humans with *unbiased* questions like "Describe the color and shape of this image" from fixed set of choices (2) paired model with *unbiased* questions. (3) Compositional model with biased questions like "Is the shape of generated image same as the given example image?" (4) paired model with biased questions.

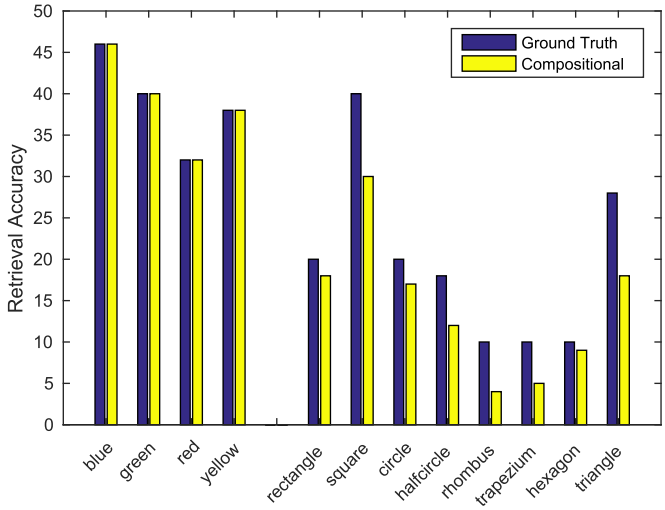


Fig. 14. Comparison of correct retrievals for language translation from English to Hindi using the compositional sparse dictionary with 50 elements.

the paired sparse model gives poor performance in both biased and unbiased human evaluation. On the generalization of shape, the compositional model again achieves much better performance over the baseline paired model in both biased and unbiased experiments. Using the internal semantics of humans, it is observed that *halfcircle* is frequently represented as *rectangle* or *trapezium* for the compositional model. It is likely because of the shape feature with invariance whose closed contour representation is not enough to distinguish perceptually similar shapes. Furthermore, *triangle* is often mistaken as *trapezium* which can be explained by a similar starting sound. It is seen that biased results give better performance denoting improved assessment after recalibration of human semantics to current experimental shapes.

4.4 Visually Grounded Translation

We demonstrate another application of the proposed model in performing vision-grounded language translation that requires no knowledge of a phrase dictionary as required by state-of-the art machine translation systems [52]. The key idea of our approach is to relate the audial utterances that refer to the same visually grounded concept by first learning the audial-to-visual concept mapping separately for different languages and then using the visual correspondence to relate audial utterances. In this paper, we demonstrate machine translation between English and Hindi (the official language of India) language.

To the best of our knowledge, there is no prior work that attempts to do audial language translation without using any prior knowledge of translation between the languages. We use vision grounding as an intermediary instead of the phrase translation dictionary. Our end-goal is similar to pictorially grounded language [53] which translates between two languages by using a data-set of single word text labeled images in both languages. For example, the dataset in [53] consists of triplets of the form (I_i, a_i, b_i) where I_i is the image and a_i, b_i are its single word text labels in two different languages. In contrast, we take sentence-like audial descriptions of an image in both the languages and use compositionality to ground concepts from both languages on to a common visual concept thereby finding the corresponding words in the two (and possibly more) languages.

Translation aims at finding a reversible mapping $T : \mathcal{A}^1 \Leftrightarrow \mathcal{A}^2$, where $\mathcal{A}^1$ and $\mathcal{A}^2$ contain audio segments on two different languages. We use the compositional dictionary learning model to map audio utterances in the first language ($\mathcal{A}^1$) to visual domain (V). The resulting visual feature is then mapped to audio feature in the second language ($\mathcal{A}^2$). We use the same reproduction evaluation as proposed in Section 4.2 to test the translation model. Fig. 14 demonstrates the performance of our compositional dictionary model for the proposed grounded language translation. Notably, we are able to retrieve the color audio segment correctly for the entire dataset and achieve a accuracy of 72.44 percent on shape audio translation.

5 Conclusion

We propose a novel model representing bimodal percepts that exploits the compositional structure of language. Our compositional sparse learning approach jointly learns the over-complete dictionaries, sparse bases, and cross-modal linking matrix. In contrast to prior work in bimodal modeling which is primarily discriminative in nature, e.g., [5], [54], our compositional sparse learning approach is generative and hence transparent. We demonstrate the effectiveness of sparsity and compositionality by both qualitative and quantitative evaluations.

In future, we will extend the compositionality by removing the constraint of one-to-one mapping which will be important for percepts grounded in more than one basic feature representation of the input. Our methods also share the full support and the sparse coefficients for the reconstruction vector in all the modalities. However, in case of significant noise/non-correlation between modalities, we will need to extend our models to allow sharing support and coefficients only for some part of reconstruction vectors.

Acknowledgments

We acknowledge the partial support for this work from the National Science Foundation NRI IIS 1522904, Army Research Office W911NF-15-1-0354, as well as the University of Michigan College of Engineering. We thank Julien Mairal and collaborators for their release of the SPAMS package (http://spams-devel.gforge.inria.fr/) as well as Niclas Börlin for his publicly available MATLAB implementation of the Hungarian algorithm (http://www.mathworks.com/matlabcentral/fileexchange/94-assignprob-zip).

Data and Software

The software for our earlier conference publication [17] (AAAI 2014) of this work is available at https://bitbucket.org/surenkum/bimodal_sparse. Upon publication both data and updated software will be released in this same location.

References

[1] N. Kyriazis and A. Argyros, "Physically plausible 3d scene tracking: The single actor hypothesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 9–16.

[2] C. Matuszek, N. Fitzgerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," in *Proc. 29th Int. Conf. Mach. Learning*, 2012, pp. 1671–1678.

[3] R. Jackendoff, *Semantics and Cognition*. Cambridge, MA, USA: MIT Press, 1983.
[4] P. Vogt, "The physical symbol grounding problem," *Cognitive Syst. Rese.*, vol. 3, no. 3, pp. 429–457, 2002.
[5] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Sci.*, vol. 26, no. 1, pp. 113–146, 2002.
[6] D. L. Chen and R. J. Mooney, "Learning to interpret natural language navigation instructions from observations," in *Proc. 25th AAAI Conf. Artifi. Intell.*, 2011, pp. 7–11.
[7] N. Mavridis and D. Roy, "Grounded situation models for robots: Where words and percepts meet," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2006, pp. 4690–4697.
[8] K. Barnard, P. Duygulu, D. Forsyth, N. de Frietas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learning Res.*, vol. 3, pp. 1107–1135, 2003.
[9] R. A. Knepper, S. Tellex, A. Li, N. Roy, and D. Rus, "Single assembly robot in search of human partner: Versatile grounded language generation," in *Proc. 8th ACM/IEEE Int. Conf. Human-Robot Interaction*, 2013, pp. 167–168.
[10] S. Tellex, et al., "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proc. 25th AAAI Conf. Artif. Intell.*, 2011.
[11] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 2634–2641.
[12] A. Barbu, et al., "Video in sentences out," in *Proc. 28th Conf. Uncertainty Artif. Intell.*, 2012, pp. 102–112.
[13] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama, "Generating natural-language video descriptions using text-mined knowledge," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 541–547.
[14] H. Yu and J. M. Siskind, "Grounded language learning from videos described with sentences," in *Proc. 51st Annu. Meet. Assoc. Comput. Linguistics*, 2013.
[15] S. Fidler and A. Leonardis, "Towards scalable representations of object categories: Learning a hierarchy of parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
[16] J. Porway, B. Yao, and S. C. Zhu, "Learning compositional models for object categories from small sample sets," *Object Categorization: Comput. Human Vis. Perspectives*, no. 1, pp. 241–256, 2008.
[17] S. Kumar, V. Dhiman, and J. J. Corso, "Learning compositional sparse models of bimodal percepts," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 366–372.
[18] H. B. Barlow, "Possible principles underlying the transformation of sensory messages," *Sensory Commun.*, no. 13, pp. 217–234, 1961.
[19] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.
[20] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.
[21] E. Knudsen and M. Brainard, "Creating a unified representation of visual and auditory space in the brain," *Annu. Rev. Neuroscience*, vol. 18, no. 1, pp. 19–43, 1995.
[22] A. Evgeniou and M. Pontil, "Multi-task feature learning," *Adv. Neural Inf. Proc. Syst.*, vol. 19, 2007, Art. no. 41.
[23] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and low-rank patterns from multiple tasks," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 4, 2012, Art. no. 22.
[24] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 113–126, Jan. 2014.
[25] S. Bahrampour, A. Ray, N. Nasrabadi, and K. Jenkins, "Quality-based multimodal classification using tree-structured sparsity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 4114–4121.
[26] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Proc.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[27] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2216–2223.
[28] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "Hog-gles: Visualizing object detection features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1–8.
[29] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *J. Found. Trends Comput. Graph. Vis.*, vol. 8, no. 2/3, pp. 85–283, 2014.
[30] I. Tošić and P. Frossard, "Dictionary learning," *IEEE Signal Proc. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.
[31] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learning Res.*, vol. 11, pp. 19–60, 2010.
[32] J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," in *Proc. Adv. Neural Inf. Proc. Syst.*, 2011, pp. 702–710.
[33] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learning*, vol. 73, no. 3, pp. 243–272, 2008.
[34] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l 2, 1-norm minimization," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 339–348.
[35] Y. Yan, et al."Event oriented dictionary learning for complex event detection," *IEEE Trans. Image Proc.*, vol. 24, no. 6, pp. 1867–1878, Jun. 2015.
[36] J. Zhou, J. Chen, and J. Ye, *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011. [Online]. Available: http://www.public.asu.edu/~jye02/Software/MALSAR
[37] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statistical Soc. Ser. B Methodological*, vol. 73, no. 3, pp. 267–288, 1996.
[38] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. 26th Annu. Int. Conf. Mach. Learning*, 2009, pp. 689–696.
[39] B. Efron, et al., "Least angle regression," *Annal. Statistics*, vol. 32, no. 2, pp. 407–499, 2004.
[40] D. Parikh and K. Grauman, "Relative attributes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 503–510.
[41] W. Wang, Y. Yan, S. Winkler, and N. Sebe, "Category specific dictionary learning for attribute specific feature selection," *IEEE Trans. Image Proc.*, vol. 25, no. 3, pp. 1465–1478, Mar. 2016.
[42] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.
[43] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, no. 1–2, pp. 83–97, 1955.
[44] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proc. Joint Conf. Empirical Methods Natural Language Proc. Comput. Natural Language Learning*, vol. 7, 2007, pp. 410–420.
[45] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, "Sample complexity of dictionary learning and other matrix factorizations," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3469–3486, Jun. 2015.
[46] F. P. Kuhl and C. R. Giardina, "Elliptic Fourier features of a closed contour," *Comput. Graphics Image Proc.*, vol. 18, no. 3, pp. 236–258, 1982.
[47] J. Elder and S. Zucker, "The effect of contour closure on the rapid discrimination of two-dimensional shapes," *Vis. Res.*, vol. 33, no. 7, pp. 981–991, 1993.
[48] H. Freeman, "Computer processing of line-drawing images," *ACM Comput. Surveys*, vol. 6, no. 1, pp. 57–97, 1974.
[49] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. Int. Symp. Music Inf. Retrieval*, 2000.
[50] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 886–893.
[51] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "Hoggles: Visualizing object detection features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1–8.
[52] P. Koehn, et al., "Moses: Open source toolkit for statistical machine translation," in *Proc. 45th Annu. Meet. ACL Interactive Poster Demonstration Sessions*, 2007, pp. 177–180.
[53] S. Borgohain and S. B. Nair, "Towards a pictorially grounded language for machine-aided translation," *Int. J. Asian Lang. Proc.*, vol. 20, no. 3, pp. 87–110, 2010.
[54] S. Roller and S. S. im Walde, "A multimodal lda model integrating textual, cognitive and visual modalities," in *EMNLP*, Seattle, WA, October 2013, pp. 1146–1157. [Online]. Available: http://www.cs.utexas.edu/users/ai-lab/pub-view.php?PubID=127403

Suren Kumar received the BTech degree in mechanical engineering from Indian Institute of Technology, Roorkee, in 2009. He received the PhD degree from SUNY Buffalo, in 2015 on the topic of "Computer vision based articulated motion understanding" working with Dr. Venkat Krovi and Dr. Jason Corso. He is a postdoctoral research fellow in University of Michigan, Ann Arbor working on the problem of long-term behavior prediction. His research interests include sparse models, probabilistic graphical model, sensor fusion, object tracking, and deep learning. He is a member of the IEEE.

Vikas Dhiman received the BTech degree in electrical engineering from Indian Institute of Technology Roorkee, in 2008, and the MS degree in computer science from SUNY Buffalo, in 2014. He is currently working toward the PhD degree in the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor. His research interests are in the fields of computer vision and robotics with specialization on mapping, localization, probablistic graphical models.

Parker Koch received the BS degree in computer engineering and mathematics from the University of Iowa, in 2015, and is working toward the PhD degree in computer engineering at the University of Michigan, Ann Arbor. His research interests lie in dictionary learning and sparse representation with application to computer vision and machine learning, particularly multimodal problems.

Jason J. Corso received the BS Degree with honors from Loyola College in Maryland and the MSE and PhD degrees from The Johns Hopkins University, in 2000, 2002 and 2005, respectively, all in computer science. He is an associate professor of electrical engineering and computer science at the University of Michigan. He spent two years as a post-doctoral fellow at the University of California, Los Angeles. From 2007-14 he was a member of the computer science and engineering faculty at SUNY Buffalo. He received the Google Faculty Research Award 2015, the Army Research Office Young Investigator Award 2010, NSF CAREER award 2009, SUNY Buffalo Young Investigator Award 2011, a member of the 2009 DARPA Computer Science Study Group, and a recipient of the link foundation fellowship in advanced simulation and training 2003. Corso has authored more than one-hundred peer-reviewed papers on topics of his research interest including computer vision, robot perception, data science, machine learning and medical imaging. He is a member of the AAAI, ACM, MAA and a senior member of the IEEE.