2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Contextual Modeling with Labeled Multi-LDA

Cheng Zhang Dan Song Hedvig Kjellström

Abstract— **Learning about activities and object affordances from human demonstration are important cognitive capabilities for robots functioning in human environments, for example, being able to classify objects and knowing how to grasp them for different tasks. To achieve such capabilities, we propose a Labeled Multi-modal Latent Dirichlet Allocation (LM-LDA), which is a generative classifier trained with two different data cues, for instance, one cue can be traditional visual observation and another cue can be contextual information. The novel aspects of the LM-LDA classifier, compared to other methods for encoding contextual information are that, I) even with only one of the cues present at execution time, the classification will be better than single cue classification since cue correlations are encoded in the model, II) one of the cues (e.g., common grasps for the observed object class) can be inferred from the other cue (e.g., the appearance of the observed object). This makes the method suitable for robot online and transfer learning; a capability highly desirable in cognitive robotic applications. Our experiments show a clear improvement for classification and a reasonable inference of the missing data.**

I. Introduction

Contextual information has been proven to improve the performance of a number of visual recognition [1], [2], [3], [4] and robotic manipulation [5], [6] tasks. Fig. 1 shows an example of a contextual classification problem. Oranges and lemons have very similar visual appearance. However, they are handled in different ways: people commonly peel oranges to eat, but cut lemons with a knife. Thus, using examples of human actions on the two types of fruit in addition to their appearance could significantly improve the performance of an orange/lemon classifier.

However, when a robot perceives a new fruit, say a lemon laying on a table, the robot will commonly only see the appearance, not the action that is associated with the object. Thus, the classifier should utilize the action information even though it is not visible at execution time. Furthermore, the robot not only needs to recognize the fruit, but it also needs to know what to do with it. It should hence be possible to infer the range of actions that an object with a given appearance afford to an agent.

Analogously, context is also important in action recognition. As an example, based on observations of visual human motion only, it is extremely hard to distinguish between the actions vacuum-the-floor and mop-the-floor. However, the classification task becomes significantly easier with knowledge about the objects involved. Such object context has been shown to improve action recognition [2], [3], [7]. Another type of context is semantic information about the action [8], [9]. Moreover, given a task to perform, the robot should know what objects it needs to search for to execute the task. Given the task mopping, for example, the robot should be able to search for a mop.

In this paper we address the robotic cognition task by proposing a Labeled Multi-Latent Dirichlet Allocation (LM-LDA) model, which marries topic modeling and contextual cue integration, encoding information from two different cues within a single principled generative framework. The *two novel aspects of the LM-LDA* emanate from the model structure: The LM-LDA models the process of generating the two cues as dependent on a common, class-dependent topic distribution. When training the model with examples of observations, the correlations between the data from the two cues will be encoded in the topic model. Put differently, the contextual cue – peeling style in the fruit example in Fig. 1 – guides the classifier to what aspects of the appearance cue are relevant for classification. In the fruit example, the difference in appearance between oranges and lemons is very small. This means that the contextual cue during training affects the representation of the appearance cue so that the classifier to a higher degree pays attention to aspects in the appearance cue that are correlated with the contextual cue. Addressing this problem, we sum up our contribution in two aspects as:

Contribution 1: LM-LDA is able to achieve significantly better performance than single cue classification by modeling the context. Even if only one cue is visible at execution time, LM-LDA outperforms single-cue classification, since the correlations between the observed and unseen cue are encoded in the model.

Contribution 2: Since the LM-LDA is generative, it is possible to synthesize information about the missing cue from a one-cue observation of a new instance (Fig. 1, left). Moreover, as a topic model, LM-LDA represents the data in terms of ”topics” in a latent space. Hence, there is a potential to use the learned topics for transfer learning of action ”intention” [6] to other robot configurations with different grasping state space.

The rest of the paper is organized as follows. In Section II, related work on contextual recognition and topic models is reviewed. The generative process and parameter estimation of the LM-LDA model is described in Section III. In Section IV, we then discuss the application of this model to two different contextual recognition tasks. The performance of the model on the two tasks is then evaluated in Section V.

This research has been supported by the EU through TOMSY, IST-FP7-Collaborative Project-270436, and the Swedish Research Council (VR).

The authors are with CVAP/CAS, KTH, Stockholm, Sweden, chengz,dsong,hedvig@kth.se.

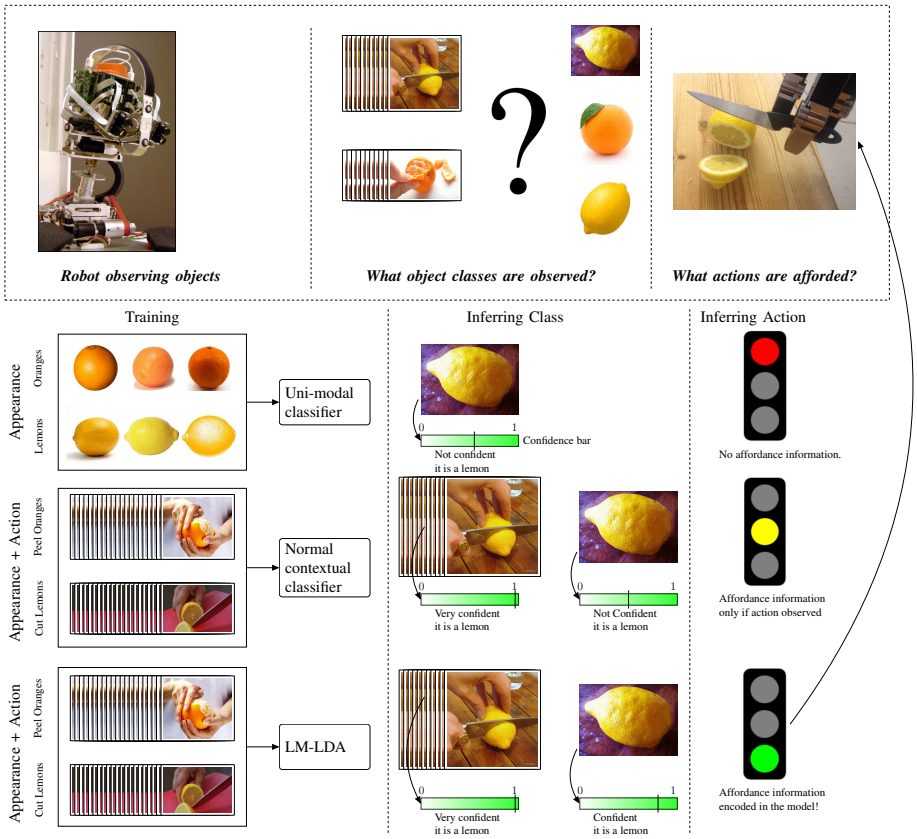


Fig. 1. Object classification for robot vision scenario. **Top row:** Standard uni-modal classification, which suffers from low classification accuracy in the case of visually similar classes with high intra-class variability. This approach is not able to infer the action, since action context is not modeled. **Middle row:** The normal contextual model, which can archive good classification performance when the contextual information is available in testing time. Commonly, it will fail, when contextual information is not present in testing time. **Bottom row:** The *contextual LM-LDA model* presented here encodes two different input modalities, in this example, object appearance and human actions associated with the object. Training data are I) appearance features extracted from each training instance, II) features encoding the human action associated with that instance. Given the trained model, a new set of appearance and action features can be classified with high confidence. In addition, the model is *able to classify a new object instance based on the appearance cue only, with considerably higher confidence than a standard appearance-based classifier*, since the LM-LDA model encodes correlations between the appearance and action cues. Furthermore, the encoded correlations give the possibility of *inferring the range of actions that an object with a given appearance afford to a human*.

II. Related Work

We propose a new model for contextual modeling in robotics, which not only inherits the strong classification properties of traditional contextual modeling, but also has stronger inference capabilities, thanks to it using a generative model with a latent space. In this section, we will briefly review related work on contextual and topic modeling.

The use of contextual information in vision recognition has shown very encouraging results in recent years [10]. A powerful contextual cue for object recognition is the scene around the object. In [11], [12], [13], the scene itself is used to guide object recognition. The scene itself is a strong prior cue as to which objects can be expected and where they are most likely to be found. Similarly, the spatial relationships between objects in the image are exploited in [14], [15].

Object recognition can also be guided by observations of human interaction with the objects; this is also explored in this paper. Moore et al. [16] provide a Bayesian framework for recognizing objects based on contextual information from other objects, human actions being performed on the object, and the scene. In [17], human actions are used to infer object class. Analogously, action recognition is supported by knowledge about the objects involved in the action. In [2] and [3], different types of sequential graphical models are used to simultaneously recognize objects and actions on those objects. They require both the action/motion cue and the object cue to be present at both training and testing time, limiting its use in practical applications.

Imitation learning, or learning from demonstration [18], [19], [20], [6], [5], [21] is a challenging problem in robotics. A large amount of context information that used in vision recognition could be well used in robotic imitation learning. An important application of imitation learning is robotic grasp selection. Song et al. [6] instead proposed a robot grasping planning method which aims to learn human intention and mapping to the robotic embodiment on this more abstracted level, instead of directly mapping the grasps from the human to the robotic grasp spaces. However, in their method, a fixed setting of high-level object and action parameters need to be manually specified. Madry and Song et

al. [4] integrated object classification and grasping generation system together. But it was processed as two separated step without modeling context. The approach proposed here learns these parameters from data, for example, using appearance features, and a latent topic space is used which enables contextual modeling. This topic space will thus contain information about the mapping between the two different data modalities – in other words, contain a high-level representation of the state (which could be a grasp, an object type, etc.) observed in the two data modalities. This high-level representation could be used for mapping of a grasp to a different robotic embodiment, but also for transfer learning: mapping an object model to a new observation setting, e.g., observing the object in a new environment.

Topic models have been applied successfully in text document retrieval [22], [23], [24], [25], modeling documents as bags (multi-set) of words, and subsequently in computer vision, modeling images as bags of visual words [26], [27], [14], [28], [29]. Fei-Fei et al. [27] show the feasibility of applying topic models to natural scene classification. New promising scene classification results are achieved using, for example, the visual concept presentation of [30]. Other extensions take spatial information into consideration [31], [14], [28]; here, spatial location is treated as contextual information to. All these extensions improve both classification and segmentation performance. However, topic modeling has not been well used in robotic community yet.

Latent Dirichlet Allocation (LDA) [23] is one of the most important approaches in topic modeling, assuming a generative process where a document is modeled as a probability distribution over topics, which in turn are modeled as distributions over words. There has been numerous extensions of LDA in recent years. Supervision is introduced for topic models in both a "downstream" manner [25] and an "upstream" manner [27], [32], [33]. Another direction to extend LDA is to use multiple cues [34], [29], [24]. Multi-Multinomial LDA (MM-LDA) [24] is successfully applied to computational linguistics, assuming that the words from both cues are multinomially distributed. In [24], web tags and web content are modeled jointly by forcing these two cues to share the same topic proportion parameters. We employ this approach since it gives an intuitive way of fusing information from different modalities. Furthermore, the topic space can be used for transfer learning, as discussed above [6].

Our LM-LDA model is based on MM-LDA, with an additional object category label. Introducing labeling to MM-LDA combines the advantages of both supervision and multiple cues in learning. Moreover, the introduction of a class label makes inference from partial observation possible. Thus, in contrast to previous contextual modeling work presented above, the present model can be used with or without observation of the second cue at test time (contribution 1, see the Introduction). The class label also gives the possibility to infer a range of possible context given just the regular visual cue (contribution 2, see the Introduction).

Next, the LM-LDA model is explained in more detail.

III. Labeled Multi-LDA

As discussed above, visual classification benefits from adding contextual cue to the traditional cue and topic model is able to extract topics/themes from documents/images. We propose the LM-LDA model, to be able to model two cues in a principled manner with prediction ability. It is to our knowledge the first LDA approach to be used for robotic cognitive reasoning. In this section, we describe the model. More details on how this model is employed in robotic applications are described in the next section.

A. Model

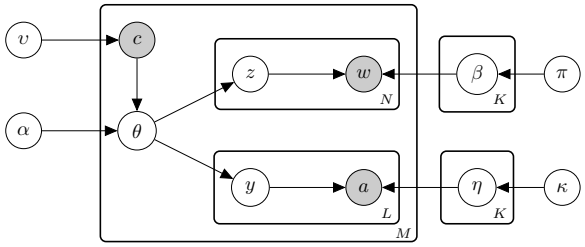


Fig. 2. Graphic Representation of Labeled Multi-LDA.

For consistency, we inherit the terminology from LDA [23] in this section to explain our proposed LM-LDA. Fig. 2 shows the graphic representation. Each document from the corpus has one categorical label c and contains two data cues. The first cue consists of N words w and the second cue consists of L terms a. In the following we use "words" to denote the first cue and "terms" to denote the second, for clarity. However, there is no principal difference between the cues, the model is symmetric. In a robotics scenario, a document is a collection of any type of information, and words/terms are information units. Each document is modeled as belonging to K different topics, with probability proportions given by θ, see Fig. 2. β/η models the word/term distribution in each topic. The generative process for a document is modeled as follows.

For each of the M documents in the corpus:

1) Choose a category label $c \sim p(c|\nu)$. $\nu \in \mathbb{R}^C$ and ν follows categorical distribution, which reflects the document category distribution in corpus.
2) Draw topic proportion θ from α by $\theta \sim Dir(\alpha)$. Here, α is a K-dimensional Dirichlet parameter, where K is the number of topics.
3) For each of the N words w from the first cue:
   a) Choose a topic assignment of a word $z \sim Multinomial(\theta)$
   b) Choose a word w from β_z. Here, β is constructed by K vectors, in which β_k is a V-dimensional vector of Multinomial distribution which reflect the distribution of words under topic k, where V is the word vocabulary size.
4) For each of the L terms a from the second cue:
   a) Choose a topic assignment of a term $y \sim Multinomial(\theta)$

b) Choose a term a from η_y. η is constructed by K vectors, in which $\beta_k \in \mathbb{R}^R$ that follows Multinomial distribution which reflect the distribution of terms under topic k, where R is the term vocabulary size.

As described above, we can see that words w and terms a are independent given the topic distribution, which makes this model suitable to model two different cues without any extra assumptions. On the other hand, by sharing the same topic space with the same class label, correlations between the cues will be encoded in the topic distribution, since the topic model is trained with instances of both cues simultaneously. The independence of the two cues given topic distribution θ, makes inference with only one cue w possible. Thanks to the correlation between the cues encoded in the topic model, the inference will benefit from the model being trained with two cues, even though only one is observed for inference. The reason for this is that the two cues make the topics more tuned to information present in both cues (e.g., object shape for a visual and a grasp cue), and less tuned to variation which only depends on one cue (e.g., object surface patterns which is uncorrelated with the grasp cue). In a related work, Zhang et al. [35] used an extra prior in the model to make the topics more correlated with object class. The additional prior can be seen as a second cue to help finding a more robust representation of the data characteristics. Furthermore, given c, θ and η, possible terms from the unobserved cue a can also be reconstructed.

B. Parameter estimation

To estimate the unknown parameters, we use Gibbs Sampling as described in [36]. In our case, there are two cues. The update equation for a word with index $i = (m, n)$ is:

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) \propto \frac{n^{(t)}_{1k,\neg i} + \pi_t}{\sum_{t=1}^{V} n^{(t)}_{1k,\neg i} + \pi_t} \cdot \frac{n^{(k)}_{m(1),\neg i} + n^{(k)}_{m(2),\neg i} + \alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m(1)} + n^{(k)}_{m(2)} + \alpha_k] - 1} \quad (1)$$

The update equation for a term with index $j = (m, l)$ is:

$$p(y_j = k|\vec{y}_{\neg j}, \vec{a}) \propto \frac{n^{(p)}_{2k,\neg j} + \kappa_p}{\sum_{p=1}^{R} n^{(p)}_{2k,\neg j} + \kappa_p} \cdot \frac{n^{(k)}_{m(1),\neg j} + n^{(k)}_{m(2),\neg j} + \alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m(1)} + n^{(k)}_{m(2)} + \alpha_k] - 1} \quad (2)$$

The parameters are estimated as:

$$\theta_{c,k} = \frac{n^{(k)}_{m(1)} + n^{(k)}_{m(2)} + \alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m(1)} + n^{(k)}_{m(2)} + \alpha_k] - 1} \quad (3)$$

$$\beta_{k,t} = \frac{n^{(t)}_{1k} + \pi_t}{\sum_{t=1}^{V} n^{(t)}_{1k} + \pi_t} \quad (4) \qquad \eta_{k,p} = \frac{n^{(p)}_{2k} + \kappa_p}{\sum_{p=1}^{R} n^{(p)}_{2k} + \kappa_p} \quad (5)$$

where $n^{(\cdot)}_{\cdot,\neg i}$ indicates that the token i is excluded. In the model, we can treat θ as a $M \times K$ matrix, β as a $K \times V$ matrix and η as $K \times R$ matrix. π, κ, α are smoothing parameters. In Equation (3), $n^{(k)}_{m(1)}$ stands for the number of words which are assigned to the topic k from the document m, and $n^{(k)}_{m(2)}$ stands for the number of terms which are assigned to the topic k from the document m. In this case, $\theta_{m,k}$ presents the probability of chosen topic k from a document. In Equation (4), $n^{(t)}_{1k,\neg i}$ stands for the number of words t assigned topic k, and the sum in the denominator shows the number for all the words which are assigned to topic k with smoothing parameter π. In this way, $\beta_{k,t}$ stands for the probability to sample the word t if topic k is assigned. Similarly, in Equation (5), $\eta_{k,p}$ stands for the probability to sample the term p if topic k is assigned. The update Equations (1) and (2) are straightforward, as a certain word/term is sampled based on the estimated probability from the rest of the data with Gibbs Sampling.

C. Inference

The inference problem is to calculate the most probable category given a new observation and the learned parameters, $\arg\max_p(c \mid \text{new observation}, \text{learned parameters})$, which is in a similar manner as [27]. All parameters are learned during the training phase. When classifying a new observation, the topic assignment is updated with Gibbs Sampling as in [36]. The maximum likelihood of the categorical label c is then computed conditioned on the topic assignment distribution. Since the two cues share the same topic space and c is inferred from the topic distribution, the inference of categorical label can be from both cues (bags of words w and terms a), or from only one of them. Furthermore, in the case where only one cue is observed, the expected value of the other cue can be inferred. The reason, as discussed above, is that the topic-word/term distributions β and η are learned in the training phase, so that they are correlated topic-wise. Given the distribution of words for a certain observed instance, the expected distribution of terms can thus be inferred via the underlying latent topic distribution. These two properties of the LM-LDA model are important for the functionality of its application to visual classification, which is described next.

IV. Contextual Modeling

As discussed above, LM-LDA is able to provided a principled framework for contextual modeling. In this section, we describe two contextual classification tasks, action recognition and functional object classification, in more detail, explaining particular difficulties with each task. In the experiments below, the application of LM-LDA in these two scenarios are studied. The bag of words representation together with LDA [27], [34], [29] requires discretization of the feature space, which can potentially lead to lost information. This problem can be addressed by using soft assignment of feature descriptors to words [37].

A. Classification of Action from Visual Motion and Object Interaction

Recognizing and reasoning about activities of daily living is an extremely important skill for a robot functioning in human environments. Traditionally, action recognition is done using motion features, e.g., STIPs [38]. Recently, [7] argue that action involving objects strongly depends on the objects involved, and propose the bag of active object (AO) feature,

based on which object classes are in contact with the human hand during the action and which are passively presented. This feature requires (manual or automatic) detection of the hand and all objects in the video sequences, whereas a visual motion feature such as STIP is easy to obtain from data without supervision. In addition, robot needs to know the tool to use and environment to carry out the task given by a human, which makes it important for robot to learn the task and involved objects mapping as well.

The two cues complement each other: While the STIP features capture differences between visually different actions, the AO features instead capture differences between actions involving different kinds of objects. This implies that it would be beneficial to classify from both cues simultaneously. The labeling requirements for the AO features implies that this feature might not be accessible for each new instance to be classified. All these together makes LM-LDA a suitable classifier.

The LM-LDA is applied in the following way: Every action clip is a document. Words w are STIP features [38], quantized according to a codebook learnt from the data. Terms a are AO features as defined in [7]. The category label c indicates action class.

Below, experiments with this LM-LDA setup are carried out, using the dataset from [7].

B. Functional Classification of Objects from Appearance and Task-Orientated Grasp

An important capability of robots functioning in unstructured environments is to be able to pick up unknown objects and use them for a specific task; this is referred to as task-oriented grasp planning [39], [4]. The planning requires categorization of previously unseen objects into functional classes based on how they should be grasped for the purpose of the current task. The classification is task-dependent: a knife is grasped differently depending on if the intention is to cut with it or to hand it over to someone else.

Conversely, how an object is usually grasped for different tasks, provides rich information to discriminate between object categories. It is therefore beneficial to use object appearance and grasp parameters in a contextual manner for object classification. Furthermore, while grasp information might be available at training time (each training image labeled with typical grasps), it is not likely that a new object instance to be classified comes with this information. All this together makes LM-LDA a suitable classifier.

The LM-LDA is applied as: Words w are SIFT features [40], classified as visual words according to a codebook learnt from the data. Terms a are grasp configurations that include global hand position, orientation and finger configuration at the single time frame when fingers contact the object (see Figure 6), classified into R grasp term categories. The categorical label c is an object class-orientated task, as in Section IV-A. A document is one object instance, which is presented with a bag of visual words w, and a bag of possible task-oriented grasps a.

A/P	Objects
Active Objects	fridge, microwave, mug/cup, oven/stove, soap liquid
Passive Objects	bed, cell, dish, kettle, monitor, soap liquid, tooth paste, book, dent floss, door, laptop, pan, tap, TV, bottle, detergent, fridge, microwave, pitcher, teabag, TV remote

TABLE I

ACTIVE AND PASSIVE OBJECTS IN THE DATASET FROM [7].

Similarly to the application in Section IV-A, the trained LM-LDA model can be used to assist classification when observing a task being performed on an object ,or when observing just an image of the object. In addition, when seeing a new object instance, the object category can be inferred, and then subsequently the most probable grasp configurations corresponding to different tasks can be predicted.

V. EXPERIMENTS

Experiments were carried out to evaluate the performance of the LM-LDA model in correspondence with the applications in Sections IV-A and IV-B separately.

A. Classification of Action from Visual Motion and Object Interaction

Action classification is evaluated in this experiment with STIP features as the regular visual cue and AO feature as the context. For this experiment, the Activities of Daily Living (ADL) dataset [7] is used – the dataset originally used to evaluated the AO feature. It contains a large number of video clips with first-person camera views of 20 people performing unscripted everyday activities. For this experiment we selected a representative subset containing 7 daily activities. The 7 actives are chosen so as to give the subset the same difficulty level as the original dataset.

The first cue are STIP features, extracted from the video clip. STIP features are quantized as visual words according to a codebook of size $V = 256$, learned from the data. Each word w in the LM-LDA model is a visual STIP word.

The second cue are AO features. Each AO feature is an 26 dimensional vector which is the scaled detection score of 5 possible active objects and 21 possible passive objects, as shown in Table I. AO features are quantized as action words according to a codebook of size $R = 30$. Each word a in the LM-LDA model is an AO word.

Parameters are $K = 25$, $\alpha = 0.5$, $\pi = 0.1$ and $\kappa = 0.1$. Here,α, π, κ, are hyper parameters which are less than one to ensure the sparsity. The model is comparably robust relating to the hyper parameters since the Dirichlet distribution has the well-known "rich-get-richer" behavior. The parameter setting here is one of the most common used setting in topic modeling [41]. The performance will be robust when the number of topics is sufficiently large to describe the data for LDA, which has shown in [41], [23], the setting here is a sufficient number in this case.

Experiments are performed by employing one person's data for testing and the rest of the data for training, in the same manner as in [7]. We run the experiments with every person as test set; the final result is an average over

Method	Performance
SVM+bag AO [7]	38.31%
SVM+pyramid AO [7]	38.58%
LM-LDA train both, test STIP	$38.23\% \pm 2.48\%$ [1]
LM-LDA train both, test both	$\mathbf{46.04\%} \pm 2.90\%$ [1]

TABLE II

THE FIRST TWO RESULTS ARE GIVEN BY RUNNING THE ORIGINAL IMPLEMENTATION OF BAG OF AO AND PYRAMID AO METHOD FROM [7]. TRAIN BOTH, TEST BOTH INDICATES THAT STIP AND AO ARE USED IN BOTH TRAINING AND INFERENCE. TRAIN BOTH, TEST STIP INDICATES THAT BOTH CUES ARE USED FOR TRAINING BUT ONLY STIP IS USED IN INFERENCE.

all the experiments for each person. Four different settings were used: a) only STIP features w during both training and testing (single cue baseline), b) only AO features a during both training and testing (single cue baseline), c) both cues w and a accessible during training but only w during testing (corresponding to the common situation that one cue is expensive to obtain), and d) both cues w and a accessible both during training and testing.

Fig. 3 shows the resulting confusion matrices with topic models. Unlike [7], where one pyramid representation of AO features is used to represent every video clip, the raw AO features are used here, where each video clip is a document with tens to hundreds of raw AO features (words), i.e., one AO feature per second in the video clip as given in the dataset. This is a weaker representation, since it does not contain any spatial information. Fig. 3(b) shows the performance with only bag of quantized STIPs features. Fig. 3(c) shows the result of training with both bags of AO and bags of STIP, but only STIPs are used for inference. In this case, no object information/detection is applied in the inference. This result correspond to the third row of Table II. The added contextual cue at training time increases the classification result by 7.74% compared to only STIPs, even though the contextual cue is not present at test time. Table II shows that it reaches the same level of performance as the SVM with pyramid presentation of AO.

Furthermore, even though the AO cue is missing in the inference, the most possible objects involved in the action can be inferred from the STIP features, without applying any explicit object detection. Table III shows the top objects inferred from the LM-LDA model, given STIP feature observations of different actions. We can see that the model is able to give noisy but reasonable inference on which objects are involved in the action. For example, there are several kitchen items inferred from *Making cold food/snack*. Furthermore, comparing *Drinking water/bottle* and *Drinking water/tap*, we see that fridge(active) is assumed to be involved in drinking from bottle since there are several videos of taking a bottle of water from fridge; however, tap is instead the first object of drinking from tap and fridge(passive) is involved since this action frequently happens in a kitchen scene. On the other hand, *Combing hair* and *Washing hands/face* have similar object presentation since they both happen in a bathroom

[1]The mean and standard deviation are derived by repeating the experiments 5 times.

Action	Top Objects
Combing hair	tap, mug/cup, dish
Washing hands/face	tap, mug/cup
Drying hands/face	mug/cup, tap, TV
Drinking water/bottle	mug/cup, dish, fridge(active)
Drinking water/tap	tap, mug/cup, door, fridge(passive)
Making cold food/snack	mug/cup, microwave, fridge(passive), fridge(active), dish
Using cell	mug/cup, TV, tap, fridge(passive)

TABLE III

THE MOST PROBABLE OBJECTS INFERRED FROM THE MODEL GIVEN DIFFERENT ACTIVITY CLASSIFICATIONS. NO DIRECT OBJECT DETECTION WAS PERFORMED, AND THE ACTION CLASSIFICATION WAS DONE FROM STIP FEATURES ONLY. THE TOP WORDS WERE GENERATED BY TAKING THE TOP 10 AO WORDS GIVEN THE CORRESPONDING ACTION AND SET A THRESHOLD. AO FEATURES ARE NORMALIZED TO $0 \sim 100$ AND THRESHOLD IS SET TO 35 HERE.

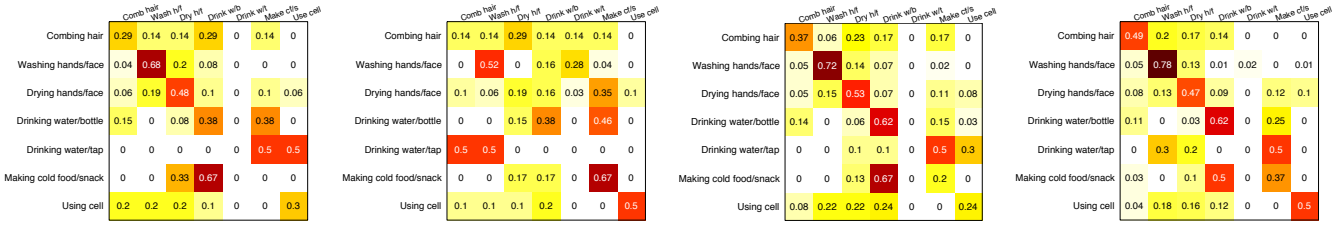
with tap and mug/cup present in the scene. This shows our second contribution.

Finally, Fig. 3(d), shows the result of using 2 cues (AO and STIPs) in both training and inference, which correspond to the last row of Table II. The second row is the current state-of-the-art on the ADL dataset, other comparisons can be found in [7]. We can see that LM-LDA outperforms the state-of-the-art on the ADL dataset, thanks to its principled way of cue-integration. This together with the result in Fig. 3(c), are experimental evidence of our first contribution.

B. Functional Classification of Objects from Appearance and Task-Oriented Grasp

The training takes place in a leave-one-out manner with three rounds, where one object instance is left out for testing.

Fig. 5 shows the resulting confusion matrices. The results cohere with those found in Section V-A, which shows our first contribution. It can be noted that the classes Knife and Screwdriver are especially hard to distinguish without grasp information (Fig. 5(a)). Adding grasp info in the training enables the topics to pick up on some of the task specific appearance aspects, increasing performance a bit (Fig. 5(b)), even though no grasp information is present at test time. However, the two classes look too similar to reach a good discrimination performance using appearance only for testing. If a grasp is observed at test time, the classification performance is much better. As noted above, the model can infer a range of allowed actions given just an object appearance. Fig. 6 gives the top 5 most probable grasps associated with a testing knife (not present in the training data) for two different tasks, Hand-over and Tool-use, which shows our second contribution. The visualized grasps correspond to intuition: when handing a knife over, the grasp can be applied from the side using a fingertip grasp or from the end of the handle, while a knife needs to be grasped with the whole hand on the handle if it is going to be used for cutting or carving. Fig. 7 gives more examples with other objects. In Fig. 7, the most probable grasp is presented in most cases, and the second most probable grasp is shown only if the first one is shared with another task. These grasps are generated by the topics, which can be understood in a similar way as intention [6]. Hence it can be easily used cross



(a) Train STIP, test STIP: 30.49% (b) Train AO, test AO: 34.4% (c) Train both, test STIP: 38.23% (d) Train both, test both: 46.04%

Fig. 3. Confusion matrices for LM-LDA classification, different settings.

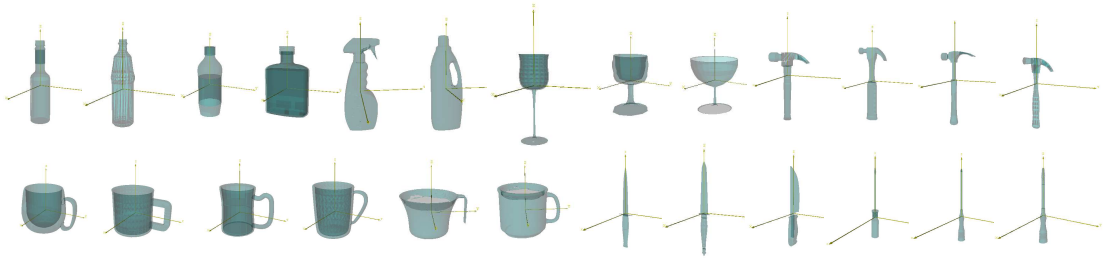
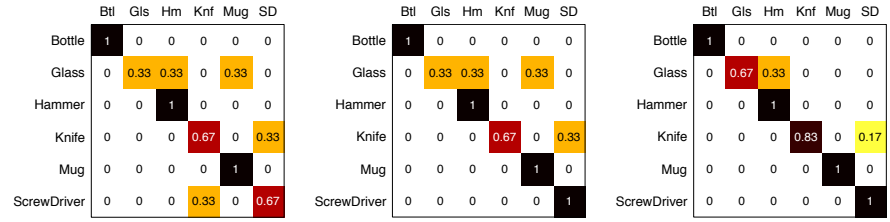
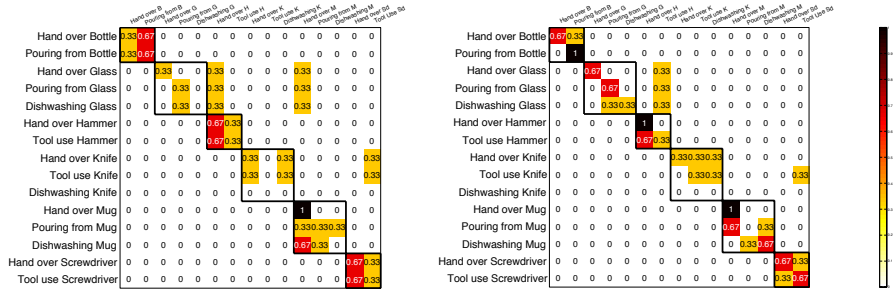


Fig. 4. All the instances used in the Appearance and Task-Orientated Grasp experiment: 6 Bottles, 3 Glasses, 4 Hammers, 6 Mugs and 6 knives.



(a) Train SIFT, test SIFT: 77.78% (b) Train both, test SIFT: 83.33% (c) Train both, test both: 91.67%



(d) Train both, test SIFT (e) Train both, test both: 59.52%

Fig. 5. Confusion matrices for the Task-Orientated Grasp experiment. (a-c) Confusion matrices for object classification (marginalized over grasp). (d-e) Confusion matrices for task classification (black square indicating tasks with the same object class, such as "Pouring from Mug" and "Dishwashing Mug", which are impossible to distinguish given appearance only). Parameter settings: $V = 192$, $R = 192$, $K = 60$ $\alpha = 0.5$, $\kappa = 0.1$ and $\pi = 0.1$

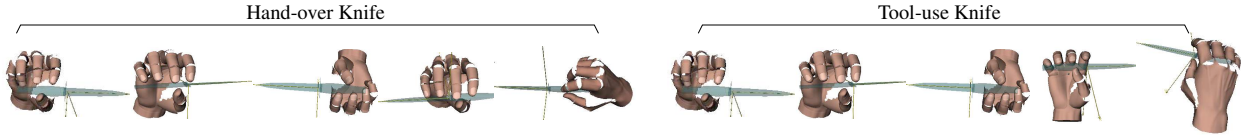


Fig. 6. The most probable grasps on an unseen knife with task-oriented grasp Hand-over and Tool-use.

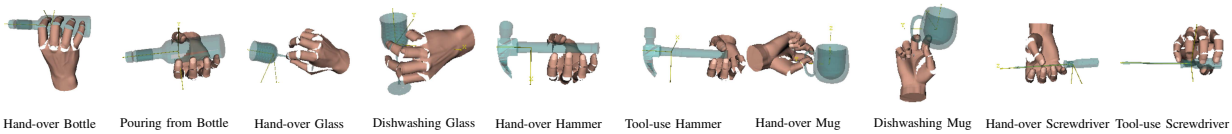


Fig. 7. Grasp examples on other objects.

different hand configurations and further generate precise grasps as in [6].

VI. Conclusions

We proposed a new model for contextual modeling, Labeled Multi-modal LDA, which models two separate cues as generated from the same topic space, for the purpose of robotic cognition. Using this model it is possible to encode information about the correlation between two contextually dependent cues, so that the classification makes use of this information even if only one of the cues are present at execution time.

Furthermore, information about the missing cue can be inferred from the model, given the other cue, available for observation. This is highly relevant for, e.g., robotic grasping applications, where suitable grasps can be inferred from object appearance, given that the model has been trained with grasps and object appearance in concordance.

Experiments showed that the LM-LDA model outperforms the state-of-the-art method on the highly challenging Activities of Daily Living (ADL) dataset [7].

Directions of future research include the addition of spatial information to the representation, as discussed in Section II. Another avenue to explore is to include private topic spaces for the two cues in addition to the shared topic space. The highly related topics will then be captured in the shared topic space and information private to one of the cues will be explained by the private topics. An example in our grasping application is appearance aspects that are unrelated to grasp: When modeling the class Mug, topics relating to the appearance of mug handles would be generated from the topic model common to the appearance and grasp cue, while topics relating to the type of pattern printed on the mug would be generated from an appearance-only topic model. We also plan to perform real world grasping test on robots, generating the grasps from the model given object appearance.

References

[1] J. Gall, A. Fossati, and L. van Gool, "Functional categorization of objects using real-time markerless motion capture," in *CVPR*, 2011.

[2] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *PAMI*, vol. 31, no. 10, pp. 1775–1789, 2009.

[3] H. Kjellström, J. Romero, and D. Kragic, "Visual object-action recognition: Inferring object affordances from human demonstration," *CVIU*, vol. 115, no. 1, pp. 81–90, 2011.

[4] M. Madry, D. Song, and D. Kragic, "From object categories to grasp transfer using probabilistic reasoning," in *ICRA*, 2012.

[5] H. Kjellstrom, J. Romero, and D. Kragic, "Visual recognition of grasps for human-to-robot mapping," in *IROS*, 2008.

[6] D. Song, N. Kyriazis, I. Oikonomidis, C. Papazov, A. Argyros, D. Burschk, and D. Kragic, "Predicting human intention in visual observations of hand/object interactions," in *ICRA*, 2013.

[7] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *CVPR*, 2012.

[8] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots: Learning a visually grounded storyline model from annotated videos," in *CVPR*, 2009.

[9] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR*, 2010.

[10] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *CVPR*, 2007.

[11] T. Malisiewicz and A. A. Efros, "Beyond categories: The visual memex model for reasoning about object relationships," in *NIPS*, 2009.

[12] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, 2007.

[13] A. Torralba, "Contextual priming for object detection," *IJCV*, vol. 53, no. 2, pp. 169–191, 2003.

[14] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *ICCV*, 2007.

[15] B. Siddiquie and A. Gupta, "Beyond active noun tagging: Modeling contextual interactions for multi-class active learning," in *CVPR*, 2010.

[16] D. J. Moore, I. A. Essa, and M. H. Hayes, "Exploiting human actions and object context for recognition tasks," in *ICCV*, 1999.

[17] P. Peursum, G. West, and S. Venkatesh, "Combining image regions and human activity for indirect object recognition in indoor wide-angle views," in *ICCV*, 2005.

[18] A. Billard, S. Calinon, R. Dillman, and S. Schaal, "Robot programming by demonstration," in *Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. New York, NY, USA: Springer-Verlag, 2008, ch. 59.

[19] V. Krüger, D. Herzog, S. Baby, A. Ude, and D. Kragic, "Learning actions from observations," *IEEE Robotics Automation Magazine*, vol. 17, no. 2, pp. 30–43, 2010.

[20] A. Herzog, P. Pastor, M. Kalakrishnan, L. Righetti, T. Asfour, and S. Schaal, "Template-based learning of grasp selection," in *ICRA*, 2012.

[21] C. H. Ek, D. Song, K. Huebner, and D. Kragic, "Task modeling in imitation learning using latent variable models," in *IEEE International Conference on Humanoid Robots*, 2010.

[22] T. Hofmann, "Probabilistic latent semantic analysis," in *UAI*, 1999.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[24] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina, "Clustering the tagged web," in *ACM International Conference on Web Search and Data Mining*, 2009.

[25] D. M. Blei and J. D. McAuliffe, "Supervised topic models, http://arxiv.org/abs/1003.0783," 2010.

[26] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *ICCV*, 2005.

[27] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, 2005.

[28] X. Wang and E. Grimson, "Spatial latent Dirichlet allocation," http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.143.4231, 2009.

[29] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *CVPR*, 2009.

[30] N. Rasiwasia and N. Vasconcelos, "Holistic context models for visual recognition," *PAMI*, vol. 34, no. 5, pp. 902–917, 2012.

[31] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Spatial-DiscLDA for visual recognition," in *CVPR*, 2011.

[32] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora," in *Conference on Empirical Methods in Natural Language Processing*, 2009.

[33] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: Discriminative learning for dimensionality reduction and classification," in *NIPS*, 2008.

[34] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[35] C. Zhang, C. H. Ek, A. Damianou, and H. Kjellstrom, "Factorized topic models," in *International Conference on Learning Representations*, 2013.

[36] G. Heinrich, "Parameter estimation for text analysis," http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.158.9854, 2008.

[37] D. W. G. L. D. Hanukaev, "Latent dirichlet allocation topic model with soft assignment of descriptors to words," in *ICML*, 2013.

[38] I. Laptev, "On space-time interest points," in *IJCV*, 2005.

[39] D. Song, K. Huebner, V. Kyrki, and D. Kragic, "Learning task constraints for robot grasping using graphical models," in *IROS*, 2010.

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[41] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.